NASA SP-300

PHYSICS OF THE SOLAR SYSTEM

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

NASA SP-300

PHYSICS OF THE SOLAR SYSTEM

Edited by

S. I. RASOOL

Goddard Institute for Space Studies

Prepared at NASA Goddard Space Flight Center

Scientific and Technical Information Office 1972

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Washington, D.C.

For Sale by the Superintendent of Documents,
U.S. Government Printing Office, Washington, D.C. 20402
Library of Congress Catalog Card Number 72-600107

Page intentionally left blank

FOREWORD

Physics of the Solar System is based on lectures given at the Fourth Summer Institute for Astronomy and Astrophysics held at the State Univeristy of New York at Stony Brook, from June 17 to July 15, 1970. The Summer Institute, sponsored by the National Aeronautics and Space Administration, was directed by Dr. Hong-Yee Chiu and Dr. S. I. Rasool.

The material covers a broad range of topics in the physics of the Sun, the structure of the planets and their atmospheres, and the origin and evolution of the solar system and of planetary atmospheres, and presents a view of current problems associated with these fields.

The editor is grateful to the authors for their respective contributions; to Dr. J. Hardorp. Department of Earth and Space Sciences at SUNY, for his valuable assistance during the summer institute; to N. Hartunian, R. Russell, and J. Warman for their assistance in preparing the lecture notes from the taped lectures; to Miss Mary Yastishak for her administrative assistance in the planning of the conference and in coordinating the manuscript; and to Miss Margaret Kavanau for the typing of the manuscript.

S. I. Rasool

Page intentionally left blank

CONTENTS

CHAPTER 1
INTRODUCTION TO SOLAR PHYSICS

J. C. Brandt
Goddard Space Flight Center
Greenbelt, Maryland

I. OVERVIEW

The purpose of this introductory chapter is to sketch the basic physics involved in the large-scale structure of the Sun's interior and atmosphere. Details and necessary refinements are given in the chapters that follow.

Our great interest in the Sun is dictated by its proximity. The Sun can influence the Earth directly; for example, storms on the Sun often disrupt radio communications and produce auroras. In addition, the Sun serves as a representative of stars in general, with the considerable advantage of nearness. If the Sun were removed to the distance of the nearest star, quantities such as flux (which decreases as the inverse square of the distance) would be smaller by a factor of 10^{11}. In addition, we can study details of surface processes on the Sun which are inaccessible to us in the distant stars.

Because we can gather so much more information about the Sun than other stars, we need to know whether or not the Sun is typical of other stars, i.e., whether we can apply knowledge of solar properties to the study of stars in general. The Sun is indeed representative of the stars that lie in the disk of the Galaxy. Its chemical composition and kinematic properties are typical of the stars that make up the flat, rotating disk of the Galaxy. The Sun lies at the inner edge of a spiral arm, at a distance of approximately 10 kpc from the center of the Galaxy. In one respect, the Sun is atypical: It is not part of a double- or multiple-star system, whereas about two-thirds of all stars in the Galaxy belong to such systems.

Some of the fundamental physical properties of the Sun are as follows:

(1) The mass of the Sun is 1.99×10^{33} g, designated by the symbol $M_\odot$. The mass of the Sun is very well determined compared to those of other stars because we can use the orbital parameters of the planets to determine it. Masses for double-star systems are difficult to determine with comparable accuracy.

(2) The mean radius of the Sun is 6.96×10^{10} cm, designated by the symbol $R_\odot$. The Sun departs very slightly from a spherical shape, as is discussed by R. Dicke.*

(3) The Sun rotates differentially, having an equatorial period of 25.36 days (tangential velocity of 2 km-s^{-1}) and a polar period of approximately 30 days.

(4) The Sun's luminosity ($L_\odot$) is determined from measurements of the *solar constant,* defined as the amount of energy received per square centimeter by a surface in space perpendicular to the Earth-Sun line at the Earth's mean distance from the Sun. Rocket and mountain-top measurements have yielded values close to 1.95 cal/cm^2-min, which integrates to a solar luminosity of about 3.9×10^{26} joule/s.

(5) The Sun's age has been determined from radioactive-dating schemes used on meteorites and lunar material. They generally yield an age of 4.5×10^9 years. Stellar-evolution calculations are consistent with a similar age for the Sun.

All knowledge of the solar interior is indirect, so that we cannot discount proposals that the Sun also rotates differentially as a function of distance from the center. Neutrino measurements hold out some hope of direct observations of the solar interior, but so far these have departed significantly from the initial predictions of measurable fluxes.

The Sun can be divided roughly into six zones (Figure 1): (1) the energy producing core; (2) a region of radiative equilibrium ending at $r/R_\odot \approx 0.86$; (3) a convective zone called the hydrogen-convective zone; (4) a thin radiative zone, called the photosphere, which is approximately 300 km thick; (5) a few-thousand-kilometer-thick zone called the chromosphere; and (6) a tenuous outer region called the corona. As the corona expands into the interplanetary medium, it is called the solar wind. The temperature of the core is approximately 15×10^6 K. Energy is transmitted in the region above the core by radiation until it reaches the hydrogen-convective zone, where convective energy transfer takes over. In the photosphere, radiative transport again becomes dominant. At the boundary between the photosphere and chromosphere, the photons are largely decoupled from the matter; the temperature here is the minimum solar temperature of about 4300 K. The boundary defined by the chromosphere and corona is at a temperature of about 10^6 K, so that the chromosphere, and particularly the "transition zone" between the chromosphere and the corona, possesses an extremely high temperature gradient. Energy is transferred to the outer atmosphere by the mechanical energy of

*See Chapter 2, "Internal Rotation of the Sun", by Dicke.

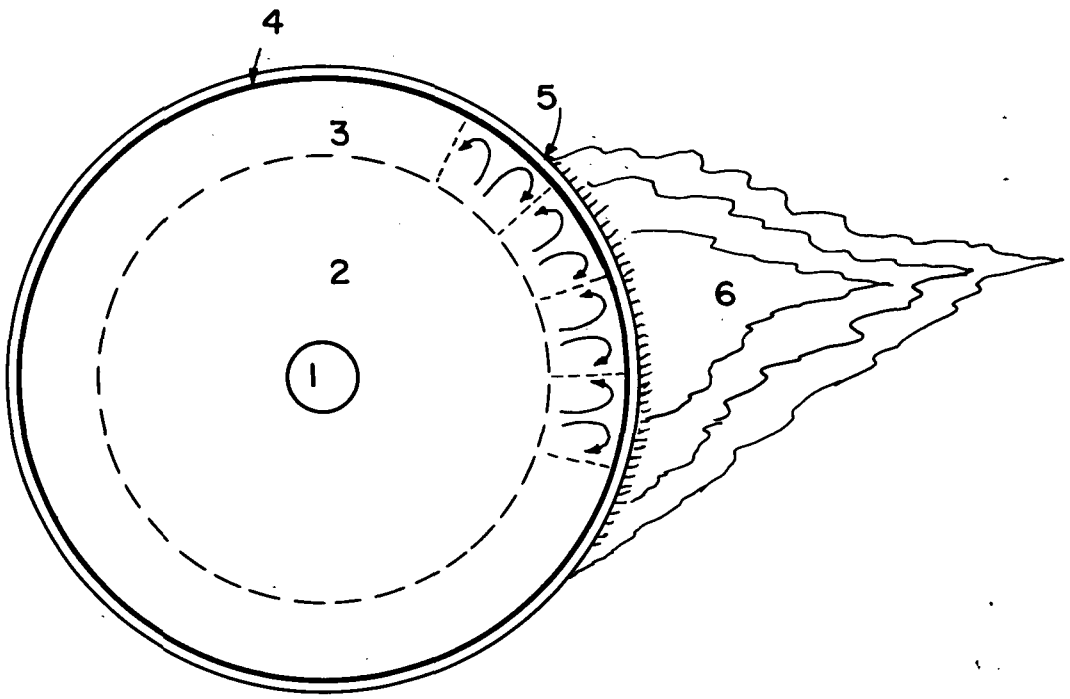


Figure 1.—Different zones of energy transport in the Sun (see Section 1).

wave motions; in the corona, conduction is important also. The corona has a temperature of about 2×10^6 K; it expands and becomes the solar wind, with a velocity of about 400 km-s^{-1}, electron and proton densities of 5 to 10 per cm^3, T_e of 2×10^5 K, and T_p of 5×10^4 K at the Earth's orbit. Although the major outward energy flow from the Sun is the photon flux, the matter outflow is also important for certain calculations such as those concerning the Sun's rotation rate.

II. INTERIOR

The structure of the solar interior is calculated on the basis of standard assumptions and equations:

(1) Hydrostatic equilibrium.

(2) The mass equation (relating the change of mass with radius to the density).

(3) The luminosity equation (relating the change of luminosity with radius to the energy generation rate).

(4) An equation of state. The perfect-gas equation is a good approximation in the solar interior.

(5) Equations specifying the rates of energy generation. The principal source of energy is the proton-proton chain, although the carbon cycle may contribute a few percent of the energy in the central region.

(6) An equation of energy transport. In the deep interior, energy transport is by radiation. The source of opacity is primarily bound-free transitions of the heavier ions which are not completely ionized. Nearer the surface, however, convective transport becomes important, as is discussed in Section III.

Analytical solutions to the equations for stellar interiors are available only for idealized cases of little practical interest. Models of the solar interior are computed with the aid of high-speed computers, subject to the conditions that the solar radius, mass, and luminosity are reproduced and that the model matches onto the solar atmosphere.

III. HYDROGEN CONVECTION ZONE

Convective energy transport can become important if the temperature gradient tends to become too large. This can be seen from the Schwarzschild instability criterion,

$$\left|\frac{dT}{dr}\right|_{\text{str}} > \left|\frac{dT}{dr}\right|_{\text{ad}} \tag{1}$$

Consider a small bubble of gas that is adiabatically displaced upward while remaining in pressure equilibrium. If the structural gradient is greater than the adiabatic gradient, the gas in the test bubble would cool less and thus be at a higher temperature than the surrounding gas. Since the bubble and the surroundings are in pressure equilibrium, the bubble must be less dense than the surroundings and, hence, subject to a buoyancy force. The result is that a bubble displaced upward produces a force that accelerates its upward motion. This situation is convectively unstable, and the process of convection tends to reduce the temperature gradient to the adiabatic value. In practice, one needs to consider the effects of viscous forces and heat exchange (i.e., the range of values of the Rayleigh number) to see if convection actually occurs.

The zone beneath the Sun's surface is convective because the opacity increases sharply. A high opacity means a short mean free path for photons, so that a high temperature gradient is needed to move the entire solar luminosity through the subsurface layers. The gradient needed is so high that the Schwarzschild criterion is met, and convection occurs.

Why does the opacity increase sharply below the photosphere? Consider the Saha equation for the abundance of H^-, which is the principal source of opacity in this region. We can write

$$N(\mathrm{H})N_e/N(\mathrm{H}^-) = \text{constant} \times T^{3/2}e^{-I/kT} = f(T)\,, \tag{2}$$

where $f(T)$ is a slowly varying function in the region of interest. Since $N(H^-)/N(H) \sim N_e/f(T)$, the opacity can vary substantially only if N_e does. For the photosphere, $N_e/N(H) \approx 10^{-4}$, since most of the electrons come from the easily ionized metals. Below the photosphere, T rises to about 10 000 K, hydrogen becomes substantially ionized, and $N(H^-)/N(H) \approx N_e$ increases by several orders of magnitude. Below the convection zone, all the hydrogen is ionized, and other opacity sources are important. It should be pointed out that the temperature gradient in the convective zone does not approach the adiabatic gradient, because of the relative inefficiency of the convective process in the Sun.

These convective motions produce observable features on the solar surface: granulation and supergranulation. Solar granulation is a small-scale pattern observed on the solar surface in high-resolution photographs. The mean diameter of the granules is about 700 km, and mean half-life is about 4 minutes; the granules are bright in the middle and dark near the edges. If a spectrograph slit is placed over this pattern, one obtains a "wiggly" line image, indicating motion toward and away from the observer. These velocities are approximately 1 km-s^{-1}. Figure 2 shows the type of systematic motion believed to be present in these convective cells. A detailed study of convection using the so-called "mixing-length" theory leads to similar velocities, but there are problems with this derivation.

Supergranulation is characterized by cells with diameters of about 30 000 km and periods of the order of 1 day. Furthermore, they seem to be related to the Ca-II chromospheric network. The Ca-II chromospheric network is studied from observations of the emission that occurs in the center of the Ca-II K absorption line. A portion of a photograph taken in the Ca-II K-line center is schematically indicated in Figure 3. One finds clumps of bright emission, called plagettes, defining an apparent "supercell". These bright clumps coincide with intense regions of magnetic field strength (as high as 100 gauss or more). Intense magnetic fields can enhance the

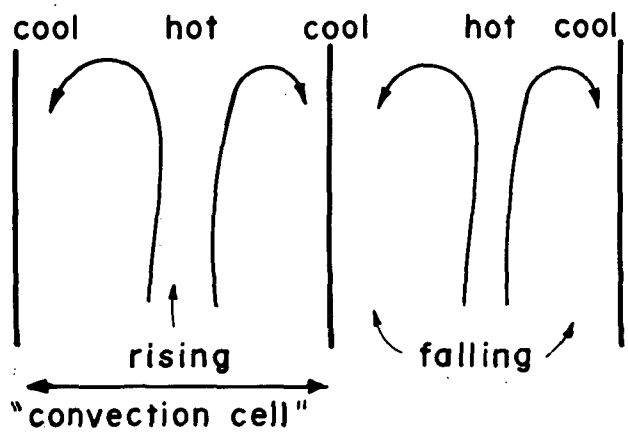


Figure 2.–Schematic of convection cells (vertical section).

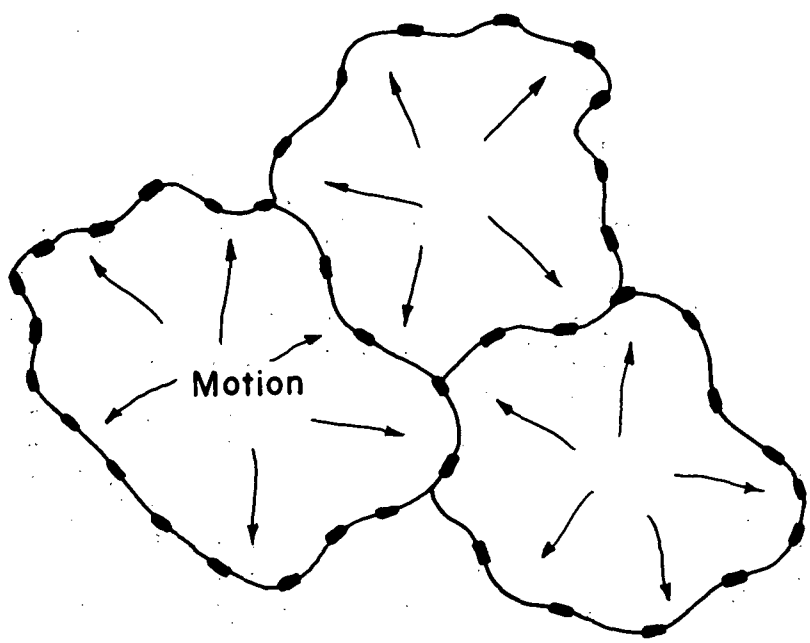


Figure 3.–Schematic of supergranulation cells on the solar surface.

generation and deposition of mechanical energy; such mechanical energy could be generated by the motions in the hydrogen convection zone and penetrate into the photosphere and the chromosphere to form the bright clumps. The magnetic fields in the clumps are "frozen" in the solar plasma and are convected to the cell boundary by the supergranulation motions themselves. Thus, the bright chromospheric regions should be found concentrated in the boundaries of the supergranulation cell, as observed.

As a final observation on convective energy transfer, it has been found that the entire photosphere moves up and down with a period of approximately 300 s and an amplitude of 0.5 km-s^{-1}. These pulsations could be driven by the hydrogen convection zone.

IV. PHOTOSPHERE

The photosphere is the source of almost all of the photons we receive from the Sun. A step in the direction of understanding the physics of how they are emitted is the construction of a model atmosphere. A crucial observational step is to determine the "temperature" of the photosphere as characterized by the radiation emitted. There are three ways to do this:

(1) We can fit a Planck curve to the continuum spectrum. This method yields a temperature of 6000 K.

(2) We determine the solar constant and use Stefan's law (flux $\sim T^4$) to get an effective temperature $T_{\text{eff}} = 5750$ K.

(3) We can look at the ionization states of various elements in the spectrum and define an ionization temperature.

Obviously, a good model of the solar photosphere must reproduce the continuum spectrum, the total flux, and the observed lines. A good model atmosphere of the Sun is very important because it is the cornerstone of our understanding of the atmospheres of all other stars.

A. Radiative Transfer

Model atmosphere construction starts with a discussion of the equations of radiative transfer. Let us begin with some definitions (Figure 4). If we consider the

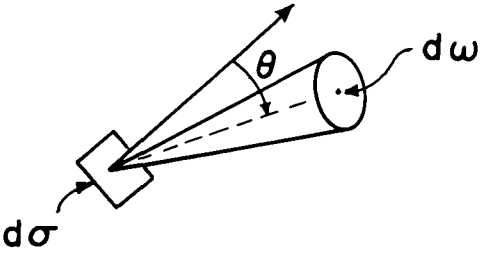


Figure 4.—Quantities used in the definition of intensity.

radiant energy dE_ν in a cone of solid angle $d\omega$ passing through a surface of area $d\sigma$ in time dt, and if axis of the cone makes an angle θ with the normal to the surface, the intensity I_ν is defined by

$$dE_\nu = I_\nu \cos\theta \, d\omega \, d\nu \, dt \, d\sigma \, . \tag{3}$$

We define a mass absorption coefficient K_ν by

$$dI_\nu = -K_\nu \rho I_\nu \, ds \, , \tag{4}$$

where dI_ν is the change in intensity of a pencil of radiation passing in the normal direction through a slab of material of density ρ and thickness ds (Figure 5).

Similarly, we may define an emission coefficient j_ν by

$$dI_\nu = j_\nu \rho \, ds \, . \tag{5}$$

Then, we may derive the equation of transfer from the change in intensity (Equations 4 and 5):

$$dI_\nu/ds = -\rho K_\nu I_\nu + j_\nu \rho \, . \tag{6}$$

We may also define the source function as

$$\mathcal{J}_\nu = j_\nu / K_\nu \tag{7}$$

and rewrite the equation of transfer (Equation 6) as

$$-dI_\nu / K_\nu \rho \, ds = I_\nu - \mathcal{J}_\nu \, . \tag{8}$$

If we further consider only problems of axial symmetry and plane parallel layers, we may introduce the optical thickness (Figure 6)

$$\tau_\nu = \int_z^\infty K_\nu \rho \, dz \, . \tag{9}$$

Thus, we may rewrite Equation 6 as

$$\mu \frac{dI_\nu}{d\tau_\nu} = I_\nu - \mathcal{J}_\nu \, , \tag{10}$$

where we have let $\mu = \cos \theta$. Note that Equation 10 is a first-order differential equation with a known solution. The solution for the radiation emerging from the top of an atmosphere is

$$I_\nu(0, +\mu) = \int_0^\infty \mathcal{J}_\nu(t, +\mu) e^{-t/\mu} \frac{dt}{\mu} \, . \tag{11}$$

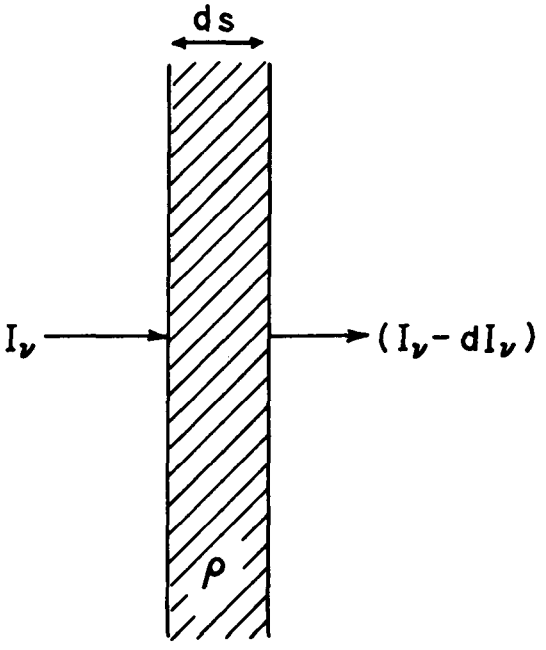


Figure 5.—Quantities used in the definition of the absorption coefficient.

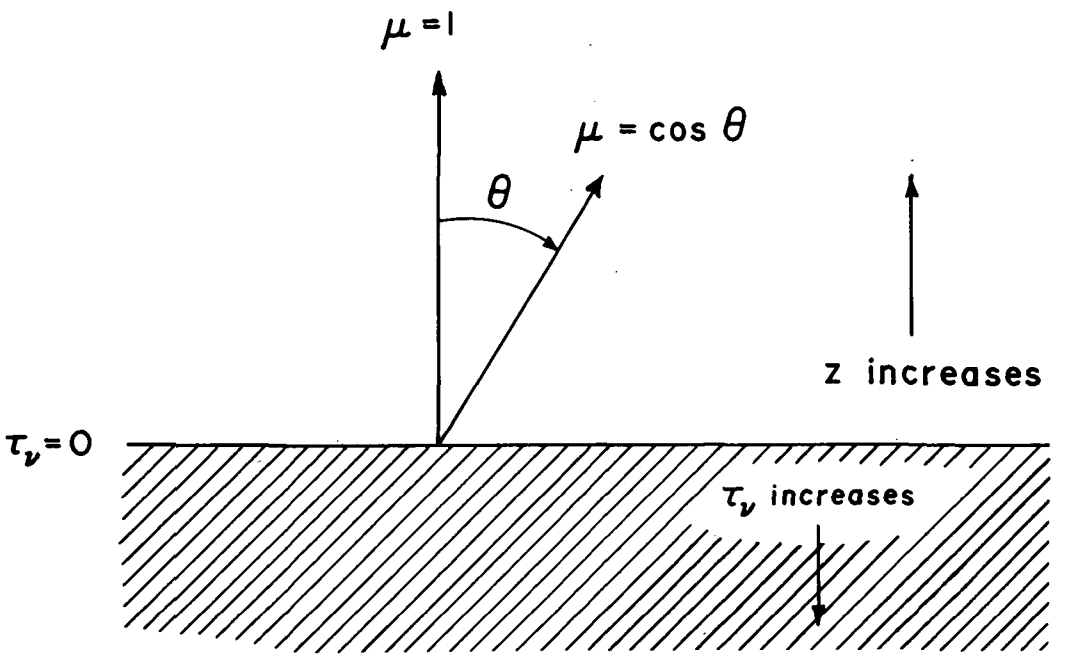


Figure 6.–Schematic showing directional quantities and the optical depth in a model atmosphere.

Equation 11 shows that the emergent radiation is simply the radiation emitted at each point decreased by the opacity between that point and the top of the atmosphere; note that a determination of $\mathcal{J}_\nu$ is a *complete solution* to the transfer problem. There are two cases of special interest:

(1) Local thermodynamic equilibrium. If at each point in an atmosphere we may write

$$j_\nu = K_\nu B_\nu(T)\,, \tag{12}$$

where $B_\nu(T)$ is given by Planck's law, then we may say the atmosphere is in local thermodynamic equilibrium (LTE). This assumption is really exact only if there are no anisotropies, i.e., reasonably deep in the interior. Since the wings of most absorption lines are formed sufficiently deep in the atmosphere, LTE is often a good assumption for this part of the line. For LTE we have the source function,

$$\mathcal{J}_\nu = B_\nu(T)\,. \tag{13}$$

(2) Scattered radiation. For this case, the emission coefficient is completely due to scattered radiation. Thus, for isotropic pure scattering we may write the source function in terms of the mean intensity J_ν:

$$\mathcal{J}_\nu = J_\nu = \frac{1}{4\pi}\int_{4\pi} I_\nu \, d\omega \,. \tag{14}$$

B. Particular Solutions

Consider an atmosphere that satisfies LTE. We may rewrite Equation 10, the equation of transfer, as

$$\mu \frac{dI_\nu}{d\tau_\nu} = I_\nu - \mathcal{J}_\nu = I_\nu - B_\nu(T) \,. \tag{15}$$

Thus, the determination of $T(z)$ determines the solution, through Equation 11.

If we then consider the photosphere to be in radiative equilibrium, and if we take a layer that has a thickness which is small compared to $R_\odot$, we may define the total upward flux as

$$\pi F = \pi \int_0^\infty F_\nu(z)\, d\nu = \text{constant} \,. \tag{16}$$

A constant net flux over all frequencies immediately implies

$$\int_0^\infty J_\nu K_\nu \, d\nu = \int_0^\infty B_\nu K_\nu \, d\nu \,. \tag{17}$$

In other words, every mass element must absorb as much energy as it emits.

C. Gray Atmosphere

We can now discuss a very idealized problem of interest. Consider an atmosphere where K_ν is independent of frequency; this is not a bad approximation for absorption by the H^- atom. This is called a gray atmosphere, and the optical thickness is given by

$$\tau = \int_z^\infty \bar{K}\rho \, dz \,, \tag{18}$$

where the method of determining the mean opacity $\overline{K}$ is left undetermined. By integrating the equation of transfer (Equation 8) over ν, we obtain

$$\mu \frac{dI}{d\tau} = I - B \,. \tag{19}$$

If the flux is constant, we find $B = J$, and the integrated intensity in a gray atmosphere is given by a solution of the constant-net-flux problem. The constant-net-flux problem is an old "war horse" of radiative transfer theory and has well-known exact solutions. The average intensity at the surface is

$$J(\tau = 0) = \frac{3^{1/2}}{4} F \,. \tag{20}$$

Since the integrated Planck intensity is related to the local temperature by

$$\pi B(T) = \pi \int_0^\infty B_\nu(T)\, d\nu = \sigma T^4 \,, \tag{21}$$

we find a relation between the effective temperature T_{eff} and the boundary temperature T_0:

$$T_0^4 = \frac{3^{1/2}}{4} T_{\text{eff}}^4 \,,$$

or

$$T_0 = 0.81 T_{\text{eff}} \,. \tag{22}$$

For the Sun, this gives us $T_0 \approx 4600$ K, which is fairly close to the minimum observed solar temperature between the photosphere and chromosphere. To determine the temperature distribution in a gray atmosphere, we utilize the source function for the constant-net-flux problem:

$$\mathcal{J}(\tau) = \frac{3}{4} F[\tau + q(\tau)] \,, \tag{23}$$

where $q(\tau) = 0.58$ at $\tau = 0$ and $q(\tau) = 0.71$ at $\tau \to \infty$. Thus, if we assume $q(\tau) \approx 2/3$, and if we use the Stefan-Boltzmann equation (Equation 21), we find

$$T^4 = \frac{1}{2} T_{\text{eff}}^4 \left(1 + \frac{3}{2}\tau\right) \,, \tag{24}$$

which is the temperature distribution in a gray atmosphere.

By using this relation and Equations 9 and 19, we find the distribution of brightness as a function of μ:

$$I(+\mu, \tau = 0) = \int_{t=0}^{\infty} B(t) e^{-t/\mu} \frac{dt}{\mu}$$

$$= \int_{t=0}^{\infty} B_0 \left(1 + \frac{3}{2} t\right) e^{-t/\mu} \frac{dt}{\mu}$$

$$= B_0 \left(1 + \frac{3}{2} \mu\right). \tag{25}$$

This produces a limb-darkening relation of the form

$$\frac{I(\mu)}{I(\mu = 1)} = 1 - u + u\mu, \tag{26}$$

where $u = 3/5$. The observational value is $u = 0.56$.

Limb darkening occurs because of the temperature gradient in the solar atmosphere. That is, an observation effectively penetrates to one unit of opacity; but because of the angle of observation, observations near the center of the disk are of deeper or hotter regions.

Thus, the simplest ideas concerning the transfer of radiation and the photospheric absorption coefficient are sufficient to produce some quantitative insight into the variation of temperature (or intensity) with optical depth in the photosphere.

D. Photospheric Structure

We have a relationship between the temperature and the optical thickness, but we need additional equations to determine the structure of the photosphere. The equation of hydrostatic equilibrium

$$dP = -\rho g \, dz \tag{27}$$

and the equation for the mean optical depth

$$d\tau = -\bar{K}\rho\, dz \tag{28}$$

can be combined to obtain

$$\frac{dP}{d\tau} = \frac{g}{\bar{K}}. \tag{29}$$

Since $\bar{K}$ is a function of temperature T, gas pressure P, electron temperature P_e, and composition A_i, we need to know these parameters everywhere. The composition is taken as known, and we already have a relation between T and τ. Furthermore, we can derive a relationship for P_e in terms of $P(T)$. Thus, we can write the mean opacity in terms of P and T. Finally, if we use the ideal-gas law as an equation of state, we can establish a relation between optical and geometrical depth by integrating Equation 29. This completes the logical specification of a model solar atmosphere.

E. Absorption Lines

The solar spectrum shows many absorption lines. The formation of the absorption lines in the photosphere is handled as is shown in Figure 7. We introduce pure absorption coefficients per unit mass K_ν and l_ν for the continuum and the line, respectively. Scattering coefficients i_ν (noncoherent) and σ_ν (coherent) are also introduced. The equation of transfer then becomes

$$\mu \frac{dI_\nu(x_\nu, \mu)}{-(K_\nu + l_\nu + i_\nu + \sigma_\nu)\,\rho\, dr} = I_\nu(x_\nu, \mu) - J_\nu(x_\nu, \mu)\,, \tag{30}$$

where the new, combined opacity is $dx_\nu = -(K_\nu + l_\nu + i_\nu + \sigma_\nu)\rho\, dr$. It is then necessary to specify the coefficients and the source function in order to solve the problem. Consider the case of no scattering, i.e., pure absorption. Then, $\sigma_\nu = i_\nu = 0$. If we assume the line is formed in LTE, then $\mathcal{J}_\nu = B_\nu$. Physically, the atoms absorbing at the line wavelength increase the opacity there, so we do not see as far into the atmosphere as we do in the continuum. Since the temperature of the photosphere increases with depth, we get a larger intensity in the neighboring continuum than in the line.

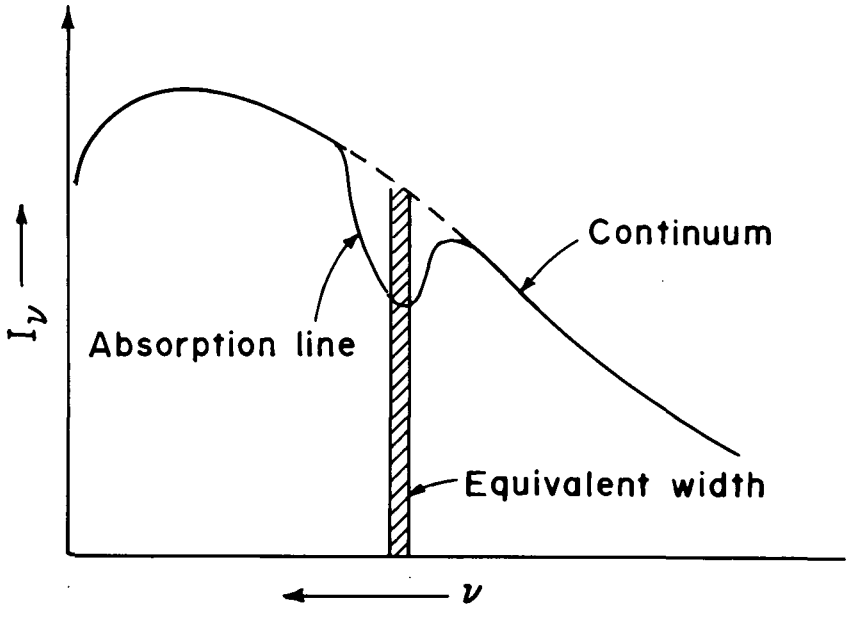


Figure 7.—Absorption line schematic.

A second case of interest is line formation by continuous absorption and line scattering. In this case, $i_\nu = l_\nu = 0$, and the source function is

$$\mathcal{J}_\nu = \frac{K_\nu}{K_\nu + \sigma_\nu} B_\nu + \frac{\sigma_\nu}{K_\nu + \sigma_\nu} J_\nu . \tag{31}$$

Solution of transfer equation with this source function leads to somewhat different predictions than with pure absorption. First, strong lines will be almost black at the center (residual intensity ≈ 0), whereas when LTE is assumed, the maximum line strength corresponds to a residual intensity given by $B_\nu(T_0)/B_\nu(T_{\text{eff}}) \approx 1/4$. Second, absorption lines formed by absorption and line scattering do not disappear at the limb, a conclusion in agreement with the observations. Absorption lines formed in LTE should disappear at the limb. In addition, the formation of lines by scattering in an absorbing atmosphere does not depend on a temperature gradient. Physically, the scattering process increases the total effective path length of a photon, thereby increasing the probability of absorption by H^- or some other absorbing constituent.

F. Abundances

For the Sun, one uses model atmospheres to determine theoretical equivalent widths for absorption lines. The theoretical equivalent widths are then compared with observed equivalent widths, by the use of the curve of growth method, to determine solar abundances. The principal uncertainties in the abundance determinations appear to arise not from the photospheric models but from the lack of reliable atomic parameters, which are needed to determine the curve of growth accurately.

A case in point is the recent redetermination of the solar Fe abundance, which had been the cause of some controversy. The coronal Fe abundance apparently was an order of magnitude higher than the photospheric determination. A recent reevaluation of the Fe-I f-value showed that the previous determinations had been in error and that the photospheric and coronal values are now in much closer agreement. In general, f-values are usually uncertain by at least a factor of 2, and this leads to similar abundance uncertainties.

V. CHROMOSPHERE

The chromosphere is a most difficult and perplexing region of the Sun. It is as much as 10^4 km thick and lies just above the photosphere. The temperature runs from 5000 K at the bottom to 10^6 K at the corona-chromosphere boundary. The chromosphere is very inhomogenous in every variable, including time.

Most chromospheric information is provided by observations of the flash spectrum during a solar eclipse when, as is shown in Figure 8, the Moon occults the chromosphere. The raw data consist of measurements on a particular atomic line or on many lines as the Moon's limb moves across the chromosphere.

Interpretation of the flash-spectrum observations have shown that the chromospheric temperature rises from about 4300 K at the lower boundary to about 6000 K at an altitude of some few thousand kilometers. At this distance, we enter a region called the transition zone, in which the temperature rises to the coronal value at the top. It is worth mentioning that because of the great difficulty in studying this zone, the width of the transition zone is not well established; however, it seems to become smaller with better data.

The chromosphere is best studied in Hα light up to about 4000 km. In the lower part of the chromosphere region, the gas can be treated as approximately homogeneous. From this lower layer emerge the brilliant streamers called spicules,

which are a property of the quiet Sun. They extend some 10 000 km above the limb and have a lifetime of about 5 minutes. They rise and seem to fall along the same path, indicating they are probably aligned with magnetic fields. It is likely that the spicules are rising material from the lower chromosphere; but as is typical of what is known of chromosphere, this fact is not well established.

In this very inhomogeneous region of the Sun, we can find CN lines and Fe-XI lines appearing simultaneously. This indicates how important the fine structure must be to an adequate understanding of the chromosphere. The fine structure manifests itself in the bright chromospheric network. These bright regions are similar to plages, and it is likely that the network arises from the breaking up of plages.

The upper chromosphere may be heated primarily by conduction down from the corona, whereas the lower chromosphere is probably heated by mechanical energy. The source of heat is very difficult to analyze, because we do not really know the shape of the boundary at the corona. Two alternative boundaries have recently been suggested and are shown in Figure 9.

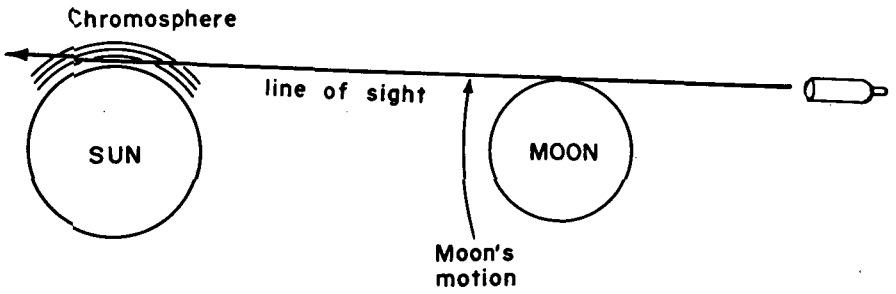


Figure 8.–Geometry of a solar eclipse.

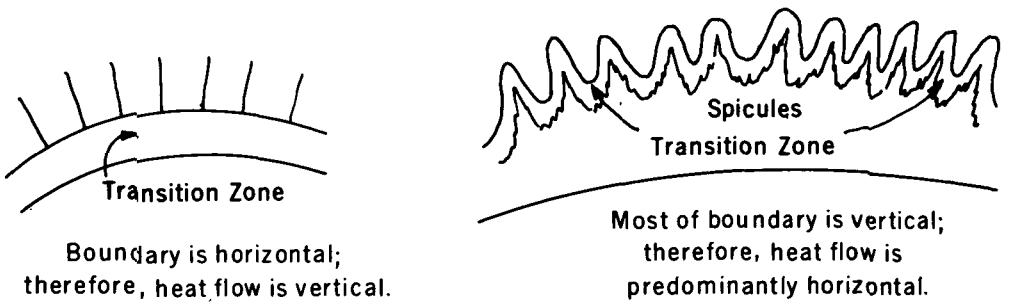


Figure 9.–Two suggestions for the shape of the transition zone.

VI. CORONA

The solar atmosphere above the chromosphere ($r \geqslant 1.03R_\odot$) is called the corona. The light of the corona is generally divided into three components:

(1) The K corona, which is the continuum due to electron (or Thomson) scattering of photospheric light.

(2) The E (or emission) corona, which is the total light of the coronal emission lines.

(3) The F corona, which is solar radiation diffracted by interplanetary dust. The K (electron-scattering) component is the most important one for determination of large scale structure, and we shall examine it more closely next.

A. Densities

One tries to determine the coronal electron density from the intensity of light from the corona. The difficulty is that we have to determine a three-dimensional distribution from two-dimensional observations. Generally, one is forced to assume a spherically symmetric distribution because almost any other assumption is too difficult to work from or too arbitrary. Another problem is that the absolute photometry is usually accurate only to a factor of 2 because of the difficulties of doing photometry on extended areas.

A relationship for the intensity can be derived (assuming spherical symmetry) as a power series:

$$I \sim \sum_n a_n / r^n \,. \tag{32}$$

From this, another power series, also in r, can be written for the electron density. In practice, we find that only two or three terms are sufficient. In particular, we find $N_e \approx 10^8$ cm^{-3} in the lower corona, whereas $N_e \approx 10^6$ cm^{-3} at $2R_\odot$.

If we look at the corona during different times in the solar sunspot cycle, we observe structural changes as shown in Figure 10. At solar maximum the corona is almost spherically symmetrical, and it is appreciably flattened at sunspot minimum. Polar streamers are also quite apparent during sunspot minumum. At intermediate stages, fans or large streamers approximately $1R_\odot$ long or greater can be seen as the principal coronal features.

From the observations cited above, it can be seen that the assumption of spherical symmetry is certainly not justified by observation. In particular, many observers report electron densities separately for the poles and for the equator. In the method generally used, an equatorial density is calculated and used to correct the high-latitude polar data. Figure 11 shows why this is necessary. The line of sight to the polar region passes through much of the corona at lower altitudes, so it is necessary to know the electron densities there to correct the so-called "polar" observations. Some eclipse observations are consistent with there being no electrons between the latitudes of 70 and 90 degrees. Thus, the polar electron data can be considered to be in question.

Before considering temperature determinations for the corona, let us briefly discuss the fan structures in the corona. Fans are the large streamers that determine the general form of the corona at any given time. Fans are usually associated with quiescent prominences, as shown in Figure 12. The base of the structure probably is

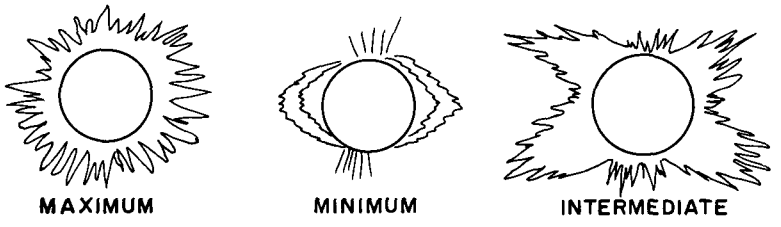


Figure 10.–Schematic showing coronal structure during different parts of the solar cycle.

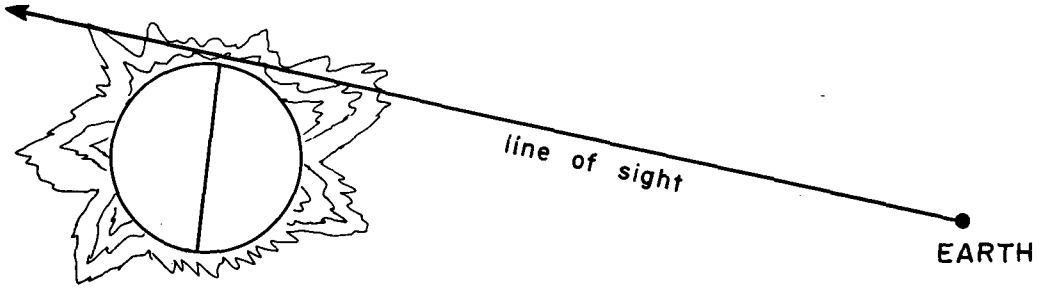


Figure 11.–Schematic line of sight through the polar regions.

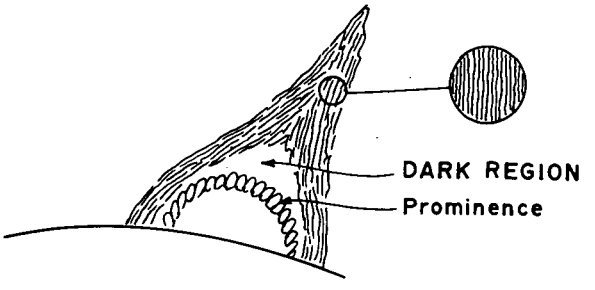


Figure 12.–Schematic showing the relationship between quiescent prominences and coronal fans.

comparable in width and length, but it quickly tapers into a two-dimensional vertical structure. The fan is composed of thin streamers, as shown in the magnified view. The prominence is probably formed by material condensing out of the fan.

B. Temperatures

The lines present in the emission spectrum of the corona are quite different from the lines of the Fraunhofer spectrum. They are primarily forbidden lines and arise from highly ionized atoms in a low-density medium. The three important lines are given in Table 1.

If we assume $N(1 - x)$ ions per unit volume in the stage p of ionization and Nx in stage $p + 1$, we may write

$$\frac{x}{1-x} = \frac{C(p \rightarrow p+1)}{R(p+1 \rightarrow p)} = f(T, \text{atomic parameters}) \tag{33}$$

for a steady-state condition, where C and R are the ionization rate and recombination rate, respectively. Note that the ratio does not depend on the electron density. The most important recombination process for the corona is dielectronic recombination. In this process, one of the atom's electrons is first excited; then recombination takes place, leaving two excited electrons. This process turns out to be more efficient than simple photo-recombination. Temperature determinations based on these results yield an average coronal temperature of 2×10^6 K.

Table 1.—Some lines in the corona with ionization potentials for the corresponding ion.

λ (Å)	Ion	Ionization Potential (eV)
5303	Fe-XIV	355
5694	Ca-XV	820
6375	Fe-X	235

The width of the coronal emission lines can also be used to determine temperature. If we assume the line is broadened only by the thermal Doppler effect, we may write an expression for the line profile:

$$I = I_0 \exp[-(\lambda - \lambda_0)^2/(\delta\lambda_0)^2], \tag{34}$$

where

$$\delta\lambda_0 = (\lambda/c)(2kT/\mu M_H)^{1/2}. \tag{35}$$

If we call the full width at half intensity h, then $h = 1.67\delta\lambda_0$ and

$$T = \frac{h^2}{\lambda^2}\mu \times 1.95 \times 10^{12}. \tag{36}$$

For the Fe-X redline, $\mu = 55.85$, $\lambda = 6375$Å, $h = 0.89$Å, and $T = 2.1 \times 10^6$ K. Actually, a temperature range of 1.2×10^6 K $< T < 4.5 \times 10^6$ K is observed using this method, indicating that the fine structure is present. The hottest temperatures are usually found over flaring regions.

When line ratios are used to determine temperature, caution must be used because there is no reason to believe both lines are fomed in the same region. Only if two lines of the same element can be found that differ by one ionization state should the method be attempted.

In principle, one can look at the thermal broadening of photospheric Fraunhofer lines to determine the coronal temperature. Here, we use the Doppler broadening equation (Equation 36), but with $\mu = 1/1836$, the electron mass. The best solar lines are the H and K Ca lines; but at 2×10^6 K, $h = 170$Å. This large width produces an exceedingly shallow and hard-to-measure depression. Thus, we can set only a limit of $T > 10^5$ K by this method.

We can also try to determine the coronal temperature from structural considerations. If we assume that the corona is isothermal, spherically symmetrical, and in hydrostatic equilibrium, its density distribution follows the generalized barometric relation

$$\frac{N_e}{N_{e,0}} = \exp\left[\frac{GM_\odot \mu M_H}{R_\odot kT}\left(\frac{1}{r} - \frac{1}{r_0}\right)\right]. \tag{37}$$

By taking logarithms, differentiating, and solving for T, we find

$$T = \frac{1.004 \times 10^7 \mu}{d(\log_{10} N_e)/d(1/r)}. \tag{38}$$

If we assume He:H = 1:10, then $\mu = 0.608$. The observed function $\log N_e(r)$ implies a temperature of 1.5×10^6 K, which apparently is a little low. The problem is probably caused by the fact we assume a spherical distribution to determine $N_e(r)$, whereas in fact the matter is somewhat confined to rays. Thus, the observed line-of-sight decrease in column electron density is caused both by a real decrease in N_e and a decrease in the fractional volume of the rays. When we take this into account, the tendency is for a slower decrease in N_e and for higher temperatures (e.g., close to 2×10^6 K).

C. Origin

The source of these high temperatures still remains in question. The convective motions of the hydrogen convection zone no doubt generate acoustic waves, or noise, which carry mechanical energy to the transition zone, where it is dissipated. This process produces high temperatures because of the low densities present in the corona. The most efficient way to lose this energy would be for the hydrogen to radiate it away; however, the hydrogen is completely ionized, so we must find other processes. One of these is conduction back into the chromosphere; the other is the expansion of the solar wind.

ACKNOWLEDGMENT

The author is indebted to Dr. Stuart Jordan for his comments on this presentation.

CHAPTER 2
INTERNAL ROTATION OF THE SUN*

R. H. Dicke
Joseph Henry Laboratories
Princeton University, Princeton, New Jersey

I. INTRODUCTION

The question of whether rotation occurs in the deep solar interior is a controversial subject of considerable importance to relativity theory as well as to solar physics. The survival of Einstein's general relativistic theory of gravitation or the establishment of the scalar-tensor theory may hinge upon the absence or presence, respectively, of rapid rotation in the deep solar interior. If a rapidly rotating solar core exists, with a rotational period under 2 days, it could be the source of angular momentum supplied to the solar wind, of internal mixing through the transport of matter along with angular momentum, and perhaps of solar activity leading to the sunspot cycle.

In prerelativity days, the observed excess motion of Mercury's perihelion ($43'' \pm 0.4''$ per century) (Clemence, 1943; Duncombe, 1958; Wayman, 1966) was a mystery and led to several unsuccessful attempts to find a perturbation that could account for the effect (Leverrier, 1859; Chazy, 1928). The suggested sources include interplanetary material, Vulcan (a hypothetical and still undiscovered planet), and the flattened mass distribution of an oblate Sun (Newcomb, 1897).

All of these suggestions must now be discarded. The interplanetary dirt and spare planet have not appeared, and a solar gravitational quadrupole moment large enough to generate the full excess centennial motion of $43''$ in Mercury's perihelion would also cause a $43''$ regression of the node on the plane defined by the Sun's equator, far too large to be allowed.

The urge to find a prosaic source for the excess motion of Mercury's perihelion disappeared with the appearance of Einstein's general relativistic theory of gravitation. This accounted for the full $43''$ motion as a relativistic effect; but with the recent increased interest in the scalar-tensor theory of gravitation (Jordan, 1948, 1959; Thirry, 1948; Bergmann, 1948; Brans and Dicke, 1961; Dicke, 1962), the

*A similar version of this paper appeared in *Annual Review of Astronomy and Astrophysics* 8:297-328, 1970.

question of a possible nonrelativistic origin for part of the 43″ motion is of interest. The scalar-tensor theory is a general relativistic theory for which the relativistic perihelion rotation of Mercury's orbit is

$$\frac{3\omega + 4}{3\omega + 6} \times 43'' \text{ per century}, \tag{1}$$

where ω is the coupling constant of the Brans-Dicke (1961) form of the theory. This constant ω has been estimated on various grounds to fall in the range $4 < \omega < 7$ (Brans and Dicke, 1961; Dicke and Peebles, 1965; Dicke, 1966).

If the "observed" excess motion, calculated from planetary perturbations only, is as accurate as claimed [43.0″ ± 0.4″ (see Duncombe, 1958)], Einstein's theory is favored; but some additional perturbation, such as that from a flattened Sun, generating a motion of 4″ per century would favor the scalar-tensor theory with $\omega \approx 5$.

With these facts in mind, the author suggested (Dicke, 1964) that the Sun may have a distorted interior induced by a rapidly rotating core, the fossil remnant of the rotation of the young Sun. (Rapid internal rotation was also discussed by Roxburgh, 1964; Plaskett, 1965; Deutsch, 1967.) It was assumed that in the density-stratified interior, below the convective zone, quasi-stable rotation of a core was possible with angular momentum diffusing to a thin shell of instability lying below the convective zone. This shell and the outer convective zone were assumed to have been rapidly rotating initially, but subsequently slowed by a solar-wind torque. In collaboration with P. J. E. Peebles, the author formulated a theory of the solar-wind torque along lines similar to the ideas of Schatzman (1959) and Cowling (1965) and used the resulting formulas in the 1964 paper to estimate the solar-wind torque. Equivalent formulas were independently derived by Modiesette (1967), Weber and Davis (1967), and Alfonso-Faus (1967).

The estimated torque (4×10^{30} dyne-cm) was based on early and preliminary measurements of the solar-wind flux and an estimated strength at the Sun of the magnetic field drawn out by the solar wind ($\approx$3/4 G). Later, when observations were available, this field strength was found to be in reasonable agreement with observations of the field in the solar wind at the Earth, if a purely radial flow of the solar wind was assumed. It was also shown in the 1964 paper that the estimated solar-wind torque was in reasonable agreement with that derived from an approximate solution to the diffusion equation applied to the diffusion of angular momentum from the core to the outer slowly rotating shell.

One interesting aspect of stellar-wind braking concerns the old problem posed by the apparent break in the rotational distribution at spectral type F5. Stars much bluer than F5 are rapid rotators, and old stars much redder than F5 are slow rotators. Schatzman (1959, 1962) pointed out that the transition between stars with deep radiative envelopes and those with deep subsurface convective zones occurs among the early F-type stars, and he developed a theory of stellar braking using magnetic fields derived from jets or flares associated with such subsurface convective zones. Kraft (1967) noted that the connection between such subsurface convection and a stellar wind could explain why stellar-wind braking of young stars is limited to stars redder than F4.

It was noted in the 1964 paper that a solar gravitational quadrupole moment large enough to induce a centennial motion of 4″ in Mercury's perihelion (needed for $\omega = 5$) would also induce an oblateness [$\Delta r/r = (r_{eq} - r_p)/r$] of 5×10^{-5} in the Sun's atmosphere. This oblateness would be in addition to the 1×10^{-5} due to surface rotation. It was also noted that the 4″ regression of the node on the solar equator expected from such a distorted Sun, when referred to the plane of the ecliptic, represents principally a centennial decrease of the inclination (0.21″). The observed residual in the rate of increase of the inclination (−0.12″ ± 0.18″; Clemence, 1943) is to be compared with the above (−0.21″). (See later discussions of this question by Shapiro, 1965; Audretsch et al., 1967; Gilvarry and Sturrock, 1967; and O'Connell, 1968.)

In the spring of 1963, H. Hill, H. M. Goldenberg, and the author designed and built a new type of instrument designed to measure the solar oblateness (Figure 1). This was installed in its own observatory at Princeton and put into operation in the late summer of that year. The summers of 1964 and 1965 were used for studying and improving the instrument. The first high-quality measurements were made in the summer of 1966 and published in an abbreviated form (Dicke and Goldenberg, 1967a). A full treatment of these measurements will soon be given. These measurements showed the Sun to have the oblateness ($5 \pm 0.7 \times 10^{-5}$) expected if it possesses such a rapidly rotating core. Associated with an oblateness of 5×10^{-5} (if our interpretation is correct) is a perihelion rotation of 3.2″, making the "observed" relativistic rotation 39.8″ ± 0.4″ consistent with the value 38.9″ expected under the scalar-tensor theory with $\omega = 5$. Unpublished measurements during the summer of 1967 gave the same oblateness with comparable precision.

The publication of these preliminary results was followed by a rash of criticisms, comments, and reinterpretations. Öpik (1967) and Ashbrook (1967) suggested that our observations may not have been as accurate as we thought. (See

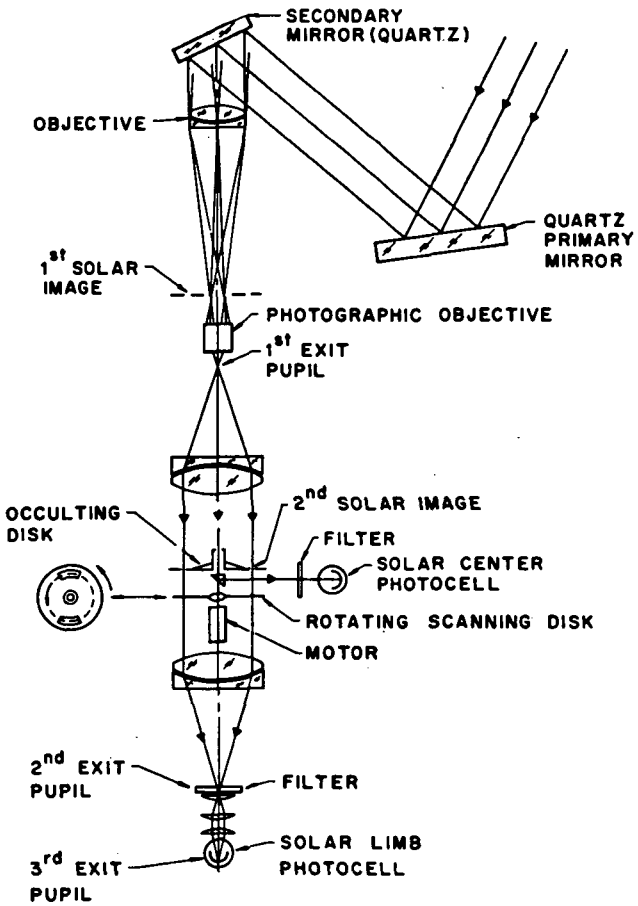


Figure 1.—The optical system of the solar-oblateness telescope. An image of the Sun is projected on a stationary occulting disk that stops all the light except that from the outer 6.5″, 12.9″, and 19.2″. Light passing the occulting disk is "chopped" at ≈ 100 cycles per second by a spinning disk. Light from the center of the Sun's disk provides a normalization signal. Calibration is carried out by replacing the circular occulting disk by one with a stepped edge.

Dicke, 1967a and 1967b, for replies to these comments.) Roxburgh (1967a, 1967b), Cocke (1967a), Sturrock and Gilvarry (1967), and Durney and Roxburgh (1969) suggested that the excess solar oblateness did not imply a gravitational quadrupole moment. (See Dicke and Goldenberg, 1967b, and Dicke, 1970a, for discussions of these suggestions.) Howard et al. (1967) and Goldreich and Schubert (1967a,

1967b) suggested that a rapidly rotating core was impossible because of spin-down by Ekman pumping or a thermally driven instability, respectively. (For comments on these papers see Dicke, 1967c, 1967d; McDonald and Dicke, 1967; Colgate, 1968; and Clark et al., 1969.) The instability argument of Goldreich and Schubert depends upon a model of the interior and is no more certain than the model.

In general, observations of the solar surface which have a bearing on the existence of a rapidly rotating core are more valuable than detailed calculations on the unknown solar interior. The important observations concern the solar oblateness; the velocity fields, including rotation in the "seen layers" of the Sun; the magnetic fields in the "seen layers"; the abundance of lithium and beryllium; and the structure of the solar wind. In the absence of magnetic and velocity fields at the solar surface, the oblateness yields the gravitational quadrupole moment unambiguously. It must be emphasized that magnetic and velocity fields *must* be in the "seen layers" of the Sun, hence observable, if they are to affect this relationship. This will be discussed later.

The history of the solar system casts some light on the solar rotation problem. Although the Sun's past may seem more hidden than its interior, the assumption that it is a typical main sequence star of 1 solar mass means that observations of young solar-type stars are capable of showing the appearance of the Sun at the same age. We make this assumption.

Kraft (1967) has shown that stars of 1.2 solar masses in the Pleiades have surfaces rotating with $\langle V \rangle \approx 40$ km-s^{-1} with the same angular velocity as that of the postulated rapidly rotating solar core. The extrapolation of the observation to G2 spectral-type stars in the Pleiades indicates that these solar-type stars may have an average surface velocity of ≈ 10 km-s^{-1}. This represents an angular velocity only one-fourth that needed in the solar core. With the assumption that such stars possess solar-type rapidly rotating cores, stars redder than F5 probably arrive on the main sequence rotating differentially. For the Hyades, approximately 5×10^8 years older, the rotation has decreased by a factor of 2. For very old stars, the rotation is almost imperceptible, as shown in Figure 2, based on Figure 17 of Kraft (1968). These observations suggest that the Sun was originally rotating with a 1- to 2-day period and that the solar wind has slowed the rotation of either the whole Sun or an outer shell.

A compelling argument in support of the contention that only an outer shell is slowed is provided by observations by Herbig (1965) and others of the abundances of lithium and beryllium in solar-type stars of various ages. (See survey article by Wallerstein and Conti, 1969, for references.) Apparently, solar-type stars arrive on the main sequence with abundances of both lithium and beryllium that are

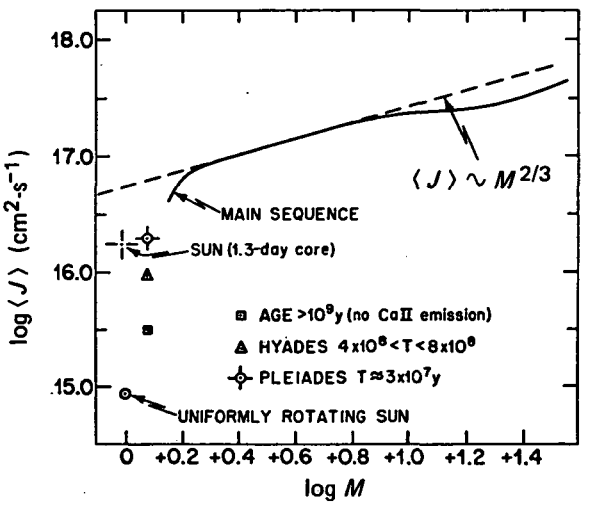


Figure 2.—The logarithm of angular momentum per unit mass of stars various ages and masses versus log M; rigid rotation is assumed. For stars bluer than F_0 (more massive than log $M = 0.2$) the surface rotation is independent of age. This figure is based on Figure 17 of Kraft (1968).

characteristic of chondritic meteorites. The lithium becomes depleted as the star ages, with a mean life of 7×10^8 years (Danziger, 1969), but the beryllium does *not* become depleted. Danziger (1967) has noted that the e-folding times for the decrease of lithium and of the rotational velocity are equal for young solar-type stars, which suggests that the two processes may be related.

The observations seem to imply that solar-type stars are mixed as deep as $r = 0.6R_\odot$ but not as deep as $r = 0.5R_\odot$ where beryllium is burned, but some type of mixing seems to be required if angular momentum is to be removed from the deep interior by the transport of material, for angular momentum will diffuse only to about $r = 0.05R_\odot$ during the life of the Sun.

Goldreich and Schubert (1967a) used this argument to conclude from the presence of lithium and beryllium in the Sun that the Sun arrived on the main sequence slowly rotating. If their argument were valid, it would imply that the late F-type stars and early G-type stars in the Pleiades must be slow rotators, for the old stars of these types are invariably slow rotators and they contain the normal amount of beryllium. Apparently, the most reasonable interpretation to make of the observations is that the solar-wind torque slows only an outer shell (Dicke, 1964) approximately $0.4R_\odot$ thick (Dicke, 1970c).

Eighty-four percent of the moment of inertia of the Sun falls inside the radius $r = 0.6R_\odot$ (see Table 1). Hence, if the solar wind slows only an outer shell, in thickness only $0.4R_\odot$, the total angular momentum is substantially unaffected, and there is no great difference between the angular momentum per unit mass of solar-type stars and more massive stars. An enigma in old theories of the origin of the solar system has disappeared. These theories could not explain the slow rotation of the solar-type stars. The approximate equality of the angular momentum of the planetary system with the assumed initial total angular momentum of the solar system, an apparently fortuitous relation, compounded the difficulty, for it seemed necessary to find a mechanism to concentrate the angular momentum in the planets. Hoyle's (1960) theory of the solar system provides an effective means of transferring angular momentum out of the proto-Sun through magnetic torque; but the observations of rapid rotation in young solar-type stars show that these stars initially have a great deal of angular momentum. Most of this remains in the deep interior if our picture is correct.

A difficult question concerns an internal magnetic field. One might expect a strong magnetic field to be trapped in the interior of the proto-Sun; but with the alternating sunspot cycle, a 22-year period, the present solar field appears

Table 1.—The slowing of rotation of a shell by a solar wind of equatorial torque density 10^8 dyne-cm^{-1}.

r_c	I_s/I	T (rigid rotation) (10^9 years)	T (10^9 years)
0.86	0.0122	0.173	0.173
0.78	0.034	0.48	0.51
0.70	0.074	1.04	1.28
0.62	0.140	1.98	2.88
0.54	0.241	3.40	6.05

Note: Decay time in years and fractional moment of inertia as functions of the inner radius of a uniformly rotating outer solar shell; it is assumed that only the shell is slowed by an equatorial solar-wind torque density of 10^8 dyne-cm^{-1}. The decay time for the whole Sun rotating rigidly is 14.1×10^9 years. In the last column the decay time is calculated for a rigidly rotating convective zone and a differentially rotating inner zone, with $\omega = (r_v/r)^2\, \omega_s$ for $r_c < r < r_v = 0.85R_\odot$.

superficially to have its origin in the convective zone. It has been suggested that convective mixing during the Hayashi phase might convert the long-lived low magnetic modes to short-lived high modes that would decay after the cessation of convection in the interior but persist in the outer convective zone (Dicke, 1964; Cowling, 1965). To eliminate low modes by this mechanism sufficiently to permit internal differential rotation no longer seems feasible to the author. At least two other possibilities remain. The internal magnetic field might be eliminated by a type of "beer foam process". If the internal field should become highly contorted during an initial convective phase, the radiative core might grow outward from the center by filling with field-free gas flowing along substantially field-free canals. Being free of the magnetic pressure, this field-free gas would tend to be denser than its surroundings and settle to the center.

A second possibility seems more likely. It will be the subject of a future publication and is mentioned here only because of its relation to a rapidly rotating core. It is possible that the Sun arrives on the main sequence with its magnetic field oriented perpendicular to the rotation axis and cut off from the convective zone by a shell of differential rotation. Such an orientation might be expected because of an instability associated with the Hoyle (1960) magnetic braking of the proto-Sun. For simplicity, assume that the magnetic field is trapped in a dipolar configuration and links the central condensation with the outer solar nebula. The rapidly rotating proto-Sun is flattened because of the tension (ρv^2 in the direction of fluid motion). As the energy of rotation is converted into toroidal magnetic energy, the negative motional tension is converted into the positive magnetic tension $B^2/4\pi$. When roughly half of the kinetic energy is lost, the proto-Sun becomes prolate and is probably unstable. It should then precess to the quasi-stable perpendicular position. This position may remain stable after the decay of the toroidal field. With a magnetic field in this perpendicular configuration, it penetrates only a few kilometers into the shell of differential rotation. Magnetic A-type stars might be exceptions where, for some reason, the shielding by differential rotation has not appeared, or has been lost, exposing the strong internal field at the star's surfaces.

The rapidly rotating core containing a magnetic field in the perpendicular orientation is capable of a torsional oscillation of high Q for which the north and south magnetic poles oscillate back and forth between the northern and southern hemispheres of the rapidly rotating core, and toroidal fields of opposite sign and alternating polarity are generated in the two hemispheres. Magnetic buoyancy might cause this toroidal field to float up to the convective zone, providing an explanation of the sunspot cycle as an effect of this oscillation. With reasonable magnetic-field

strengths, the period of this oscillation can be made to 22 years. Owing to the stability of the frequency of the oscillating core, this model has implications that can be tested with observations of the sunspot cycle. These analyses have been carried out and will be reported elsewhere. (See Dicke, 1970c, for a brief discussion.) A torsional oscillation in the rapidly rotating core may have implications for the Goldreich-Schubert instability, for the θ-component of velocity can be as great as 1 m-s^{-1}, eliminating these slow-growing axial modes (Goldreich and Schubert, 1967a).

II. THE SOLAR OBLATENESS

In earlier days, the Sun's oblateness was measured photographically (see discussion by Schaub, 1938) and, more accurately, with the heliometer, a telescope with a split objective (e.g., the work of Schur and Ambronn, 1895, at the Göttingen Observatory). Opposite limbs of the two solar images could be brought into contact by adjusting the two halves of the objective, whose separation provides a measure of the corresponding diameter.

Several possible difficulties with these measurements are apparent. The solar limb frequently has a width of 3″ to 5″ when the Sun is high in the sky, where it should be if atmospheric refraction is to be manageable, but the anticipated oblateness represents a difference between equatorial and polar radii of only 0.05″, 1 percent of the "seeing" width. Owing to an expected small anisotropy in the seeing disk associated with anisotropy of the turbulence near the ground, the northern and southern limbs of the Sun might be more (or less) diffuse than the eastern and western limbs near the meridian, causing a systematic error. Furthermore, problems of personal bias are very difficult when the measured effects are so small relative to seeing widths. Also, the heliometer may not have been free of systematic errors associated with gravitational distortion. These instruments required a 90-deg rotation of the objective system in the Earth's gravitational field. Any gravitational distortion would change with such a rotation.

The instrument designed by H. Hill, H. M. Goldenberg, and the author incorporated a number of improvements. This system is shown in Figure 1. Instead of measuring the position of the solar limb, the light flux was integrated from the edge of an occulting disk, a position near the limb, outward beyond the limb to an aperture stop a few tens of seconds of arc beyond the limb. Anisotropic seeing induced near the ground would be expected to spread the light but not change the

flux. Anisotropy can still introduce some error because of the gradient in limb darkening at the edge of the occulting disk, but the effects are much reduced. The telescope was vertically mounted to avoid a change in gravitational distortion with rotation.

The problem of separating the signal from the noise was solved by measuring photoelectrically. A rapidly spinning wheel perforated by two apertures of different sizes at the ends of a diameter scanned the light flux passing the occulting disk. The photoelectric signals were analyzed electronically in an impersonal way to measure the amplitudes of the sine and cosine terms of the second harmonic of the rotation frequency of the wheel. One-min averages were recorded on a punched magnetic tape, and the vertical telescope was rotated through 90 deg between the 1-min runs. The results recorded on the punched tape were analyzed by a computer. The scanning wheel also provided an error signal fed back to the main mirror to servolock the Sun's disk to the occulting disk, causing the Sun's disk to be accurately centered.

During a typical day of 6 hours, the Sun's image rotated through 90 deg relative to the telescope. The sine and cosine amplitudes were combined linearly to give the north-south (or vertical) component of the oblateness $(\Delta r/r)$ cos $2P$ and the northeast-southwest (or diagonal) component $(\Delta r/r)$ sin $2P$, where P is the angular position of the Sun's rotational north pole measured eastward from the north point of the disk.

The two mirror cells were rotatable about the mirror axes, and the mirrors were cycled with a 2-day period through all four combinations of positions to permit the elimination of errors due to mirror astigmatism. The only astigmatic error not eliminated is the off-axis error associated with a slight curvature of the main mirror viewed obliquely. This error contributed to the vertical component of oblateness only; the diagonal component was unaffected. Except for the effect of this off-axis astigmatism, the instrument is believed to be free of significant systematic errors, and measurements of the diagonal component are believed to be reliable. The telescope aperture was about 6.35 cm and was stopped below 2″. The instrument probably has the largest ratio of pounds of electronics to pounds of telescope of any telescope in existence.

The instrument permitted checks for systematic errors, and there was an accurate and reliable means of calibration that was repeated several times during each day. In addition, several internal checks of the data were possible. The contribution from atmospheric refraction to the measured diagonal component of the oblateness is usually large, ranging for September 1 from -4×10^{-4} at 9 a.m. to

zero at noon to $+4 \times 10^{-4}$ at 3 p.m. On July 16, the corresponding values are -2×10^{-4}, 0, and $+2 \times 10^{-4}$. This refractive contribution to oblateness is large, but it can be computed from atmospheric conditions measured in the observatory. For a laminar atmosphere, it is independent of conditions above the ground. After subtracting the computed values, the residuals are observed to be constant through the day (except for a small time-varying residual late in the summer when the Sun was low in the sky). On July 16, the residual in the diagonal component, $(\Delta r/r)$ sin $2P$, is approximately 7×10^{-6}, or only 3 percent of the refractive effect. That this small residual should be constant during the day indicated that the instrument has been correctly calibrated.

The most important internal check of the data is based on the change in orientation of the Sun's axis through the summer season. On July 7, the axis is in the north-south direction and the diagonal component should be zero. Also, for any assumed constant oblateness along the rotation axis, the variation of the diagonal component with date (through the term sin $2P$) is predictable. For an oblique distortion axis the oblateness should vary with the solar-rotation period. This variation is not present to any marked degree. The change of the observed diagonal component with time through the summer of 1966 is shown in Figure 3. The curve is calculated with the assumption that the solar oblateness is equal to $\Delta r/r = (r_{eq} - r_p)/r = 5 \times 10^{-5}$. During 1967, the observational period was longer, but the weather was substantially worse. The same oblateness was obtained with comparable precision.

One interesting interpretation of the data of Figure 3 is based on least-squares fits of the curve shown in Figure 3 to data representing different amounts of exposed limb averaged in various ways. In permitting the curve to float up and down, the date for crossing the abscissa will vary, and this change in crossing date can be interpreted as equivalent to an angle between the rotation axis and oblateness axis. Based on 15 different analyses of the data, the average crossing date is July 5.4, ± 3.4 days, whereas it should be July 7. This corresponds to the oblateness axis leading the rotational axis by 0.7 ± 1.4 deg as they rotate together counterclockwise on the sky through an angle of 40 deg. It is difficult to believe that these results are fortuitous, that instrumental and atmospheric effects would so conspire as to yield the curve of Figure 3 with a crossing date differing only a few days from July 7.

Measurements were made with three different distances from the edge of the occulting disk to the limb. This permits a separation of a signal due to the variation of brightness with latitude from the oblateness signal. The oblateness signal is proportional to the brightness of the photosphere at the edge of the occulting disk,

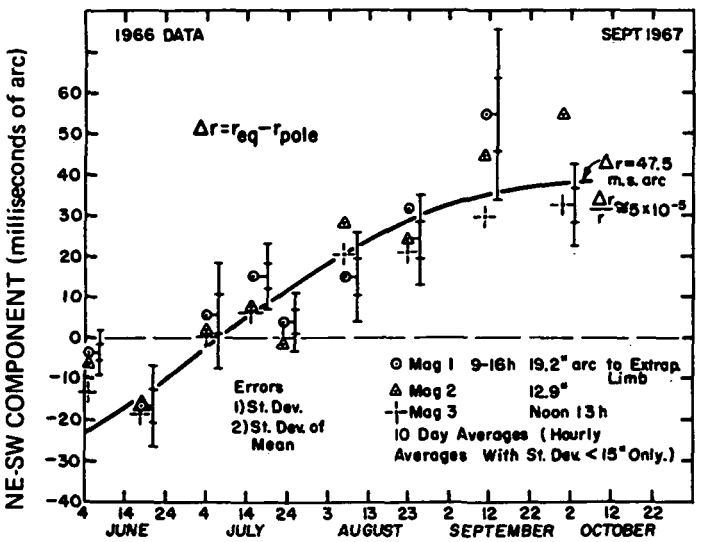


Figure 3.—The solar oblateness, diagonal component, for 1966. Observations are averages over a 7-hour day. Points are approximately 10-day averages at three different amounts of exposed limb. The curve was calculated by assuming $\Delta r/r = 5 \times 10^{-5}$.

but the intensity signal (associated with a variation with latitude of the photospheric brightness) is proportional to the integrated flux passing the occulting disk. For the edge of the occulting disk to be at the three distances 6.5", 12.8", and 19.1" from the Sun's limb, the photospheric brightness at the edge obtained from a limb-darkening curve is proportional respectively to 0.380, 0.400, and 0.432. The integrated flux is respectively 3.1, 5.4, and 8.34 (Dicke, 1970a). Measurements were made at these positions referred to the "extrapolated limb". With the measured values of edge brightness and light flux, the signals obtained at any two of the three positions are easily separated into the two parts.

Later measurements made with an annulus on the solar disk near the limb permitted a determination of oblateness from the limb-darkening effect and eliminated the chromosphere as a significant source of signal. The annulus technique was also used far from the limb to investigate the dependence of solar brightness upon latitude. None was found. Measurements of the solar oblateness were made with two broadband filters in the red and green; in 1967, these filters were

frequently switched. No systematic dependence of oblateness on color was found. A few oblateness measurements were also made with an Hα filter as a separate check on the contribution to the oblateness signal from the chromosphere and from chromospheric lines on the disk.

One frequently raised question about the solar oblateness concerns solar activity and the effect of active patches, faculae, and sunspots on the measurements. There are two aspects to this question: (1) Is the shape of the solar surface distorted by the solar activity a distortion not mirrored in the gravitational quadrupole moment, and (2) does an active patch at the Sun's limb adversely affect the measurement of solar oblateness? The first question will be discussed in the next section; the answer to the second question is yes. A large sunspot lying on the limb can induce a 20- to 30-percent error in the results for that day, but this happened only infrequently in 1966 and 1967, when the Sun was reasonably quiet. Furthermore, the error would be interpreted largely as an intensity signal, and the absence of a significant signal of this type provides an internal check for the insignificance of this effect. The systematic error in average oblateness from this effect is negative in sign (i.e., prolate) but is too small to be significant.

The conclusions from these observations are as follows:

(1) The values of the solar oblateness during 1966 and 1967 were equal, and $\Delta r/r = 5 \times 10^{-5}$.

(2) This photospheric oblateness was independent of color.

(3) The contributions from the chromosphere and corona to the oblateness were unimportant.

(4) The contribution from chromospheric lines on the solar disk was unimportant.

(5) During 1966 and 1967 the effective temperature of the photosphere was remarkably free of variation with latitude. There was no convincing stationary dependence on latitude (< 3 K).

III. SOLAR OBLATENESS AND THE GRAVITATIONAL QUADRUPOLE MOMENT

In this section it will be assumed that the solar oblateness has been measured to be 5×10^{-5}. The implication of this measurement for the existence of a quadrupole moment will be considered. This question has been discussed in detail (Dicke, 1970a).

The direct and umambiguous relation between the oblateness of the Sun and the solar gravitational quadrupole moment under certain conditions hinges upon a functional relationship used by von Zeipel (1924). As used here, it can be stated: *In the absence of magnetic and velocity fields in the "seen layers" of the Sun, surfaces of constant* P, ρ, T, *and* φ *(gravitational potential) coincide; i.e., pressure, density, and temperature can be considered to be functions of gravitational potential.* Also, as will be discussed below, for uniform rotation (or rotation on cylinders), these relations hold with φ replaced by φ plus a centrifugal potential. This functional relation was used by von Zeipel (1924) to derive his well-known paradox. Unlike von Zeipel, we are here interested in the validity of these relations only in a limited region of the Sun, the part actually seen. If only a part of the solar surface is free of magnetic and velocity fields, the functional relation is applicable to this part.

The proof of this relation is trivial. In the absence of magnetic and velocity fields,

$$0 = \nabla P + \rho \nabla \varphi \,. \tag{2}$$

Thus, the vector normal to a surface of constant P coincides with the normal to a surface of constant φ, which implies that two such surfaces coincide everywhere if they touch anywhere. Taking the curl of Equation 2 shows that surfaces of constant ρ and φ coincide. For a uniform composition, T is a function of P and ρ, and surfaces of constant P, ρ, T, and φ coincide; or P, ρ, and T are functions of φ. This result follows for any patch on the Sun's surface, simply connected or not, that is free of these fields.

Inasmuch as the gravitational potential has a simple, layered structure, the atmosphere must have the same layered structure in pressure, density, and temperature wherever the theorem applies. It has been verified that the effect on the shapes of surfaces of constant φ due to the presence of the gravitational quadrupole moment in question is so minor as to not affect noticeably this simple layered structure (Dicke, 1970a). Hence, the limb-darkening curve would be substantially independent of latitude. Furthermore, an analysis of the factors affecting the position of the solar limb shows that if such a quadrupole moment exists and von Zeipel's assumptions are valid, the position of the solar limb is determined by the position of a surface of constant density with an accuracy of approximately 3 m (Dicke, 1970a). The expected brightness at the limb under these conditions should be free of any noticeable dependence on latitude. In summary, when von Zeipel's relations are applicable, and with the oblateness of a surface of constant

gravitational potential at the Sun's surface being of the order of 10^{-5}, the location of the limb of the Sun is determined to high precision by a surface of constant density and, hence, gravitational potential.

In an expansion of the gravitational potential outside the Sun in spherical harmonics the gravitational quadrupole moment is determined by the second zonal harmonic, falling off inversely as the cube of the distance from the Sun's center. The oblateness of surface of constant gravitational potential at the Sun's surface is given by

$$(\Delta r/r)_g = \varphi_2/rg = \frac{3}{2}J_2 , \tag{3}$$

where

$$\varphi_2 = \frac{3}{2}J_2(GM_0/r^3)r_0^2$$

is the coefficient of the term $(1/3 - \cos^2 \theta)$ in the expansion. The limb temperature is expected to be remarkably uniform, and the variation of disk brightness with latitude should be less than 0.01 percent (Dicke, 1970a).

The limb occurs at an optical depth of about 0.004 and a density of approximately 2×10^{-8} g-cm^{-3}. One might think that a very strong magnetic or turbulent velocity field just below this layer would affect the oblateness of the limb; but as shown above, such is not the case. Similarly, one might think that the tension of the strong magnetic field of a sunspot would depress the level of the photosphere over a large area surrounding the sunspot; but as shown above, such a strong field cannot affect the height of the photosphere anywhere except at the location of the magnetic field. Also, the only significant effect of the sunspot on the measurement is induced by the darkening of the sunspot, not the change in level (Dicke, (1970b).

As was noted above, von Zeipel's functional relations include the effects of rotation whenever the rotation of the surface layers is on cylinders, i.e., the angular velocity ω is a function only of distance from the rotation axis. If, and only if, purely rotational motion is on cylinders, the "centrifugal force" term is derivable from a potential, and the inertial term can be added to Equation 2 by including a centrifugal potential. Then, the equation is still valid with φ replaced by

$$\Phi = \varphi - \int_0^{r \sin \theta} \omega r \sin \theta d(r \sin \theta) . \tag{4}$$

Including the effects of rotation of the Sun's surface, the oblateness of a surface of constant ρ is also the oblateness of a surface of constant Φ. From this oblateness, the contribution of the centrifugal potential (8×10^{-6}) must be subtracted before the oblateness of the surface of constant φ is obtained.

Only one thing more is needed for a complete description of the connection between the observed solar oblateness and the gravitational quadrupole moment. When magnetic and velocity fields exist in the "seen layers" of the solar surface, von Zeipel's relations are not exactly satisfied, and the effects of these fields must be included. Such fields may contribute generally different amounts to the oblateness of surfaces of constant density and pressure. The surface rotation discussed above is a special case for which these two contributions are equal and the temperature hence constant on the surface.

The surface fields usually induce different oblateness in density and pressure surfaces, and the contribution to the oblateness of a constant density surface is usually accompanied by a variation of brightness with latitude. Only for a carefully selected stress distribution in the surface layers is the surface brightness independent of latitude.

For an arbitrary distribution of magnetic and velocity surface stresses, these contributions to the oblateness of constant density and pressure surfaces are known (Dicke, 1970a). They are conveniently expressed in terms of integrals over the surface of a set of basic stresses. These are P_f, the field pressure; S_r, the radial shear; S_t, the transverse shear; and S_m, the meridional shear. The defining equations are

$$\begin{aligned} P_f &= \frac{1}{8\pi} B^2 + \frac{1}{3}\rho v^2 , \\ S_r &= -\frac{1}{4\pi}(2B_r^2 - B_\theta^2 - B_\varphi^2) + \rho(2v_r^2 - v_\theta^2 - v_\varphi^2) , \\ S_t &= -\frac{1}{8\pi}(B_\theta^2 - B_\varphi^2) + \frac{1}{2}\rho(v_\theta^2 - v_\varphi^2) , \end{aligned} \tag{5}$$

and

$$S_m = -\frac{1}{4\pi} B_r B_\theta + \rho v_r v_\theta .$$

In Legendre polynomial expansions of ρ and P over a surface of constant φ, $\delta\rho$, and δP, the coefficients of the second Legendre polynomial $(1/3 - \cos^2\theta)$ are given by

$$\delta P = \frac{45}{8}\int_0^{\pi}\left(P_f - \frac{1}{6}S_r\right)\left(\cos^2\theta - \frac{1}{3}\right)\sin\theta d\theta - \frac{15}{8}\int S_t \sin^3\theta d\theta$$

$$-\frac{15}{8}r^{-2}\frac{d}{dr}r^3\int S_m \cos\theta \sin^2\theta d\theta \tag{6}$$

and

$$g\delta\rho = \frac{45}{16}\frac{1}{r^2}\frac{d}{dr}r^2\int S_r\left(\cos^2\theta - \frac{1}{6}\right)\sin\theta d\theta + \frac{15}{8}\frac{d}{dr}\int S_t \sin^3\theta d\theta$$

$$+\frac{15}{8}\left(\frac{4}{r} + r^{-1}\frac{d^2}{dr^2}r^2\right)\int S_m \cos\theta \sin^2\theta d\theta\,. \tag{7}$$

The quantities δP and $\delta\rho$ can also be interpreted as contributions to the equatorial excesses of pressure and density on surfaces of constant φ. The contributions to the oblatenesses of surfaces of constant pressure and density induced by these fields are

$$(\Delta r/r)_P = \delta P/rg\rho \tag{8}$$

and

$$(\Delta r/r)_\rho = (\lambda_\rho/r)(\delta\rho/\rho)\,, \tag{9}$$

where λ_ρ is the density scale height and g is the gravitational acceleration. The temperature excess at the equator on a surface of constant density is

$$\delta T = -[(\Delta r/r)_P - (\Delta r/r)_\rho]\mu g r R^{-1}\,, \tag{10}$$

where μ is molecular weight and R is the gas constant.

The oblateness of the Sun is observed at the limb at an optical depth of approximately 0.004, but the whole of the Sun is observed to be remarkably free of a variation of brightness with latitude. This implies that any acceptable distribution of surface fields must generate equal oblateness in the observed surfaces of constant ρ and P, and this oblateness must be substantially independent of optical depth over the range of 0.004 to 1. Thus,

$$(\Delta r/r)_\rho = (\Delta r/r)_P = \text{constant} . \tag{11}$$

Equation 11, in turn, implies that such a distribution of surface fields induces a force per unit volume of the form $-\rho\nabla W$, where W is some scalar function of the polar angle θ and is only weakly dependent on r. A variation of W with latitude by an amount δW generates an outward displacement of a surface of constant ρ, P, and T by an amount

$$\delta r = -\delta W/g . \tag{12}$$

In order that the surface fields induce an oblateness $(\Delta r/r)_f$,

$$W = gr(\Delta r/r)_f\left(\cos^2\theta - \frac{1}{3}\right). \tag{13}$$

Equations 6 and 7 can be solved subject to the constraint of Equation 11 to give the three independent solutions. The first is

$$P_f - \frac{1}{6}S_r = \frac{3}{2}P_f = \rho W , \tag{14}$$

with W given by Equation 13 and with S_t and S_m zero. Note that these equations refer to second Legendre polynomials. Other nonzero terms could be present. The second solution is

$$S_t = -\frac{1}{2}\rho gr(\Delta r/r)_f \sin^2\theta , \tag{15}$$

with $P_t = S = S_m = 0$ and W given by Equation 13. The third is

$$S_m = \rho gr(\Delta r/r)_f(\lambda_\rho/r)\sin^2\theta , \tag{16}$$

with $P_f = S = S_t = 0$ and W given by Equation 13. Any linear combination of Equations 14, 15, and 16 are also suitable.

Roxburgh (1967a), Cocke (1967a), and Sturrock and Gilvarry (1967) suggested, respectively, that the observed excess oblateness was due to variation with radius of the angular velocity, meridional currents acting against turbulent viscous forces, and magnetic fields. Equations 6, 7, and 14 to 16 have been used to analyze these suggestions. Roxburgh (1967a) suggested that Coriolis forces acting on the convective zone induce a heat-transfer imbalance that in turn causes a strong increase of the angular velocity with radius. The stresses associated with convective transport can be ignored since they occur below the optical depth of 0.004 of the limb. At the limb, only the rotation is significant. From equations similar to Equations 6 and 7, it was shown that Roxburgh's assumed rotational distribution reduces, rather than increases, the oblateness, and a large variation of brightness with latitude is generated, contrary to the observations (Dicke and Goldenberg, 1967b). (See later rediscussion by Roxburgh, 1967b.) In similar fashion, through the use of Equations 14, 15, and 16 it was shown that meridional currents and magnetic fields strong enough to generate the excess oblateness without generating a variation of brightness with latitude were incompatible with the observations (Dicke, 1970a). It is concluded that the observed surface stresses do not generate the observed excess oblateness.

The great uniformity in surface brightness is probably no accident. Any inequality between the radiation rate at the surface and the rate of heat transport to the surface from the interior would generate circulation currents that would transport magnetic fields over the Sun's surface to redistribute surface stresses. Probably, surface stress distributions are automatically adjusted by this feedback mechanism until the stress distribution is sufficiently uniform for the surface to be uniformly bright (except when sunspots appear where the magnetic field is strong enough to inhibit the convective transport of heat from below). Whatever the mechanism, the observations show the random-velocity field and the background magnetic field to be independent of latitude. For the weak-background magnetic fields of the quiet Sun, this uniformity is strikingly shown by Livingston's (1966) pictures. An analysis of one of his pictures shows no systematic variation of the background field with latitude (Dicke, 1970a).

In summary, the oblateness of the undisturbed Sun's surface is very nearly that of a surface of constant density. After subtracting the contribution from surface rotation, the remainder represents the oblateness of a surface of constant gravitational potential. This oblateness in turn uniquely determines the gravitational

quadrupole moment. Magnetic and velocity fields in the "seen layers" could change these conclusions, but no such fields capable of seriously affecting these relations are found.

IV. THE SPIN-DOWN PROBLEM AND OTHER QUESTIONS OF STABILITY

It has been claimed (Howard, Moore, and Spiegel, 1967) that the Sun probably could not have a rapidly rotating core, because of loss of angular momentum by dynamically driven circulation currents associated with the formation of an Ekman layer, i.e., the spin-down effect. It has also been claimed that because of an instability due to a thermally driven turbulence, associated with a thermal diffusivity large compared with that of angular momentum, a rapidly rotating core is precluded (Goldreich and Schubert, 1967a, 1967b; Fricke, 1968). This firm position of Goldreich and Schubert was later modified somewhat (Goldreich and Schubert, 1968) after Colgate (1968) showed that a compositional gradient in the zone of differential rotation could stabilize a rapidly rotating core.

The spin-down effect is easily seen in a cup of tea where a rotation of the stirred tea ceases in a time short compared with the diffusion time. The rapid slowing is due to pumping of the tea through a thin (Ekman) layer at the bottom of the cup.*

In attempting to relate this phenomenon to the solar interior it is essential to understand the significance of an important difference between a cup of tea and the solar interior. The density of tea is constant, whereas the solar medium is compressible. Furthermore, there is an important difference between the outer convective zone of the Sun and its radiative core. In the convective zone, pressure and density are functionally connected through the adiabatic condition, but not in the radiative core. If we neglect viscous forces, purely rotational motion is on cylinders in the convective zone. This is seen by noting that for $P = P(\rho)$ and for purely rotational motion with the angular velocity ω,

$$\nabla P + \rho \nabla \varphi + \rho \omega \times (\omega \times r) = 0 , \qquad (17)$$

*For additional treatment of the teacup analogy, see Chapter 3, "A History of Solar Rotation", by Spiegel.

and, dividing through by ρ,

$$\nabla \int_0^P \frac{1}{\rho} dP + \nabla\varphi + \omega \times (\omega \times r) = 0 , \tag{18}$$

Equation 18 is valid also for the incompressible tea. From Equation 18, $\omega \times (\omega \times r)$ must be the gradient of a scalar, and normal to circular cylinders:

$$\omega \times (\omega \times r) = -\nabla \int \omega^2 r \sin\theta d(r \sin\theta) . \tag{19}$$

Thus, ω is constant on circular cylinders.

In the stirred cup of tea, because of the boundary condition on the bottom of the cup, purely rotational motion on cylinders is impossible. The inclusion of the viscous-force term significantly affects the motion in a thin layer (Ekman) at the cup bottom and permits the boundary condition to be satisfied. Owing to the reduction of the centrifugal-force density in this thin layer, fluid is propelled inward, pumping the whole content of the cup through the Ekman layer where viscous dissipation is great. This causes a slowing of the rotation in a time approximately equal to $a/(\omega\nu)^{1/2}$, where a is the cup dimension, ω is the angular velocity of the fluid, and ν is the kinetic viscosity of the fluid. This is much less than the diffusion time a^2/ν.

Spin-down would be expected in the convective zone of the Sun, but other complications are associated with the functional relation between pressure and density. Angular momentum per unit mass could not increase inward toward the rotation axis because of the Kelvin-Helmholtz instability. Also, for an appreciable variation of angular velocity, turbulence would be excited, the Reynolds criterion being easily satisfied.

In the presence of turbulence, the spin-down phenomenon becomes modified, a phenomenological turbulent viscosity playing a role similar to molecular viscosity. The presence at the Sun's surface of differential rotation in spite of the enormous turbulent-viscous force has long been something of a mystery. To drive this differential rotation, a large torque is required. The best candidate for this force seems to be the viscous force itself, the extra component added through anisotropic turbulence providing the driving force (Kippenhahn, 1963; Cocke, 1967b).

Just below the convective zone, the temperature gradient is nearly adiabatic, but a few thousand kilometers deeper the temperature gradient has greatly decreased

and the gas is strongly density stratified. The appropriate criterion for stability against turbulence for an angular-velocity gradient normal to constant-density surfaces is not that of Reynolds, but rather that of Richardson:

$$\left(\frac{r}{\omega}\frac{d\omega}{dr}\right) \leqslant \frac{4}{\gamma}\left(\frac{g}{\omega^2}\right)\left[(\gamma-1)\frac{d}{dr}\ln\rho - \frac{d}{dr}\ln T\right]. \tag{20}$$

This stability criterion holds only for rotation on spheres $\omega = \omega(r)$. From the Weymann (1957) solar model, the values of $+r/\omega\, d\omega/dr$ for equality in Equation 20 are -90, -1230, and -1730 at $r = 0.84$, 0.76, and 0.64, respectively. If ω is not constant on spherical surfaces (i.e., if it depends upon θ), turbulence should be excited. It might be expected that this turbulence would eliminate the dependence of ω on θ.

Pressure and density are uncoupled in the radiative core. The functional relation between P and ρ, which leads to Equation 18 and forces purely rotational motion to be on cylinders, is relaxed. For an arbitrary choice of $\omega(r, \theta)$, the dependences of both P and ρ on θ are separately determined by the function $\omega\,(r, \theta)$ (Dicke, 1967c). If the θ component of Equation 16, defined to lie in a surface of constant gravitational potential φ, is integrated over the surface, the dependence of P on θ over this surface is obtained. By taking the curl of Equation 16, a similar integration can be carried out for ρ. Thus, the θ dependences of P and ρ (and hence of T if the mean molecular weight is constant) are determined by the rotational distribution.

Purely rotational motion (no spin-down) is possible in the density-stratified solar interior if the temperature distribution is appropriate, but the adopted rotational distribution need not be stable. The importance of density stratification on spin-down has been discussed several times from different viewpoints by Holton (1965), Pedlosky (1967, 1969), Dicke (1967c), Holton and Stone (1968), Sakurai (1969a, 1969b), and Clark et al. (1969).

The effect of density stratification on spin-down was exhibited experimentally (McDonald and Dicke, 1967). A density-stratified fluid was established in corotation with a steadily rotating cylindrical dish. The rotational rate of the dish could be changed by a fractionally large amount without inducing spin-down if the change were made slowly in very small steps. If the angular velocity were changed discontinuously by as much as 1 percent, spin-down would occur through a series of complex events, starting with the excitation of gravity waves, followed by mixing in

two layers and separate spin-down of each of the mixed layers. The sudden change in the angular velocity of the dish imposes on the density-stratified fluid a rotational distribution that, for purely rotational motion, is incompatible with the actual density distribution.

It is concluded that dynamically driven spin-down currents do not occur in the density-stratified solar interior. The time scale, of the order of 10^{10} years for slowing the solar rotation by the solar wind, is extremely long compared with the rotation period, and the inertial effects of the circulation currents that maintain the correct density distribution are negligible. Thus, there is adequate time for the solar temperature to automatically adjust itself to satisfy the dynamical requirements of a purely rotational motion. Whereas dynamically driven Ekman-type currents probably do not exist, Eddington-Sweet thermally driven circulation currents should occur, unless there is a gradient in molecular weight in the zone of differential rotation. In general, the dual requirements of the rotational distribution, on pressure and density, lead either to a variation of molecular weight or else to a temperature distribution that is incompatible with the requirements of heat balance. The velocities of Eddington-Sweet currents associated with differential rotation can be orders of magnitude greater than the more familiar thermally driven currents associated with uniform rotation (Schwarzschild, 1958). It is a common mistake to apply the time scale associated with Eddington-Sweet currents under uniform rotation to situations with differential rotation.

The instability discussed by Goldreich and Schubert (1967a, 1967b) and Fricke (1968) takes place through the development of axially symmetric angular-velocity variations on spherical surfaces. Thin toruses, approximately 1 km thick, are continuously generated and destroyed, moving upward and downward and transporting angular momentum. It was noted (Dicke, 1967b) that the theory of this instability assumed the absence of a magnetic field in the deep interior of the Sun. Goldreich and Schubert had noted that a negligibly small θ component of velocity was required. Clark et al. (1969) have proposed that the oscillating motion of internal gravity waves driven by turbulence in the convective zone could provide the θ velocity component that would stabilize the flow. Fricke (1969) has investigated the effects of magnetic fields on the instability. He finds that a strong toroidal magnetic field ($\approx 10^5$ G) in the zone of differential rotation can stabilize the rotation if the field strength increases outwardly. Colgate (1968) had shown that this instability could be eliminated by the existence of a molecular weight gradient in the shell of differential rotation (also see Goldreich and Schubert, 1968). The required gradient in the mean molecular weight is slight and could be established by

the production of helium in the core if a sufficiently rapid means of mixing the core were available. The gradient in mean molecular weight necessary to stabilize the core satisfies the equation (Goldreich and Schubert, 1968)

$$\frac{r\,d\ln\mu}{dr} < 2\left(\frac{\omega^2 r}{g}\right)\frac{1}{r\omega}\frac{d}{dr}(r^2\omega)\,. \tag{21}$$

For the model of a rotating core to be discussed below, the right side of Equation 21 is roughly 6×10^{-2}, and the fractional increase in mean molecular weight μ in the core over that of the exterior need be only 3×10^{-3}. Roughly 5×10^{7} years of nuclear burning with the products mixed uniformly through the core would be required to increase μ by this amount in the core.

The ordinary Eddington-Sweet thermally driven currents associated with a rigidly rotating core are too slow to mix the core, even if the core rotates as rapidly as we postulate (Schwarzschild, 1958). But differential rotation in the core, or a strong poloidal magnetic field buried in the core in the perpendicular rotator configurations, could greatly increase the circulation rate, by several orders of magnitude. Mixing by thermally driven currents might occur for a few hundred million years but then be choked by the accumulated molecular weight gradients. If this happened, the evolutionary tracks in the H-R diagram might be only slightly modified.

If the Goldreich-Schubert-Fricke instability does occur, what is the limiting velocity distribution, assuming an initially uniform and rapidly rotating Sun slowed by the solar wind? This instability is very effective at transporting angular momentum as long as the angular momentum per unit mass increases inward toward the rotation axis, but ordinary viscosity-driven turbulence would be expected to develop if ω were a function of θ. Thus, the quasi-stable limiting distribution would be expected to be of the form $\omega \sim r^{-2}$ below the convective zone, with ω constant in that zone except for the above-mentioned differential rotation generated perhaps by anisotropic turbulence in the differentially rotating zone. The long-term stability of the distribution below the convective zone depends upon the effectiveness of thermally driven currents, upon whether or not they have been choked by gradients in molecular weight, and possibly upon other complications, such as an internal magnetic field.

One possible distribution, particularly interesting because it can be tested observationally, is a rapidly rotating core inside a shell of differential rotation (of

thickness $\delta r/r_0 \approx 0.05$) through which the angular momentum leads by molecular diffusion. Outside of this to the convective zone is a thick shell through which the angular momentum is transported by the Goldreich-Schubert process and in which $\omega \sim r^{-2}$. Outside the shell is the convective zone.

This model differs from the one first proposed (Dicke, 1964) in that the zone of molecular diffusion could lie substantially deeper. The observation that lithium, but not beryllium, is depleted with time suggests that the outer radius of the zone of molecular diffusion may fall below $r = 0.58R_\odot$, where ^{7}Li is quickly burned, but outside $0.5R_\odot$ if the Weymann solar model is correct. This new model will be discussed in some detail below.

To summarize, it is the lack of observations of the deep solar interior that makes conclusions about instabilities uncertain. Because of strong density stratification below the convective zone and the mild nature of braking by the solar wind, dynamically driven spin-down currents probably do not exist, but fairly rapid thermally driven circulation currents in the zone of differential rotation are possible, though they would be very easily choked by gradients in mean molecular weight. The Goldreich-Schubert-Fricke instability is easily inhibited by such a molecular weight gradient or by oscillatory motion in the θ direction. Density stratification stabilizes purely rotational motion on spheres, and it is concluded that ω is a function of r if a stable rapidly rotating core exists. Of more importance than these theoretical arguments concerning the deep interior are the observations of the solar surface.

V. THE SOLAR-WIND TORQUE

Although the structure of the solar wind is not directly of concern to us, observations of the solar wind can provide a measure of the solar-wind torque, and this is of importance to the problem of internal solar rotation. If we make the questionable assumption that the solar wind blows substantially radially out to the vicinity of Venus and the Earth, measurements of solar-wind flux (performed with the Mariner space probes) and of the magnetic-field strengths in the wind permit an evaluation of the solar-wind torque. The assumption of radial flow when the Sun is quiet may be questionable in the light of the appearance of the corona.

The solar-wind torque density on the solar surface at the equator is (Dicke, 1964; Modiesette, 1967; Weber and Davis, 1967; Alfonso-Faus, 1967)

$$K = Jr^2\omega_0 , \tag{22}$$

where J is the mass flux density at the solar surface, ω_0 is the angular velocity, and r is a "critical radius" for which

$$\rho v^2 = \frac{1}{4\pi} B^2 , \tag{23}$$

namely, the radius at which v, the radial component of the solar-wind velocity, equals the Alfvén velocity calculated from B, the radial component of the magnetic field.

The magnetic field is trapped in the wind, and B falls off inversely as the square of the radius. Equation 23 can be written as

$$\rho = 4\pi J^2 / B_0^2 , \tag{24}$$

where $B_0 = (r/r_0)^2 B$ is the strength of the trapped magnetic field referred to the solar surface. This is a particularly interesting way to express the field, for the magnetic field at the solar surface cannot be too strong or trapping is impossible. The physical properties of the chromosphere and lower corona limit the strength of the trapped field. Before the strength of the interplanetary field was measured, B_0 had been estimated to be 0.75 G (Dicke, 1964). This estimate was based on the assumption that the field strength would lie near its upper limits.

Measurements of the magnetic-field strength at 1 AU with the Mariner 2 space probe gave an rms value for the radial component of roughly 3.5×10^{-5} G, $B_0 \approx 1.4$ G (Coleman, 1966). If we make the simplified assumption that near the Sun's surface substantially cylindrical magnetic flux tubes are stretched out from the solar surface by the solar wind, the magnetic pressure $(1/8\pi)B^2$ cannot exceed the gas pressure outside these tubes. From Allen's model of the solar corona at the equator, the rough upper limit for B_0 takes on the values 0.8, 0.7, and 0.5 G from gas pressures at $r = 1.01R_\odot$, $1.1R_\odot$, and $1.4R_\odot$, respectively. The concentration of the magnetic field toward the equatorial plane may have increased the field strength at the Earth's radius.

The mass flow in the solar wind is determined by the rate of heating of the corona. This heating is believed to be caused by acoustic noise generated by turbulence in the convective zone. It would be expected to be more or less constant in time, depending upon the relative importance of magnetohydrodynamic waves in coupling the corona to the convective zone. There are observational reasons for believing that young solar-type stars are more active magnetically than older stars (Wilson, 1966).

The magnitude of the surface density of the mass flux in the solar wind, from the Mariner 2 space probe, is $J \approx 1.7 \times 10^{-11}$ g-cm^{-2}-s^{-1} (Neugebauer and Snyder, 1966). From Equation 24, combining the results obtained with the space probe with Allen's (1963) model of the corona gives $r = 20r_0$ for the critical radius. From Equation 22, $K = 9 \times 10^7$ dyne-cm^{-1}. For the total solar-wind torque,

$$\frac{8\pi}{3} r_0^2 K = 3.8 \times 10^{30} \text{ dyne-cm}^{-1} . \tag{25}$$

One should not be misled by the apparent precision of this number, which may be uncertain by a factor of 2. This precision and those of similar apparently accurate numbers below are introduced only to make the arithmetic well defined.

The torque (Equation 25) is proportional to the angular velocity of the solar surface. With the assumptions that the magnetic field B_0 lies near its maximum value and that the solar-wind strength J has been reasonably constant, the torque density per unit angular momentum, K/ω_s, would be reasonably constant over the life of the Sun, but the greater magnetic activity in young solar-type stars would be expected to increase the torque. If magnetic coupling to the corona provides the dominant means of heating the corona, the solar-wind flux in the young Sun could have been substantially greater.

VI. THE EVOLUTION OF THE RAPIDLY ROTATING CORE IN THE SUN

In this section, the picture developed above will be adopted as a working hypothesis, and a quantitative history of the Sun's rotation will be developed. The radius of the core will be assumed to be $r_c = 0.54R_\odot$, permitting rapid burning of lithium at the core boundary, but not of beryllium (see Section VII). It will be assumed that the solar core, of radius $0.54R_\odot$, is rotating uniformly with the angular velocity $\omega = 20\omega_0$ needed to generate the quadrupole moment associated with the solar oblateness ($\omega_0 \approx 2.87 \times 10^{-6}$ s^{-1}). The angular velocity of $20\omega_0$ corresponds to a 1.27-day period; at a core radius of 0.8, an angular velocity of $15\omega_0$ is needed.*

It is impossible to transport angular momentum from the outer bounds of such a core to the bottom of the convective zone at $r_v = 0.86R_\odot$ (Weymann, 1957) by

*B. E. McDonald, 1969, in a private communication.

molecular diffusion. It will be assumed that the thermally driven turbulence discussed by Goldreich and Schubert (1967b, 1968) permits the transport by turbulent diffusion in the range $r_c < r < r_v$, a thin shell of molecular diffusion limiting the flow of angular momentum from the core. As discussed above, thermal turbulence is easily inhibited. It will be assumed that the core boundary is stabilized, probably by a molecular weight gradient or by oscillatory motion in the θ direction.

Thermal turbulence is so effective when it occurs that the angular velocity gradient cannot appreciably exceed the threshold value $r(d\omega/dr)/\omega = -2$. This gradient will be assumed, or ωr^2 = constant in this shell. As discussed above, it will be assumed that ordinary mechanically driven turbulence keeps the angular-velocity gradient parallel to the density gradient, i.e., angular velocity is a function of r only.

It will be assumed that the Sun is a typical star. Thus, solar history might be illuminated by Kraft's (1967) observations of rotation in young stars. He finds that the surface rotation $20\omega_0$ of stars of $1.2M_0$ in the Pleiades, 3×10^7 years old, has dropped to $10\omega_0$ in the Hyades, 5×10^8 years old. Observations are missing for G2 stars, but extrapolation curves suggest angular velocities as low as $5\omega_0$ and $2.5\omega_0$ for the Pleiades and the Hyades, respectively. If the above picture is correct, solar-type stars, with their deep convective envelopes, arrive on the main sequence either with a substantial amount of differential rotation already present or with a strong stellar wind acting to slow the outer shell in a time as short as 3×10^7 years.

We assume that the angular velocity is substantially constant at the surface value ω_s in the convective zone (down to $r_v = 0.86R_\odot$) and that for a fully developed turbulent zone, $\omega = (r_v/r)^2 \omega_s$ for $r_c < r < r_v$. A stellar-wind equatorial torque density $K_s = 10^8 \ (\omega_s/\omega_0)$ dyne-cm^{-1} (substantially the same as that observed for the solar wind) must act for 1.63×10^9 years on an initially uniformly rotating star to develop the thermal turbulent zone down to the core radius $r_c = 0.54R_\odot$. At this time $\omega_s = (r_c/r_v)^2 \omega_c = 0.394\omega_c$. Subsequent slowing of the fully developed shell, together with the convective zone, occurs with an e-folding time of 6.05×10^9 years. This neglects the (initially small) contribution to the solar-wind torque from angular momentum leaking out of the core (Table 1).

To account for the factor of 2 decrease of ω_s from the Pleiades to the Hyades in 5×10^8 years would require that the solar-wind torque density $K_s = 8.35 \times 10^8 \ (\omega_s/\omega_0)$. The solar-wind torque may have been even greater in the first 3×10^7 years. If the torque density were as great as $K_s = 150 \times 10^8 \ (\omega_s/\omega_0)$, the Sun could initially have been uniformly rotating on the main sequence. For reasons discussed above, a torque this great seems unlikely, and a strong torque during the late

Hayashi phase seems more likely. It will be assumed that the Sun arrived on the main sequence with the surface rotating with an angular velocity $\omega_s \approx 5\omega_0$.

The solar-wind torque density K_s can be decomposed into two parts, K_t and K_d. Here, $K_t = -6.05 \times 10^8 \ (\dot{\omega}_s/\omega_0)$ represents the contribution from the deceleration of the outer shell ($\dot{\omega}_s$ means time derivative with 10^9 years as the unit of time); and K_d, the contribution from the loss of angular momentum from the core, is evaluated by solving the diffusion equation

$$\frac{\partial}{\partial r}\left(\rho\nu r^4 \frac{\partial\omega}{\partial r}\right) = \rho r^4 \frac{\partial\omega}{\partial t} \tag{26}$$

as a boundary-value problem, assuming that $(r_v/r_c)^2\omega_s$, the angular velocity at the core boundary, is known as a function of the time. If $\rho\nu r^4$ varies slowly enough through the shell of molecular diffusion (thickness $\approx 0.05R_\odot$), a good approximation is obtained by replacing it by its (constant) value at $r_c = 0.54R_\odot$. At this point ν, 10 percent of the kinematic viscosity (14.3 $cm^2 \cdot s^{-1}$) is due to radiation transport, and the remaining is due to transport by ions. In this approximation, for $r < r_c$,

$$\omega(r,t) = -(r_v/r_c)^2 \int_{-\infty}^{t} (d\omega_s/d\tau)\{\mathrm{erf}\,[(r_c - r)/\sqrt{2\nu(t-\tau)}] - 1\}\, d\tau + \omega_c \,. \tag{27}$$

From Equation 27, the radial derivative at r_c is calculated, giving

$$K_d = -r_0^{-2} r_c^4 \rho_c \nu_c (\partial\omega/\partial r)_c \tag{28}$$

as the core's contribution to the solar-wind torque density. If ω_s decreases exponentially to zero with a decay constant λ,

$$K_d = r_0^2 (r_c/r_0)^2 \rho_c \omega_c (\nu_c/\pi t)^{1/2} \sqrt{\lambda t}\, W(\lambda t) \,, \tag{29}$$

where

$$W(x) = e^{-x} \int_0^x e^y y^{-1/2}\, dy \,. \tag{30}$$

The function $D(x)$ has the value

$$D(x) = x^{1/2} W(x) \approx 1.3x\left(1 - \frac{1}{4}x\right) \qquad \text{for} \quad x < 2$$

$$\approx 1 + 1.2/(x+2) \qquad \text{for} \quad x > 2 \,.$$

Substituting numerical values gives

$$K_d = 1.38 \times 10^8 t^{-1/2} D(\lambda t) \text{ dyne-cm}^{-1}, \tag{31}$$

where t is in units of 10^9 years. For $\lambda t > 0.5$, K_d varies slowly with time. The above formalism is easily generalized to cover the case

$$\omega_s(t) = \int_0^\infty A(\lambda) e^{-\lambda t} d\lambda,$$

for which $D(\lambda t)$ in Equation 31 is to be replaced by

$$\bar{D} = \int A\, D d\lambda \Big/ \int A\, d\lambda. \tag{32}$$

The time dependence of the surface rotation is obtained as a solution of the differential equation $K_s = K_t + K_d$, namely,

$$N(t) \times 10^8 (\omega_s/\omega_0) = -6.05 \times 10^8 (\dot{\omega}_s/\omega_0) + K_d(t), \tag{33}$$

where $N(t)$ is the ratio of the solar-wind torque density to surface angular velocity, expressed in units of $10^8\ \omega_0$ g-s^{-1} characterizing the present solar wind. The general solution to this equation is easily written if we assume that $N(t)$ and $K_d(t)$ are known. For sufficiently large values of N and slow variation of K_d/N, ω_s falls until $|K_t| \ll K_d$. These conditions seem to be approximately satisfied for N varying in such a way as to give a satisfactory account of the surface rotations adopted for the Pleiades and Hyades, and of that observed in the Sun. As a result, it is to be expected that the present value of the solar-wind torque is approximately equal to

$$K_d \approx 0.8 \times 10^8 \text{ dyne-cm}^{-1}, \tag{34}$$

which is in satisfactory agreement with the observations of the solar wind.

Two variations of N with time seem to be particularly interesting. The results of numerical integrations with these choices are plotted in Figure 4. For the curves A of Figure 4, N is assumed to fall exponentially with time, and the three parameters characterizing N and the initial value of ω_s are adjusted to give a satisfactory account of the three "known" values of ω_s. For the curves B, N is the sum of a constant and an exponential. The mean life of the exponential is arbitrarily

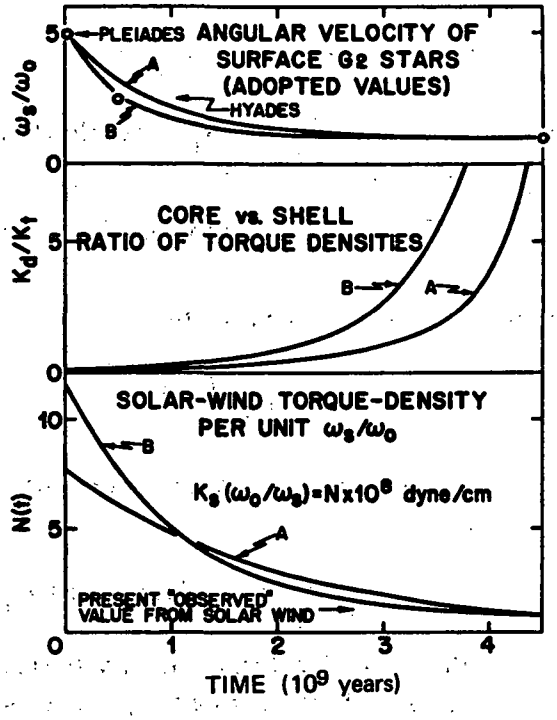


Figure 4.—The slowing of the Sun's rotation, assuming for curve *A*, a solar-wind torque with *N* decreasing exponentially in time; and for curve *B*, *N* decreasing exponentially to a constant value.

taken to be of the order of 10^9 years. This term is introduced to represent the decay of the initially strong solar activity. The decay with time of magnetic activity in solar-type stars, as exhibited by Ca-II emission (Wilson, 1966), might be due to the decay of short-lived magnetic modes originally trapped in the Sun. As noted above, increased magnetic activity could result in a stronger solar wind, particularly if magnetic coupling to the corona is the primary source of coronal heating. The initial values of the two terms are adjusted to give surface rotations agreeing with the "observations".

It should be remarked that the integrations *A* and *B* require the present solar-wind torque density to be approximately 0.85×10^8 dyne-cm^{-1}. This is in satisfactory agreement with the observations of the solar wind.

VII. LITHIUM DEPLETION IN SOLAR-TYPE STARS

The depletion of lithium in solar-type stars can be related to the deceleration of their surface rotations if the model developed above is reasonably correct. In the thermal-turbulent zone, the diffusivities of angular momentum and lithium are equal. But this diffusivity is determined by two requirements: The angular momentum flux must be correct, and ωr^2 must be constant in this zone, with ω independent of latitude.

Integrating Equation 26 from r_c to r and using the relation

$$r^2\omega = r_v^2\omega_s \tag{35}$$

gives

$$\rho v r = \frac{1}{2\omega_s}\left(\frac{r_0}{r_v}\right)^2\left[K_d + K_t\frac{M(r) - M(c)}{M(v) - M(c)}\right]. \tag{36}$$

In Equation 36, $M(r)$, $M(c)$, and $M(v)$ are stellar masses inside the designated radii; the minor contribution from the convective zone has been omitted from K_t (as defined in the paragraph above Equation 26).

The diffusion of lithium is controlled by the equation

$$\frac{\partial}{\partial r}\left(\rho v r^2\frac{\partial F}{\partial r}\right) = \rho r^2\frac{\partial F}{\partial t}, \tag{37}$$

where F represents the fractional abundance of ^{7}Li or ^{6}Li (by mass or number).

The solution to Equation 37 is eased by the simplifying assumption that the zone of burning has a sharp boundary. This requires $F = 0$ as a condition on the boundary. There is also a condition to be satisfied at the inner boundary of the convective zone. This is determined by the requirement that the radial derivative of F, which determines the flow of lithium from the convective zone, be proportional to the time derivative of F, which gives the loss of lithium in the convective zone.

A normal solution to Equation 37, with Λ as the decay constant, satisfies the eigenvalue equation

$$\frac{\partial}{\partial r}\left(\rho v r^2\frac{\partial F}{\partial r}\right) + \Lambda\rho r^2 F = 0. \tag{38}$$

This is to be integrated subject to the above-described boundary conditions. Equation 36 is first substituted in Equation 38. If the zone of burning is $r < r_c$, the normal solutions depend upon the parameter K_d/K_t. For the lowest mode, we shall need the slope-to-value ratio $r(\partial F/\partial r)/F$ evaluated at r_ν (as a function of K_d/K_t); this is given in Table 2.

All higher normal modes decay rapidly, an order of magnitude faster for the second mode, and the subsequent fractional decay rate is that of the lowest mode, providing K_d/K_t varies slowly with time.

The decay rate of the lowest mode is conveniently expressed in terms of the solar-wind torque. Integrating Equation 38 from r_ν to r_0, the solar surface, gives

$$-\left(\rho\nu r^2 \frac{\partial F}{\partial r}\right)_\nu = \int \rho r^2 \dot{F}_\nu \, dr = \frac{1}{4\pi} \dot{F}_\nu [M(0) - M(\nu)] . \tag{39}$$

Substituting Equation 36 gives

$$\Lambda = 2\pi \frac{K_s}{\omega_s} \left(\frac{r_0}{r_\nu}\right)^2 \left(\frac{r\partial F/\partial r}{F}\right)_\nu \frac{1}{M(0) - M(\nu)} . \tag{40}$$

Taking K_s/ω_s and K_ν/K_t from Figure 4 and using Table 3 gives Λ as a function of time. This permits the integration of $\dot{F} = -\Lambda F$. The resulting curves *A* and *B* are plotted in Figure 5, which is based on Figure 2 of Danziger (1969). Note that these curves contain no adjustable constants.

It should be emphasized that the above integration is based on the assumption that the boundary of lithium burning is sharp and that it occurs at the bottom of the zone of turbulent diffusion. This may be reasonable if we assume that the outward-moving boundary is initially somewhat below r_c. Two corrections tending slightly to lower the upper ends of curves *A* and *B* of Figure 5 have been omitted. The first is caused by the delay in arrival of the outward-moving boundary of

Table 2.–The slope-to-value ratio of F versus K_d/K_t.

K_d/K_t	0.02	0.05	0.10	0.2	0.4	0.8	1.6	3.2	6.4
$r\frac{dF}{dr}/F$	0.074	0.096	0.118	0.150	0.193	0.234	0.273	0.304	0.319

Table 3.—Fractional depth r_b/r_o at which burning takes place with indicated mean life (based on Fowler et al., 1967, and the solar model by Weymann, 1957).

Element	3×10^6 years	3×10^7 years	3×10^8 years
^{6}Li	0.57	0.60	0.63
^{7}Li	.51	.55	.58
^{9}Be	.42	.45	.47

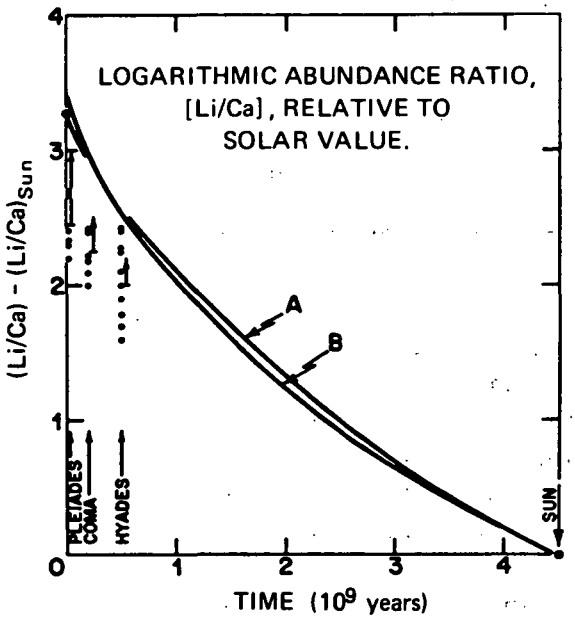


Figure 5.—The depletion of lithium; the same turbulence viscosity as that associated with the transport of angular momentum is assumed (curves *A* and *B* of Figure 4). The turbulence is assumed to be driven *in part* by the thermal effect of Goldreich and Schubert (1967b). The plotted points represent individual stars (Danziger, 1969). The three arrows are Danziger's corrections for "curve-of-growth effects".

lithium burning at r_c, possibly a time approximately 10^8 years. The second is caused by the delay in the start of depletion caused by the necessity of the decay of the higher decay modes of F, a time approximately 10^8 years.

If the boundary of burning extends somewhat outside r_c, the upper ends of the curves are raised slightly, but explicit integrations have not been carried out. If the boundary of lithium burning occurs at least $0.1r_0$ above r_c, the right side of Equation 36 is approximately independent of r, which simplifies the solution of Equation 38. This case has previously been discussed by Dicke (1970c), assuming that the zone of burning occurs at $r = 0.58R_\odot$, for the adopted radii of the convective zone $r_v = 0.86$, 0.78, and $0.70R_\odot$. For the "observed" present solar-wind torque-ratio density the mean decay times for ^{7}Li are found to be roughly the same as Danziger's value of 7×10^8 years. But the decay times are almost a factor of 2 too small for a solar-wind torque density adequate to slow the outer shell with $r_c = 0.54R_\odot$.

VIII. SUMMARY

The types of observations having a bearing on the rotation of the deep solar interior are as follows:

(1) The observation of the solar oblateness of $(r_{eq} - r_p)/r = 5 \times 10^{-5}$.

(2) The independence of latitude of the solar photospheric brightness.

(3) The magnetic- and velocity-field distributions in the photosphere.

(4) The structure of the solar wind, suggesting a present solar-wind torque density, at the equator, of 10^8 dyne-cm^{-1}.

(5) The rotation observed in F- and G-type stars in young clusters.

(6) The lithium and beryllium abundance in the Sun and in solar-type stars of young clusters.

In the light of the uniformity of solar photospheric brightness, the solar oblateness seems to require a gravitational quadrupole moment sufficient to advance the perihelion of Mercury's orbit by about 4″ per century.

A reasonable account of all of the above observations can be given by assuming the following:

(1) That the Sun is a typical star of 1 solar mass.

(2) That it possesses a core of radius $\approx 0.55R_\odot$ rotating at an angular velocity 20 times greater than that of the Sun's surface.

(3) That it arrived on the main sequence with the surface rotation already slowed to ≈5 times the present surface value, but with the core rotating uniformly at 20 times the present surface rate.

(4) That the young Sun had a ratio of solar-wind torque to surface angular velocity roughly an order of magnitude greater than its present value (an increase probably associated with the increased magnetic activity of the young Sun).

(5) That this enhanced torque decayed with time, either quickly or more slowly.

It should be explicitly stated that if the overall picture is qualitatively correct, the radius of the core is rather accurately fixed by the requirement that it fall outside $0.5R_\odot$ where beryllium is burned, but inside $0.58R_\odot$ where ^{7}Li is burned [expressed in terms of Weymann's (1957) solar model as given by Schwarzschild (1958, see p. 259)].

The radius of the rapidly rotating core is fixed by the requirement that the outer parts of the Sun be mixed down to $r = 0.58R_\odot$ (to destroy lithium) but not to $0.5R_\odot$ (to avoid destroying beryllium). The rotation of the core is fixed by the observation of the solar oblateness, assuming that the oblateness implies a quadrupole moment due to a rapidly rotating solar interior. The past slowing of surface rotation in the Sun is crudely fixed by observations on young stellar clusters, assuming that the Sun is a typical star. To slow the surface rotation by this amount yields a present value for the solar-wind torque in agreement with observations of the solar wind, but only if the Sun possesses such a rapidly rotating core. The present value of the rate of loss of angular momentum from the core is substantially independent of the time scale for slowing the young Sun. Without any adjustable parameters, the depletion of lithium in solar-type stars is determined by the loss of angular momentum, if the model is correct. The resulting losses are found to be in reasonable agreement with observations of lithium in young clusters and the Sun.

REFERENCES

Alfonso-Faus, A., *J. Geophys. Res.* 72:5576, 1967.
Allen, C. W., *Astrophysical Quantities*, Athlone Press, London, 1963, 2nd ed.
Ashbrook, J., *Sky Telescope* 34:229, 1967.
Audretsch, J., Dehnen, H., and Hönl, H., *Astrophys. J. (Letters)* 150:L127, 1967.
Bergmann, P., *Ann. Math.* 49:255, 1948.
Brans, C., and Dicke, R. H., *Phys. Rev.* 124:925, 1961.
Chazy, J., *La Théorie de la Relativité et de la Mécanique Celéste*, Gauthier-Villars, Paris, 1928.

Clark, A., Thomas, J. H., and Clark, P. A., *Science* 164:290, 1969.
Clemence, G. M., *Astron. Pap., Amer. Ephem.* 11:1, 1943.
Cocke, W. J., *Phys. Rev. Lett.* 19:609, 1967a.
Cocke, W. J., *Astrophys. J.* 150:1041, 1967b.
Coleman, P. J., Jr., *J. Geophys. Res.* 71:5509, 1966.
Colgate, S. A., *Astrophys. J. (Letters)* 153:L81, 1968.
Cowling, T. C., in *Stellar Structure*, L. H. Aller and D. B. McLaughlin, eds., University of Chicago Press, Chicago, 1965, Chap. 8.
Danziger, I. J., *Astrophys. J.* 150:733, 1967.
Danziger, I. J., *Astrophys. Lett.* 3:115, 1969.
Deutsch, A. J., *Science* 156:236, 1967.
Dicke, R. H., *Phys. Rev.* 125:2163, 1962.
Dicke, R. H., *Nature* 202:432, 1964.
Dicke, R. H., and Peebles, P. J. E., *Space Sci. Rev.* 4:419, 1965.
Dicke, R. H., *Stellar Evolution*, R. F. Stein and A. G. W. Cameron, eds., Plenum Publishing Corporation, New York, 1966, p. 319.
Dicke, R. H., *Int. Astron. J.* 8:29, 1967a.
Dicke, R. H., *Sky Telescope* 34:371, 1967b.
Dicke, R. H., *Astrophys. J. (Letters)* 149:L121, 1967c.
Dicke, R. H., *Science* 157:960, 1967d.
Dicke, R. H., and Goldenberg, H. M., *Phys. Rev. Lett.* 18:313, 1967a.
Dicke, R. H., and Goldenberg, H. M., *Nature* 214:1294, 1967b.
Dicke, R. H., *Astrophys. J.* 159(1), 1970a.
Dicke, R. H., *Astrophys. J.* 159(1), 1970b.
Dicke, R. H., in *IAU Colloq. No. 4, Stellar Rotation*, A. Slettebak, ed., Reidel, Holland, 1970c.
Duncombe, R. L., *Astron. Pap., Amer. Ephem.* 16:1, 1958.
Durney, B. R., and Roxburgh, I. W., *Nature* 221:646, 1969.
Fowler, W. A., Caughlan, G. R., and Zimmerman, B. A., *Ann. Rev. Astron. Astrophys.* 5:525, 1967.
Fricke, K., *Z. Astrophys.* 68:317, 1968.
Fricke, K., *Astron. Astrophys.* 1:338, 1969.
Gilvarry, J. J., and Sturrock, P. A., *Nature* 216:1283, 1967.
Goldreich, P., and Schubert, G., *Science* 156:1101, 1967a.
Goldreich, P., and Schubert, G., *Astrophys. J.* 150:571, 1967b.
Goldreich, P., and Schubert, G., *Astrophys. J.* 154:1005, 1968.
Herbig, G. H., *Astrophys. J.* 141:588, 1965.
Holton, J. R., *J. Atmos. Sci.* 22:402, 1965.
Holton, J. R., and Stone, P. H., *J. Fluid Mech.* 33:127, 1968.
Howard, L. N., Moore, D. W., and Spiegel, E. A., *Nature* 214:1297, 1967.
Hoyle, F., *Quart. J. Roy. Astron. Soc.* 1:28, 1960.
Jordan, P., *Astron. Nachr.* 276:1955, 1948.
Jordan, P., *Schwerkraft und Weltall*, Vieweg, Braunschweig, 1959.
Kippenhahn, R., *Astrophys. J.* 137:564, 1963.

Kraft, R., *Astrophys. J.* 150:551, 1967.
Kraft, R., in *Stellar Astronomy*, H. Y. Chiu, R. Warasila, and J. Remo, eds., Gordon and Breach Science Publishers, Inc., New York, 1968, p. 2.
Leverrier, U. J., *Ann. Obs. Paris* 5:104, 1859.
Livingston, W. C., *Sci. Amer.* 215:107, 1966.
McDonald, B. E., and Dicke, R. H., *Science* 158:1562, 1967.
Modiesette, J. L., *J. Geophys. Res.* 72:1521, 1967.
Neugebauer, M., and Snyder, C. W., *J. Geophys. Res.* 71:4469, 1966.
Newcomb, S., *Suppl. Amer. Ephem. Naut. Alm.*, 1897.
O'Connell, R. F., *Astrophys. J. (Letters)* 152:L11, 1968.
Öpik, E. J., *Int. Astron. J.* 8:29, 1967.
Pedlosky, J., *J. Fluid Mech.* 28:463, 1967.
Pedlosky, J., *J. Fluid Mech.* 36:401, 1969.
Plaskett, H. H., *Observatory* 85:178, 1965.
Roxburgh, I. W., *Icarus* 3:92, 1964.
Roxburgh, I. W., *Nature* 213:1077, 1967a.
Roxburgh, I. W., *Nature* 216:1286, 1967b.
Sakurai, T., *J. Phys. Soc. Jap.* 26:840, 1969a.
Sakurai, T., *J. Fluid Mech.* 37:689, 1969b.
Schatzman, E., in *IAU Symp. No. 10*, J. L. Greenstein, ed., 1959, p. 129.
Schatzman, E., *Ann. Astrophys.* 25:18, 1962.
Schaub, W., *Astron. Nachr.* 265:161, 1938.
Schur, W., and Ambronn, L., *Astronomische Mitteilungen* (Göttingen Obs.), 1895, Part 4 (also see 1905, Part 7, for discussion of results).
Schwarzschild, M., *Structure and Evolution of the Sun*, Princeton University Press, Princeton, New Jersey, 1958, pp. 175-184.
Shapiro, I. I., *Icarus* 4:549, 1965.
Sturrock, P. A., and Gilvarry, J. J., *Nature* 216:1280, 1967.
Thirry, Y. R., *C. R. Acad. Sci.* 226:216, 1948.
von Zeipel, H., *Mon. Notic. Roy. Astron. Soc.* 84:665, 1924.
Wallerstein, G., and Conti, P. S., *Ann. Rev. Astron. Astrophys.* 7:99, 1969.
Wayman, P. A., *Quart. J. Roy. Astron. Soc.* 7, June 1966.
Weber, E. J., and Davis, L., Jr., *Astrophys. J.* 148:217, 1967.
Weymann, R., *Astrophys. J.* 126:208, 1957.
Wilson, D. C., *Science* 151:1487, 1966.

CHAPTER 3
A HISTORY OF SOLAR ROTATION*

E. A. Spiegel
Columbia University
New York, New York

I. THE PREHISTORY

Stars form in the turbulent, magnetized interstellar gas in a way that is only partly understood. However, it is clear that they tend to form in clusters and associations that contain a spectrum of stellar masses and rotation rates. Thus, for an individual star such as the Sun, we do not know the initial angular momentum. This, of course, makes it difficult to produce a deductive account of the evolution of solar rotation, even if the theoretical problems were not so formidable. Of course, at this stage, one may even wonder whether rotational evolution occurs at all, and, even if it does, why it should be worthy of the attention we are about to devote to it.

As to the occurrence of rotational evolution, there is, however, little doubt, since solar-type stars show a clear correlation between parameters indicating their ages and their surface rotation rates. Moreover, we know that the Sun loses mass, which carries off angular momentum; unfortunately, we do not know how active this process was early in the Sun's lifetime, and this is one of the great uncertainties of the subject. Finally, we know that stars change their radii with time (sometimes abruptly), and this too must affect their angular velocities.

My own interest in solar rotation is related to questions of stellar evolution which will be mentioned briefly later. However, in our present concerns with the solar system, rotation has mainly nuisance value. Solar rotation produces solar oblateness and a quadrupole moment in the density distribution. This is a small effect, but in an accurate subject such as celestial mechanics, it might make itself noticeable. This is especially true if one is interested in distinguishing among various proposed theories of gravity. Hopefully, in time, when other tests of gravity are used to make these distinctions, celestial mechanics might help us unravel the vexing

*Research supported by the National Science Foundation (NSF GP 18062).

question of solar rotation; but the part of the problem that will be with us much longer is concerned with the early stages, especially when the planets formed. If it is true that the planets originated in the solar nebula, then from the standpoint of cosmogony, we surely need a knowledge of the dynamics of the nebula. The rotational part of the problem probably will not be the hardest to solve, but it will certainly be crucial, and there is no promise of direct observational guides.

The problem of describing the Sun's early years is at present approached, in the first approximation, by neglecting rotation. A spherically symmetric gas cloud that is in its own gravitational well is considered. It is assumed to be cool so that its pressure cannot support it against its own gravity. How the cloud got that way is by no means understood, but this model seems to be a reasonable starting point. Among astronomers, it is hoped that these initial conditions are not misleading. However, the masses are important, and, unfortunately, it appears that the initial radii also matter.

The gas cloud, having no support, will collapse with essentially free-fall velocity. Kinetic energy is dissipated into thermal energy, and as long as the cloud is tenuous, and hence transparent, the thermal energy is radiated away. In this phase the matter remains cool, and the pressure does not rise enough to impede the collapse; but fairly soon, the density becomes large enough to trap photons, and the temperature and pressure increase. The pressure then halts the collapse, and contraction begins.

In the contraction phase, the cloud (or star) is nearly in hydrostatic equilibrium, but it radiates, and this leads to a slow contraction. The contraction rate is governed by the rate at which radiation can leak from the star. The rate of energy loss is the luminosity, and if this is divided into the energy available for radiation (the thermal energy) the characteristic time for contraction is obtained—the so-called Kelvin-Helmholtz time. This period of time is typically millions of years.

For a theoretical discussion of solar rotation, a quantity of interest is the time spent in what is called the Hayashi phase of stellar contraction. This phase is characterized by an opacity sufficiently large that photons diffuse very slowly through the star. In that case, the star "prefers" to transport the heat out by convection: Hot material rises to the surface, radiates away much of its energy, cools, and sinks back into the star. In the Hayashi phase, the star is fully convective, and because the energy is removed so efficiently, the subsequent Kelvin-Helmholtz contraction is accelerated. (For discussions of these questions, see Larsen, 1969, and Graham, 1969.)

The importance of this matter arises from the fact, to be explained later, that convective stars lose mass and angular momentum. Hence, the time spent in the Hayashi phase will probably be vital in determining the angular momentum that is left in a star when its contraction phase ends. The situation is such that until we sort out these problems and carry through the contraction calculation for a rotating star, we cannot predict the angular velocity distribution in, for example, the Sun at the end of its contraction phase. The story of solar rotation could not, therefore, be told deductively even if other difficulties did not intervene.

Let us assume, for the present purposes, that at the end of its Hayashi phase a star rotates rigidly. This initial rotation rate Ω_0 corresponds to the beginning of the phase of evolution in which the Sun, for example, now finds itself. The question we shall consider here is how the Sun's rotation rate evolves from such an initial state of rigid rotation.

II. OBSERVED HISTORY OF SOLAR ROTATION

The Hayashi contraction phase of the Sun ends when the core becomes hot enough to produce nuclear reactions. These stabilize the Sun at a fixed radius since a further contraction makes the reactions go faster, thus heating the Sun so that it reexpands. This is true for all stars above a given mass (approximately $0.07M_\odot$). They arrive at a static state with values of luminosity L and surface temperature T_e that depend on their masses. For a given chemical composition, there is a relation between L and T_e, as is indicated in Figure 1. The solid curve, called the zero-age main sequence (ZAMS), shows L versus T_e, with the mass as a parameter on the curve. Shortly after the Hayashi phase, a star arrives on the main sequence, where it stays as long as it has hydrogen available for conversion into helium. This first nuclear stage lasts about 10^{10} years for the Sun, during which the Sun changes very little but moves slightly away from the ZAMS.

Most stars fall very near to the main sequence, but the band they occupy is fairly broad. This is due partly to observational scatter, partly to a spread in ages, and somewhat to differences in initial composition, rotation, and magnetic fields. For the Sun, observations and theory are good enough that its departure from the ZAMS can be used to infer its age ($\approx 6 \times 10^9$ years). For other stars, individual ages cannot readily be found, but for clusters of stars the ages can be determined from the distortion of their main sequences, on the supposition that the stars in a cluster all have the same age. If we can then measure the rotations of stars, a study of rotation versus age is possible.

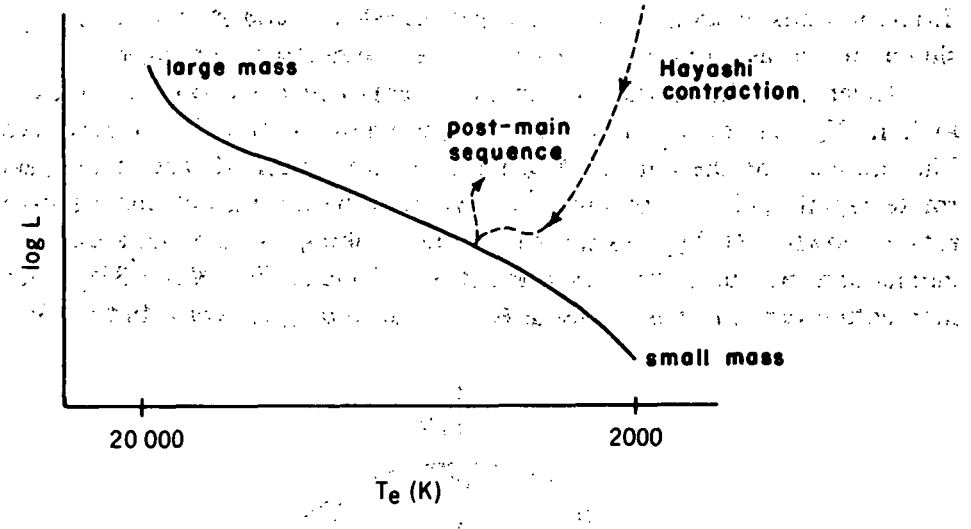


Figure 1.—Typical relation between luminosity L and surface temperature T_e.

The rotation rates of stars (or at least of the stellar surfaces) can be inferred from the widths of absorption lines in their spectra. Suppose that if there were no rotation the lines would be narrow. Then, for a star with rotation axis normal to our line of sight and with constant surface angular velocity, a strip of projected surface parallel to the rotation axis would by itself produce a line shifted in wavelength from that produced by the central strip (Figure 2).

The contribution of a given strip to the total line is proportional to the area of the strip for a uniform brightness over the face of the star. In that case, if the individual contributions are very narrow, the resulting profile is an ellipse whose width indicates the rotational velocity at the star's equator. Now we should back up and worry about the effects of line widths of the individual strips and of the darkening of the stellar disk near the edge. These effects blur the profiles and the various effects have to be unraveled; suffice it to say that this can be done reasonably well.

A difficulty with this solution of the problem lies in our assumption that the axis of rotation is in the plane of the sky. In general, one really measures not the equatorial velocity, V_e but $V_e \sin i$, where i is the angle the rotation axis makes with the sky plane. Except for eclipsing binary stars, which are of no help here, we have no way to find i. Thus, if we want to measure and compare rotation rates of

different kinds of stars, we must proceed statistically and attempt to average out the sin i under the assumption that rotation axes are randomly oriented.

Given that the appropriate measurements and corrections are made, one can look at $\langle V_e \sin i \rangle$ or $\langle V_e \rangle$ if you assume random orientation, for main sequence stars. The summary of the data by Kraft (1970) is very useful. The more massive stars rotate rapidly [$\langle V_e \rangle \approx 200$ km-s^{-1}], whereas the less massive main sequence stars rotate slowly [$\langle V_e \rangle \lesssim 30$ km-s^{-1}]. More striking is the variation of angular momentum per unit mass as a function of mass. For $M \gtrsim 1.3 M_\odot$, the angular momentum per unit mass varies as $M^{0.57}$, according to Kraft; but for $M \lesssim 1.3 M_\odot$,

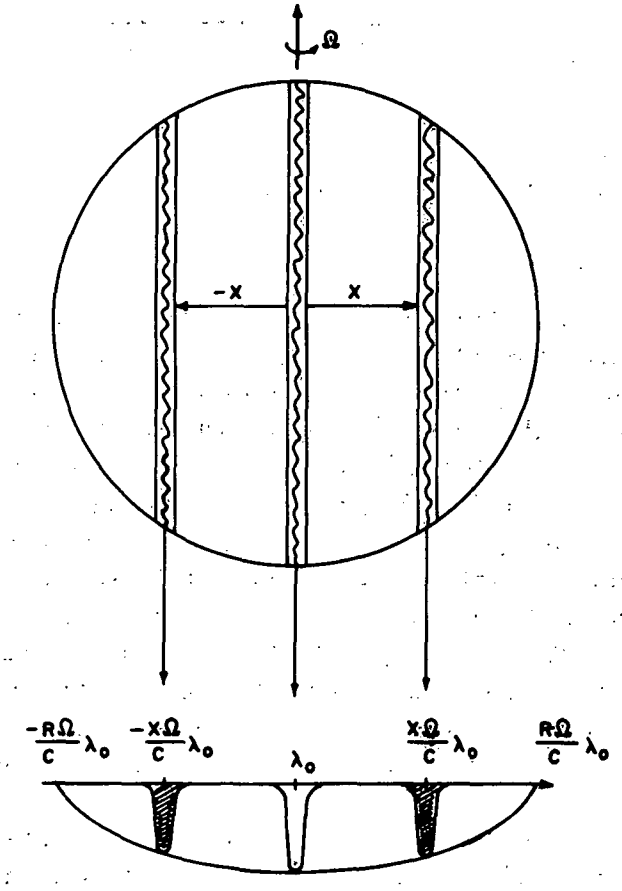


Figure 2.—Shift in wavelength of a line, produced by rotation of projected strip of surface parallel to the stellar axis of rotation and normal to viewer's line of sight.

the angular momentum J per unit mass falls very sharply with mass. The break in the curve occurs at a surface temperature $T_e \approx 8000$ K, and it is significant that stars with $T_e \lesssim 8000$ K have powerful convective zones in their outer layers. Before this fact was stressed and its importance realized (see the next two sections), it was suggested that the break in J/M versus M was a result of planetary formation. This remark was based mainly on the solar system, for which it was observed that if the angular momenta of the planets were added to that of the Sun, the value of J/M for the Sun would fall fairly near the curve extrapolated from data for stars with large masses. (It should be noted here that the curve given by Kraft is based on the assumption of rigid rotation; Dicke has proposed another model which is given elsewhere in this volume.*)

An interesting application of observations of stellar rotation is the attempt to see how stars like the Sun but of differing ages differ in rotation rate. To obtain stars of given age we must go to clusters. We need clusters with enough stars like the Sun to be able to average out the sin i factor and also to obtain a good value for the stellar ages. To get enough stars to average over, we must include stars somewhat different in mass from the Sun. At present, we have data for this problem for the Hyades and the Pleiades which were given by Kraft (1967), and we adopt the averages suggested by Conti (1968) (Table 1). In the case of the Sun, there is no problem with sin i; but, on the other hand, if stars of a given age have a spread in V_e, only one star is dangerous to use. Still, it seems an inescapable conclusion that stars like the Sun rotate more slowly (at least at their surfaces) as they age. Let us now see why.

Table 1.—Stellar ages and rotational velocities.

Object	Age (years)	$\langle V_e \rangle$ (km-s^{-1})
Pleiades	5×10^7	19
Hyades	5×10^8	9
Sun	5×10^9	2

*See Chapter 2, "Internal Rotation of the Sun", by Dicke.

III. SOLAR FLUID DYNAMICS

Perhaps the most basic fluid dynamical activity in the Sun is convection. This is a familiar phenomenon and occurs when a fluid is heated sufficiently intensely from below. The criterion for the onset of convection is that

$$dT/dr < -g/C_p , \tag{1}$$

where dT/dr is the vertical temperature gradient, g is the local acceleration of gravity, and C_p is the specific heat at constant pressure. The term adiabatic gradient (or adiabatic lapse rate) is used for g/C_p. The meaning of this criterion (called the Schwarzschild criterion by astronomers) is that the potential energy per unit mass, gdr, gained by a parcel displaced vertically by an amount dr should be more than the internal energy $C_p dT$ it loses in adiabatic expansion. Equation 1 must be modified to allow for dissipation, rotation, and magnetic fields (Chandrasekhar, 1961), but in the case of the Sun, these modifications are generally small (with the main exception of corrections for sunspots).

Near the surfaces of relatively cool stars like the Sun, Condition 1 is easily met because the temperatures just below the surfaces are such as to partially ionize hydrogen. This causes C_p to be large. Moreover, the material near the surface is highly opaque because negative hydrogen ions are formed; this would force $-dT/dr$ to be large if the photons are to get through. It is also worth noting that some stars have convection in their central regions. This would occur when the rate of nuclear burning is very sensitive to the temperature, as it is in the CNO cycle. In that case, the temperature gradient becomes large, and efficient convection is produced. On these grounds, we do not expect convection in the core of the Sun, but it should occur in the cores of massive stars. However, it is important to realize that according to Iben's (1966) calculations, the Sun supported the CNO cycle during its first 30 million years on the main sequence, and thus had a convective core during this period.

The effect of convection on the structure of the Sun is difficult to compute, and astronomers do not always agree even on its qualitative aspects; but it is clear that the instability is very pronounced and that in a fairly extensive part of the outer layers of the Sun, convective heat transfer dominates over radiative transfer; this is likely to be true for about 150 000 km into the Sun. When convective transfer dominates, a large value of $-dT/dr$ is not needed to force out the photons, and the gradient drops. It cannot go below the value g/C_p, since that would stabilize the

convection, so the gradient takes on the value of g/C_p over the bulk of the convection zone.

At the edges of the convection zone, transition layers occur in which convection is not very efficient and the gradient goes somewhat above g/C_p. This is especially true at the top, where the material is becoming more transparent and less able to carry the heat. There, a fairly marked instability is maintained. This has two effects. First, the convective velocities become moderately large (of the order of a kilometer per second), so the turbulent stresses can deform the medium and generate waves. Second, even without such stresses, sound waves may become unstable in the transition layer and thus can be self-generated from thermal noise. I do not know which process dominates, but I suspect that in different frequency ranges different processes win out. In either case, such waves are believed to travel outward in the Sun, steepen into shocks, and dissipate their mechanical energy into heat in the low-density regions (see the discussion by Schatzman and Souffrin in *Annual Reviews of Astronomy and Astrophysics*). This gives rise to the chromosphere and the corona.

Though the waves carry off relatively little energy from the convection zone, this small amount is adequate to raise the corona to a high temperature ($\approx 10^6$ K) since the density is so low. The corona then becomes ionized, and heat is transported by electron conduction. The failure of various people to find a static solution for the corona led Parker to propose a dynamic solution with an expanding corona (see Cowling, 1969). Schematically, we may say that the waves continually bring energy into the corona, and it is unable to get rid of this energy fast enough. Thus, it heats up and expands off the Sun, giving rise to the solar wind.

The picture outlined here seems qualitatively reasonable, and different parts of it have been calculated. However, a complete theory is still lacking; in particular, a prediction of the mass loss rate from the Sun cannot be made with any accuracy. Observations indicate that if the Sun continued to lose mass at its present rate, it would lose all of its 2×10^{33} g in about 10^{13} years.

IV. BRAKING OF SOLAR ROTATION

We have seen, schematically at least, how strong convective instability near the surface of a star gives rise to a wind. The association between slow rotation and low surface temperature, and hence strong surface convection, might then be explained by the loss of angular momentum resulting from the wind. But the mass-loss rate

involved in creating the solar wind is so slight that it might not produce a significant braking were it not for magnetic effects, as Schatzman recognized. The role of the magnetic field is to impart rigidity to the gas for some distance above the solar surface and thus keep it in approximate corotation with the solar surface even after it has left it. (For a summary of these matters see Brandt, 1970, and Mestel's discussion in the *16th Liège Symposium.*) Of course, after the matter is far enough from the Sun, the field becomes too weak to maintain corotation with the Sun, and Kepler's law of areas is obeyed.

The importance of the corotation is that the gas has the angular velocity of the solar surface Ω out to a distance R_c, which is generally greater than $R_\odot$. Therefore, at R_c the gas has an angular momentum per unit mass $= \Pi^2\Omega$, where $\Pi = R_\odot \cos\varphi$ and φ is the solar latitude. Hence, the solar wind removes angular momentum from the Sun at a rate

$$dJ/dt = -\dot{M}_\odot R_c^2 \Omega \,. \tag{2}$$

Here,

$$J = \alpha M_\odot R_\odot^2 \Omega \,, \tag{3}$$

where α is a pure number and $M_\odot$ and $R_\odot$ are the solar mass and radius. Normally, α would be taken to be of order 0.1, but Dicke's (1967) proposal that only the outer layers of the Sun slow down would imply $\alpha \ll 0.1$; this matter will be discussed below. In any case, we then have

$$d\Omega/dt = -\alpha(\dot{M}_\odot/M_\odot)(R_c/R_\odot)^2\Omega \,. \tag{4}$$

Effective braking results when $(R_c/R_\odot)^2 \gg 1$, and we must now estimate this quantity (see Mestel's discussion in the *16th Liège Symposium*).

We assume that the flow in the solar wind is steady and that the flux of mass is uniform over the solar surface. Conservation of mass requires

$$4\pi r^2 \rho u = \dot{M}_\odot \,, \tag{5}$$

where u is the radial component of velocity. Conservation of magnetic flux demands that

$$4\pi r^2 B = \text{constant} = Q = 4\pi R_\odot^2 B_s \,, \tag{6}$$

where B is the radial component of magnetic field and B_s is its surface value. This field can keep the gas rotating rigidly as long as the magnetic pressure exceeds the dynamic pressure, or, qualitatively speaking, as long as

$$B^2/4\pi \gtrsim \rho u^2 . \tag{7}$$

In the present solar wind, the matter is accelerated off the solar surface by pressure; thus, thermal energy is converted to kinetic energy. The main thrust of this acceleration should thus be accomplished when u climbs above the sound speed c, which we will take to be constant. Thus, in the neighborhood of R_c we have the crude estimate, for $u \approx c$,

$$\rho_c \approx \dot{M}_\odot/4\pi R_c^2 c . \tag{8}$$

If we take the equality in Equation 7 and use it with Equation 6, this leads us to the estimate

$$(R_c/R_\odot)^2 \approx \frac{R_\odot^2 B_s^2}{\dot{M}_\odot c} \approx 5B_s^2 , \tag{9}$$

with B in gauss, for $c \approx 10^7$ km-s^{-1}. With values of B_s in the neighborhood of 10 gauss, we obtain a reasonable agreement with more precise estimates of the enhancement factor. This does not mean that we should take this discussion too literally. It is merely intended to show the various parameters that enter the calculation and to stress the uncertainties. In any case, current theories used in connection with satellite measurements of the solar-wind particle flux and the field in the Earth's neighborhood suggest that $(R_c/R_\odot)^2$ is approximately 10^2 or 10^3.

To solve Equation 2 we need also to know how B_s varies with time. Most experts assume that the field in the convection zone is caused by some dynamo mechanism. Cowling (1969) has given dimensional arguments that suggest that in this case $B_s \sim \Omega^2$, but I am not sure that these arguments properly allow for the possible effects of a strong rotation on the convective motions that must be involved in the process. Another problem is that we do not know how c depends on B_s and, hence, on Ω, but since the waves that heat the corona probably have hydromagnetic aspects, such a dependence is to be expected. It is likely that c goes up with B_s and, hence, with Ω, but this suspicion is based only on the observed increase of calcium emission with enhanced fields, and the interpretation here is very complicated. Thus,

about all we can conclude is that B_s^2/c probably increases with Ω but does so more slowly than Ω^4. Other uncertainties in the estimates exist, but we have seen enough to shake our confidence in any conclusion about what these processes might have been during epochs for which the parameters cannot be determined from observation.

Thus, surrendering all pretense of a deductive theory, we must admit that all we can say is

$$d\Omega = -f(\Omega, \ldots)\Omega \,, \tag{10}$$

where the braking factor f depends on Ω and the ellipsis indicates a possible dependence on other parameters that describe stellar structure, such as the mass of the star. As mentioned previously, the hydromagnetic theory of the solar wind, when combined with observations, yields at present

$$\frac{1}{f} = \left(\frac{1}{\Omega}\frac{d\Omega}{dt}\right)^{-1} \approx 5\alpha \times 10^{10} \text{ years} \,. \tag{11}$$

For $\alpha \approx 0.1$, the half-life of the solar angular momentum is comparable to the age of the Sun. (The value 0.1 comes from estimating the moment of inertia of the Sun, allowing for its density and structure.) I feel that we must consider the possibility that this agreement is not just a chance one, but is a clue to the nature of the process. To see how this can be so, let us make the simple assumption that f varies as a power of Ω (Spiegel, 1968) so that $f = \beta\,\Omega^n$, where β is a constant and the crude physics of the problem indicates that n is in the range of about 1 to 4. Then,

$$d\Omega/dt = -\beta\Omega^{n+1} \,, \tag{12}$$

and we find that, for $n \neq 0$,

$$\Omega = \Omega_0/(1 + t/T)^{1/n} \,, \tag{13}$$

where Ω_0 is an integration constant (the initial rotation rate) and the time

$$T = (n\beta\Omega_0^n)^{-1} \,. \tag{14}$$

Now, if these formulae are to represent the change in Ω from Pleiades to Hyades (assuming an approximately constant Ω_0), we should have $T \approx 10^8$ years, for n in the range of values suggested by the theory. Moreover, from Equation 13 we find a

half-life for Ω, i.e.,

$$\left|\left(\frac{1}{\Omega}\frac{d\Omega}{dt}\right)^{-1}\right| = n(t+T)\,, \tag{15}$$

and for the present Sun, this half-life is nt. Thus, the apparent coincidence of the solar age and angular momentum follows from Equation 10 and the crude estimates for n. However, for Dicke's model, $\alpha \ll 1$, and we would require $n \ll 1$, which does not work. (The value $n = 0$ gives an exponential decay curve which does not fit the data and is not really consistent with our primitive understanding of the process; but at the present level of understanding, nothing is reliable.) Of course, this merely implies that if only a small fraction of the solar mass is acted on by the solar wind, it would long ago have stopped rotating. However, the bulk of the Sun would then be rotating with essentially its original angular velocity. To save his model, Dicke would require some coupling between the inner and outer parts of the Sun, in order to keep the surface angular velocity from going to zero. Thus, the problem raised by his views is one of the coupling between the inner and outer parts of the Sun when the surface is being decelerated by the solar wind torque. We now turn to a discussion of this question.

V. SOLAR SPIN-DOWN

Several years ago Temesvary (1952) suggested that hydromagnetic braking of the rotation of the solar surface would produce instabilities in the solar interior. His discussion does not explicitly have what is now called the solar wind, nor does it have the process called "spin-down", but its general outlines are sound, and it anticipated by many years some of the current discussions of motion in the solar interior. Here, the chief departure from his point of view will be the manner in which the deceleration of the solar surface makes itself felt inside the Sun. To see this in a more homely context, we will consider first the problem of how tea in a cup slows down after being stirred.

The kinematic viscosity ν of tea (i.e., viscosity/density) is approximately 0.01 $cm^2\text{-}s^{-1}$. This viscosity prevents the tea in contact with the wall of the cup from moving relative to the cup. The tea has a sharp velocity gradient at the cup, and the action of this gradient would stop the tea in a time approximately equal to d^2/ν, where d is the diameter of the cup. For a typical teacup, this time is about 40 minutes. Clearly, the tea stops more quickly than this; hence, another process must be acting. This process is called spin-down (see Greenspan, 1968); for good physical discussions, see Einstein's (1934) paper on the meandering of rivers in *Mein Weltbild* and the discussions in the "Miscellaneous Topics" chapter of Prandtl, 1952.

The spin-down process for the tea is most simply discussed in the case where the cup is set rotating at an angular velocity Ω and a steady rigid rotation is achieved in the tea. The rotation rate of the cup is then changed to $\Omega(1+\epsilon)$, with $\epsilon \ll 1$, and we need to know how soon the tea adjusts to the new rotation rate of its container.

When the cup is slowed, the fluid in contact with bottom slows too. This effect is spread by viscosity over a thin layer at the bottom called the Ekman layer. In this layer, the centrifugal potential $\frac{1}{2}\Omega^2\rho\Pi^2$ (where Π is distance from the symmetry axis) is reduced by an amount

$$\Delta\left(\frac{1}{2}\Omega^2\rho\Pi^2\right) \approx \Omega^2\Pi^2\rho\epsilon \,. \tag{16}$$

The gradient of this change is the residual centrifugal force

$$\frac{\partial}{\partial\Pi}(\Omega^2\Pi^2\rho\epsilon) \approx 2\Omega^2\Pi\rho\epsilon \,, \tag{17}$$

which represents a radial force and causes an inward flow u (which carries the tea leaves with it). The inward motion is opposed by a viscous force

$$\eta\nabla^2 u \approx \eta\frac{u}{h^2} \,, \tag{18}$$

where η is the viscosity and h is the thickness of the Ekman layer. The approximate balance of the viscous force and the residual centrifugal force leads to the estimate

$$u \approx \frac{2\Omega^2\Pi h^2\epsilon}{\nu} \,, \tag{19}$$

where $\nu = \eta/\rho$.

To estimate h we note that because of the radial motion the fluid suffers an azimuthal coriolis force of approximately $\rho u\Omega$. Moreover, the fluid is moving with an azimuthal speed $v \approx \Pi\epsilon\Omega$ with respect to the cup such that the azimuthal force balance is

$$\eta\nabla^2 v \approx \eta v/h^2 \approx \rho\Omega u \,, \tag{20}$$

whence we find

$$h \approx (\nu/\Omega)^{1/2} \tag{21}$$

as the Ekman layer thickness.

Now, the fluid in the Ekman layer converging toward the center of the cup has nowhere to go but up, so it develops an upward velocity w. The mass flux inward in the Ekman layer is approximately $\rho u(\pi\Pi)h$, and this balances the flux out of the Ekman layer, which is approximately $\rho w(\pi\Pi^2)$. Thus,

$$w \approx (h/\Pi)u \approx (2\Omega^2 h^3/\nu)\epsilon \,. \tag{22}$$

With the value of h given by Equation 21, we have

$$w \approx (\nu/\Omega)^{1/2}\Omega\epsilon \,, \tag{23}$$

which is the rate at which fluid is "pumped" into the interior of the cup from the Ekman layer.

The flow upward into the cup causes a circulation in the tea upward, outward toward the rim, and down the sides back into the Ekman layer (Figure 3). For a cup of height of the order of d the horizontal motion in the main body of the tea is comparable to the pumping velocity w of Equation 23. The resulting circulation alters the angular velocity of the tea. To see how this happens, consider a ring of fluid in the interior of the tea.

The outward component of the circulation in the tea will cause the radius R of the ring to increase. In such an expansion, the angular momentum will be approximately conserved, so the angular velocity of the ring will decrease. The change in angular velocity $\delta\Omega$ for a change in ring radius δR is obtained by setting

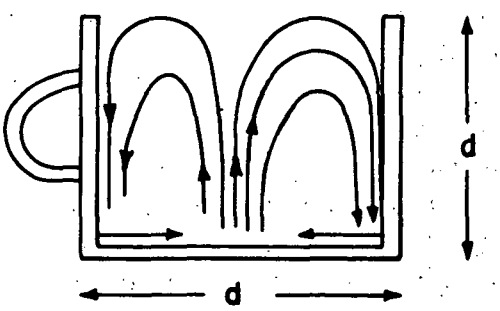


Figure 3.–Flow of tea in a teacup.

the change in angular momentum per unit mass equal to zero since on the large scale of the interior, viscous forces are negligible. That is,

$$\delta(R^2\Omega) = 0$$

implies

$$\delta\Omega \approx (\Omega/R)\delta R \,. \tag{24}$$

To slow the ring down to the new speed of the container, we need $\delta\Omega = \epsilon\Omega$, for which an expansion $\delta R \approx \epsilon R$ is required.

If $R \approx d$, the velocity of expansion of the ring u is comparable to w as given by Equation 23. Hence, the time required for the tea to spin down to the speed of the container is

$$\tau \approx \delta R/u \approx d/\sqrt{\nu\Omega} \,, \tag{25}$$

which is approximately 1 minute for a teacup. This time is called the spin-down time (Greenspan and Howard, 1963).

However, one may still ask the question, if the angular momentum is conserved, will the ring of fluid speed up after passing through the Ekman layer and recontracting? The answer, of course, is that as the tea comes down the side of the cup and rubs against it, the angular momentum is destroyed.

The spin-down time given here is based on the response to a slight but sudden change in the rotation of the cup. The actual motion is time dependent, and various waves are generated that take longer to die out. In the case where the cup is slowed down continuously but slowly, a quasi-steady circulation such as the one discussed here is set up. The angular velocity in the tea lags that of the cup by the spin-down time τ. [This was shown by Bondi and Lyttleton (1948) in what is probably the first derivation of Equation 25.]

Now, in applying these ideas to the Sun, we must take into account three special difficulties: (1) there is no teacup, (2) the interior of the Sun is stratified, and (3) the Sun is spherical. These are not all the differences, but they are the main ones, and several comments should be made about them.

The slowing down of the surface of the Sun can be thought of as the result of a torque exerted by the corotating solar wind on the Sun through the magnetic lines of force. These force lines extend into the convection zone, which then is slowed down. The eddying motions in this zone keep the zone nearly rigid, and thus the

convection zone is bodily slowed down. The suggestion has been made that the near-rigidity of the convection zone causes it to act like a solid wall in setting up an Ekman layer (Howard, Moore, and Spiegel, 1967). This may indeed be true, but it appears that a more powerful mechanism may act (Bretherton and Spiegel, 1968). The mechanism, which we may call convective pumping, is simply that the convection zone itself acts like an Ekman layer. When the zone is braked by the solar wind, its loss in centrifugal potential starts a weak circulation, just as occurred with the tea in the Ekman layer. Thus, if the rotation of the convection zone is changed from Ω to $\Omega(1 + \epsilon)$, the arguments that led to Equation 19 still hold, except that ν is to be replaced by the eddy viscosity in the convection zone, and u is the north-south velocity component. Now, however, we do not use Equation 21 for h, but instead, we use the thickness of the convection zone. Finally, the argument that gives Equation 22 also holds true with ν replaced by ν_{eddy}, and it gives a reasonable estimate of the speed with which fluid is pumped into the solar interior. The enhancement over the Ekman pumping is about a factor of 10^5 for the case of the Sun.

If the interior of the Sun were not stably stratified, the convective pumping would result in a spin-down time of about 1 month. Of course, the argument given here and the mathematical model used are both rather simplified, and there might be some other way for the fluid to take up partially the spin-down stress. However, as far as the calculations have gone, they indicate a rather powerful convective pumping.* What seems the most powerful inhibiting factor the spin-down process must face is the stable stratification of the interior, which we now discuss.

Since the interior is stratified, it tends to reject the fluid that the convection zone is pumping into it. The fluid is mainly pumped in from the poles and initially moves toward the center of the Sun. The stratification turns it aside, and it flows in a layer just under the convection zone to reenter the convection zone in the equatorial regions. This layer accounts for the greater share of the velocity difference between the convection zone and the deep interior; it is analogous to a layer in the ocean called the thermocline, and to emphasize this analogy, I shall call it the tachycline.

Perhaps the first person to consider the effect of stratification on spin-down was Holton (1965). When Howard, Moore, and I discussed spin-down in the Sun, his results were duplicated by a physical argument and extended to include the effects of radiative transfer. However, at about the same time, Pedlosky (1967) concluded

*Indeed, there is even the possibility that it is amplified by convective instability.

that the effect of sidewalls of the container, which had been neglected in the stratified case, actually completely suppressed the Ekman pumping. Since this went against the physical arguments, many people did the calculation independently and the matter has been straightened out. The result is that Holton's results are qualitatively correct. [Some of the other papers on this point are by Walin (1968) and Sakurai (1969); several others have done related calculations, but many of these have not been published.] Thus, the results seem quite good, so as above, only an outline of the physics will be attempted.*

Suppose, for this discussion, we regard the Sun as a cylinder. At the top and bottom are convective layers (Figure 4), and we suppose that the rotation of the convective layers is bodily changed from Ω to $\Omega(1+\epsilon)$. Fluid will then be pumped into the interior at the rate (see Equation 22)

$$w = (2\Omega^2 h^3/\nu_{\text{eddy}})\epsilon \,. \tag{26}$$

The fluid will penetrate a distance l before being turned aside. The speed $\hat{u}$ with which it moves out radially is given by mass conservation as

$$\hat{u} \approx (a/l)w \tag{27}$$

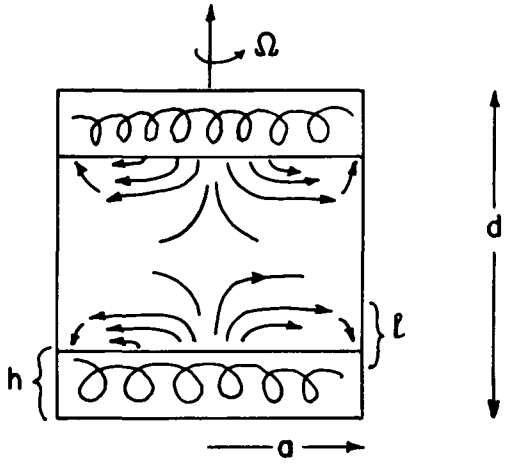


Figure 4.—Fluid flow in a Sun-like cylinder.

*Following an unpublished manuscript by D. W. Moore and me.

if the penetration layer does not extend over too great a density range. This radial outflow generates an azimuthal coriolis force. Since the viscosity in the solar interior is negligible and the situation is assumed to be axisymmetric, this force is unopposed and leads to a buildup of azimuthal velocity with time:

$$\hat{v} \approx \hat{u}\Omega t \,. \tag{28}$$

The azimuthal velocity produces a radial component of Coriolis force which is quickly counteracted by a pressure gradient:

$$\rho\hat{v}\Omega \approx \partial p/\partial \Pi \approx p/a \,. \tag{29}$$

It is understood that this pressure represents the deviation from the preexisting static pressure and that it includes the centrifugal potential. Since the pressure contains a centrifugal part, and since $\hat{v}$ varies with Π differently at different heights, a vertical pressure gradient must also develop across the layer. Since the vertical velocities in the stable regions are small, the pressure gradient is opposed chiefly by buoyancy; i.e., we can suppose a nearly hydrostatic situation.

The vertical pressure gradient is of the order of p/l, and if the density fluctuation is $\delta\rho$,

$$p/l \approx g\,\delta\rho \,. \tag{30}$$

For small fluctuations, we can write

$$\delta\rho \approx \rho\alpha\,\delta T \,, \tag{31}$$

where δT is the temperature perturbation and α is the coefficient of thermal compressibility. (If variations of chemical composition occur, an additional term is needed in Equation 31.) The combination of Equation 29, which is usually called the geostrophic condition, with Equations 30 and 31 gives

$$g\alpha\,\delta T/a \approx \hat{v}\Omega/l \,, \tag{32}$$

which is known as the thermal-wind equation. It shows how a horizontal temperature gradient in a rotating, stratified fluid forces motions, and it plays an important role in meteorological and stellar circulation problems.

As fluid is convected through the vertical gradient, it alters its temperature both convectively and by radiative diffusion. The rate of change of temperature is

$$\delta T/\partial t \sim \delta T/t \sim \mathbf{u} \cdot \nabla T + \kappa \nabla^2 (\delta T) , \tag{33}$$

where T is the unperturbed temperature and κ is the thermometric conductivity. With $\nabla^2 \approx -l^{-2}$, we obtain

$$(1/t + K/l^2)\delta T \approx w\, dT/dz , \tag{34}$$

where z is the vertical coordinate. By using Equation 34 to eliminate δT from Equation 32, introducing Equation 30, and using Equation 27 to eliminate $\hat{u}$, we are led to the relation

$$\frac{a^2}{l^2} \frac{\Omega^2}{g\alpha\, dT/dz} \left(1 + \frac{Kt}{l^2}\right) \approx 1 . \tag{35}$$

If the fluid is compressible, we must replace dT/dz by $dT/dz - g/c_p$ to take into account adiabatic expansion or contraction. We now introduce the quantity

$$N^2 = g\alpha(dT/dz + g/C_p) , \tag{36}$$

where N is the so-called Brunt frequency. If a parcel of fluid in a stably stratified medium is displaced vertically and adiabatically, it will oscillate with frequency N. In a hydrostatic star, $N \approx \sqrt{G\bar{\rho}}$, where $\bar{\rho}$ is the mean density.

Now consider the situation for an adiabatic medium with $\kappa = 0$; we find

$$l \approx (\Omega/N)a \tag{37}$$

as the thickness of the tachycline. If we take the present surface rotation of the Sun for Ω, we have $\Omega/N \approx 10^{-2}$. Thus, the solar tachycline would occupy 1 percent of the solar radius and be rather less than a local scale height in thickness. This crudely justifies treating the density as nearly constant; earlier in the Sun's history (or even now if Dicke were right) this would not be the case. The limit $\Omega \approx N$ is sometimes called the breakup velocity.

The tachycline is set up essentially instantaneously when the pumping begins. It spins down by the process discussed above since it is the region in which the stratification does not inhibit pumping. The argument for the adjustment of the angular velocity of a fluid ring, based now on $\hat{u}$ as given by Equations 26 and 27 leads to

$$\tau \approx \delta R/\hat{u} \approx \epsilon a/\hat{u} \approx l\nu_{\text{eddy}}/2\Omega^2 h^3 \approx (a/h)(\nu_{\text{eddy}}/h^2)1/N\Omega\,, \tag{38}$$

where h^2/ν_{eddy} is the characteristic time scale of the eddies in solar convection. Hence, on the long time of solar braking, the tachycline can be considered to be rotating with the convection zone. Of course, the tachycline is not rigid and represents only an exponential penetration layer, rather like the Hartmann layer of hydromagnetics; it supports the shear between the convection zone and the deep interior.

When the motion is not adiabatic, the picture is somewhat altered since radiative conduction can reduce δT and $\delta\rho$. If radiative conduction were very rapid, the fluid pumped from the convection zone could penetrate into the interior without much inhibition; but even for modest conduction, the flow can move slowly into the interior at a rate governed by the conduction. We may refer to Equation 35 and determine that the value of t for which $l = d$ is

$$t \approx [(N^2/\Omega^2) - 1]\, d^2/K\,. \tag{39}$$

Thus, for $N = \Omega$, the tachycline fills the cylinder in about a rotation period. But the present solar conditions imply $N^2 > \Omega^2$, and we have

$$t \approx (N^2/\Omega^2) t_{\text{KH}} \equiv t_{\text{ES}} \tag{40}$$

as the time required for the tachycline to spread through the Sun. Here, t_{KH} represents the thermal time, which is the same as the Kelvin-Helmholtz time. The time scale t_{ES} is, in other contexts, known as the Eddington-Sweet time.

Now, for the Sun, $t_{\text{KH}} \approx 10^7$ years, and if we take the surface value for Ω, $t_{\text{ES}} \approx 10^{11}$ years. However, if we take a value 15 times larger for Ω, as suggested by Dicke's model, we find $t_{\text{ES}} \approx 10^9$ years. Hence, the possibility of a really large time lag between interior and exterior is not unlimited.

There is one possible way out of these conclusions: The density stratification may in part be a result of molecular weight gradients. In that case, κ is replaced by

the diffusivity of chemical constituents, which is much smaller than κ. Of course, during the Hayashi phase, gradients of chemical composition would be destroyed. Once the nuclear reactions in the core start, they would tend to set up a gradient of helium concentration that would shield the core from incursions of material from the envelope, as Mestel pointed out some time ago. Even this part of the problem is not clear. First, Iben (1966) has asserted that the solar core was convective during its first 30 million years on the main sequence. Second, it is possible that the spin-down currents may initially keep the Sun mixed and prevent a gradient from forming; this effect would not go on indefinitely, but the time it lasts may be crucial. These processes can be studied in reasonable detail, but for the reasons given in the next section, I believe that the laminar spin-down circulation given here provides only a lower bound on the mixing processes and that far more remains to be done.

For the early phases when Ω is large and the Eddington-Sweet time quite short, it may be safe to assume that the core and the surface rotated at similar speeds. Then, we might risk using Equation 13 to infer the angular velocity for the Sun. From this formula, we can estimate the number of Eddington-Sweet times that have passed at a given solar age:

$$A = \int_0^t \frac{dt}{t_{ES}} = \left(\frac{n}{n-2}\right)\frac{T}{t_{KH}}\left(\frac{\Omega_0}{N}\right)^2\left[\left(\frac{t}{T}+1\right)^{(n-2)/n} - 1\right]. \tag{41}$$

Fitting Equation 13 to the sparse data gives $T/t_{KH} \approx 10$ and $\Omega_0/N \approx 1/3$. Hence, the solar age in units of t_{ES} is

$$A \approx \left(\frac{n}{n-2}\right)\left[\left(\frac{t}{T}+1\right)^{(n-2)/n} - 1\right]. \tag{42}$$

The limited available understanding of the process involved is not, as we saw, adequate to fix n, but a value of approximately 3 was plausible. This suggests that at present $A \gg 1$ and hence that the internal Ω would not lag the surface value by one or two orders of magnitude. On the other hand, Dicke's proposal calls for small n, as we saw, so that this last result is not really inconsistent with his picture. However, a small n does not seem likely and is not really called for by the data. If this were all that was involved, it could be concluded that the model of a rapidly rotating core seems unlikely, but it is not excluded; but there is much more to be understood, as we shall see in the next section.

The third aspect of the elementary solar spin-down process that is not the same as the teacup process is the spherical geometry. Thus, the gravity is radial, so there is a tendency toward spherical density contours. However, the rotation gives a cylindrical symmetry to the problem. This mixture of symmetries makes the solar case messy. Bretherton and I have been looking into this problem and have found that, physically, it is rather like the cases we have discussed. In the meridional planes, there is a circulation similar to that in Figure 5, but the azimuthal velocity is a little curious. If this picture is compared to the cylinder (Figure 4), it can be seen that, topologically, the equator at the bottom of the convection zone is equivalent to the sidewalls. It may also be remembered that the angular momentum is destroyed as the fluid rubs against the wall. Here, in this inviscid model (except for ν_{eddy} in the convection zone), the angular momentum "accumulates" at the equator and produces a jet under the convection zone. We do not know what happens then but suspect that the fate of the jet is involved with the instabilities mentioned next.

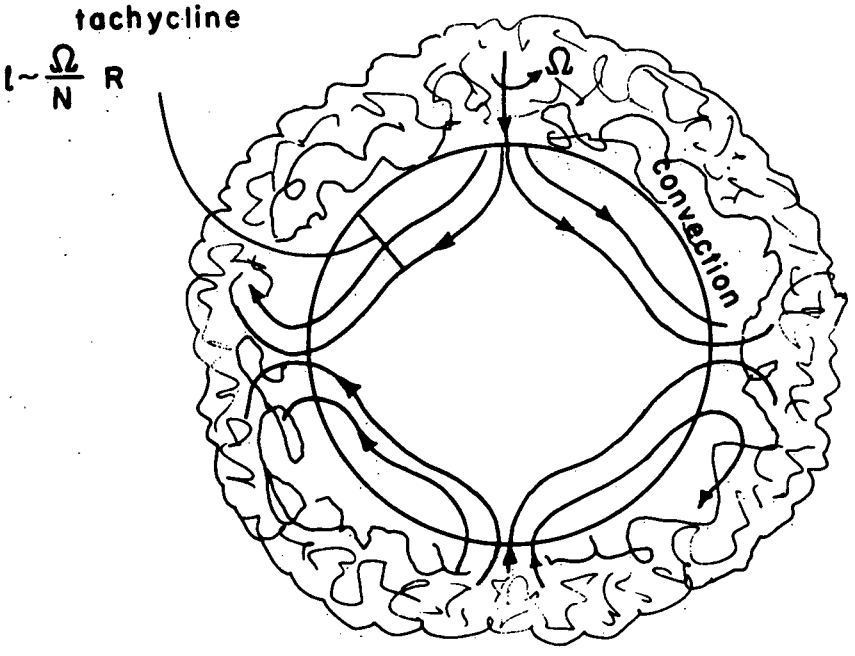


Figure 5.—Solar circulation pattern in the meridional planes.

VI. EFFECTS OF INSTABILITIES

When Temesvary discussed the internal motions of the Sun, he pictured that the effect would produce an instability in the Sun which would eat its way in. Instabilities are also the basis of the criticisms leveled at Dicke's models by Goldreich and Schubert (1967) and Fricke (1968). In our spin-down discussion, Howard, Moore, and I suggested that the seat of instability would be in the tachycline. The picture there was that on a fairly long time scale, the tachycline would retain its form, but the amplitude of the velocity difference across it would grow until an instability would necessarily arise.

From all these discussions, it is not clear what form of instability would predominate, nor is it clear what exactly would happen when the instability started. Qualitative suggestions can be made, but no model calculations exist, as they do for what we have discussed so far. In the previous considerations, essentially all that was discussed was based on fairly precise solutions for simplified models. One may not like all the simplifications, but at least such models give confidence that the process discussed can occur in appropriate circumstances. However, once instability sets in, we are on no such firm ground. For this reason, I will not give details, and only some of their more important aspects will be mentioned.

The question as to which instability strikes first is not closed. Zahn and I recently made a listing of the ones we think may be relevant (Spiegel and Zahn, 1970). It seems that the most likely candidate is the so-called inflection point instability discussed by Lord Rayleigh (1955). This instability is one that occurs when the velocity field has an inflection point in space. Of course, in the tachycline we must include the effects of rotation and stratification, but even then an analogy to the inflection-point instability exists in a form of baroclinic instability (Pedlosky, 1964). The discussion of the effect of this instability on a reasonably correct tachycline model (jet and all) has not been carried out. There is no doubt that the tachycline becomes unstable; the questions are when and in what way. Crude estimates seem to indicate that the tachycline is almost always unstable, but this is not established.

What the instabilities do to the tachycline is not clear. If they are sufficiently weak, they probably distort it in a nonaxisymmetric way; but as the shear across the layer builds, a weak turbulence probably develops. The turbulence transports out the angular momentum carried off by the solar wind; however, a more complicated process might occur.

If the tachycline is weakly turbulent, we can associate with it an eddy viscosity. This eddy viscosity may then keep the tachycline nearly rigid and cause it to pump fluid farther into the interior. That is, the turbulent tachyline causes the fluid just inside it to spin down in a second thin layer. I anticipate that this second layer becomes unstable too and spins down a third layer. In this way, layer by layer, the Sun spins down. As one goes farther into the Sun, the local value of $N \approx \sqrt{G\bar{\rho}}$ becomes larger since $\bar{\rho}$ is the mean density interior to the radius considered. Thus, the layers become thinner; but, accordingly, they become more and more unstable. Crude estimates for how long this process takes to reach the center indicate 10 to 10^3 years, but at the present state of understanding, we cannot attach much significance to such estimates. In any case, once the process has reached the center, it would tend to maintain a weak turbulence in the Sun over much of its main-sequence lifetime. Once the braking is slow enough, however, the process should die out. Once it does, a small differential rotation becomes possible, but a variation in Ω from surface to center of an order of magnitude seems unlikely from this turbulent-onion model of the Sun.

In this inconclusive tone, three aspects of the problem should be raised that are of interest for solar physics and which provide the basis of a continuing interest in the solar spin-down problem.

A. The Lithium Problem

There is good observational evidence that the surface lithium abundance of solar-type stars decreases with time with a half-life of the order of 10^9 years. Destruction of lithium by thermonuclear process requires that it be raised to a temperature of about 2.3×10^6 K. Present estimates of the depth of the convection indicate that the temperature at the bottom of the zone is less than this value. On the other hand, the most promising explanation for lithium depletion is that it is slowly mixed to depths at which the temperature is large enough to destroy it. Various processes have been considered (e.g., Spiegel, 1968). Perhaps the most evident is that the convective motions penetrate slowly into the stable regions and do the mixing. Such calculations as exist indicate that this is not the case; but the matter is not closed. It is also possible that the unstable spin-down currents do the job, and this makes it interesting to understand spin-down. Similar questions exist for beryllium.*

*For a discussion of this aspect, see Chapter 2, "Internal Rotation of the Sun", by Dicke, and Chapter 11, "Origin of the Solar System", by Schatzman.

B. Magnetic Fields

If a strong magnetic field exists in the solar interior, the discussion of the interior motions must be severely modified. The usual argument against a strong interior field is that the surface field seems to flip every 11 years. If this field were attached to a strong internal field, such a short time scale would not be possible. Thus, a strong internal field would have to be detached from the field of the convection, or at least it would have to be detached and reattached every 11 years. No definite model of this kind has been introduced, and even if it were, it would face stability problems (Schubert, 1968; Goldreich and Schubert, 1968). Nevertheless, it seems quite likely that there is some kind of nonnegligible magnetic field in the solar interior.

One reason for believing that fields exist in the solar interior is that the mixing of lithium somewhat into the "stable" regions of the Sun implies an interchange of matter. Material descending from the convection zone should pull magnetic fields with it. Initially, such a field would be passive, in that it is completely controlled by the motion; but as the field stretches out, it will strengthen and feed back on the motion. It is likely that the strong field becomes unstable and buckles into a small scale and is dissipated. The picture is thus rather like many current ones of solar activity, but the difference is that the process takes place under the convection zone and is driven by spin-down currents. Moore and I tried to make a crude model of this process; the main difficulty was that the manner in which the field behaved after it became unstable was unknown. However, the qualitative conclusion was that the convection zone field and the interior field might couple in a relaxation oscillation with a period equal to the turbulent spin-down time. One outcome of these considerations is that the surface angular velocity should reflect these oscillations, with a very small amplitude. Such a variation, if detected, would provide an excellent clue to the spin-down time.

C. Mixing of Helium

As the reactions in the solar core convert hydrogen into helium, the Sun slowly evolves, because the hydrogen supply in the core diminishes; but if spin-down currents were powerful enough, they could replenish the hydrogen supply in the core and thus alter the course of solar evolution. (Howard, Moore, and Spiegel, 1967). Such a modification of the nuclear history of the Sun bears strongly on the problem of detectability of solar neutrinos (Ezer and Cameron, 1968). It seems

likely that helium mixing took place when the Sun first reached the main sequence; but as the Sun slows down, this process will choke off, primarily at the center. Bahcall, Bahcall, and Ulrich (1969) suggest that the helium mixing could not last for an appreciable time, since a gradient of chemical composition would inhibit the mixing process. However, they did not consider baroclinic instability, which could work even in the presence of such gradients when the rotation is strong enough. Thus, this important question is unsettled and, as with the problem of solar rotation, hangs on the occurrence and final development of complicated instabilities.

REFERENCES

Bahcall, Bahcall, and Ulrich, *Astrophys. J.*, 1969.

Bondi, H., and Lyttleton, R. A., "On the Dynamical Theory of the Rotation of the Earth", *Proc. Cambridge Phil. Soc.* 44:345-359, 1948.

Brandt, John C., *Introduction to the Solar Wind* (from a series of books in astronomy and astrophysics by Geoffrey and Margaret Burbidge), W. H. Freeman & Co., San Francisco, 1970.

Bretherton, F. P., and Spiegel, E., "Effect of the Convection Zone on Solar Spin Down", *Astrophys. J. (Letters)* 153:L77, 1968.

Chandrasekhar, S., *Hydrodynamic and Hydromagnetic Stability*, Clarendon Press, Oxford, 1961.

Conti, P., "The Li/Be Ratio in Main-Sequence Stars", *Astrophys. J.* 151:567, 1968.

Cowling, T. G., *Observatory* 89:217-224, 1969.

Dicke, R. H., "Sun's Rotation and Relativity", *Nature* 202:432, 1964.

Dicke, R. H., *Astrophys. J. (Letters)* 149:L121, 1967.

Einstein, A., *Mein Weltbild*, Querdo-Verlag, Amsterdam, 1934.

Ezer, D., and Cameron, A. G. W., "Solar Spin Down and Neutrino Fluxes", *Astrophys. Lett.* 1:177-179, 1968.

Fricke, K., "Instabilität Stationärer Rotation in Sternen", *Z. Astrophys.* 68:317, 1968.

Goldreich, P., and Schubert, G., *Astrophys. J.* 150:571, 1967.

Goldreich, P., and Schubert, G., "The Theoretical Upper Bound to the Solar Oblateness", *Astrophys. J.* 154:1005, 1968.

Graham, E., doctoral thesis, University of Manchester, 1969.

Greenspan, H. P., and Howard, L. N., "On a Time-Dependent Motion of a Rotating Field", *J. Fluid Mech.* 17:385, 1963.

Greenspan, H. P., *The Theory of Rotating Fluids*, Cambridge University Press, London, 1968.

Holton, J. R., "Influence of Viscous Boundary Layers on Transient Motions in a Stratified Rotating Fluid", Part I, *J. Atmos. Sci.* 22:402, 1965.

Howard, L. N., Moore, D. W., and Spiegel, E., "Solar Spin Down Problem", *Nature* 214:1297, 1967.

Iben, I., in *Stellar Evolution,* R. F. Stein and A. G. W. Cameron, eds., Plenum Press, New York, 1966.

Kraft, R., "Studies of Stellar Rotation: V. The Dependence of Rotation on Age Among Solar-Type Stars", *Astrophys. J.* 150:551, 1967.

Kraft, R., in *Spectroscopic Astrophysics* (memorial volume to Struve), G. Herbig, ed., University of California Press, Berkeley, 1970, pp. 385-422.

Larsen, R. B., "Numerical Calculations of the Dynamics of a Collapsing Protostar", *Mon. Notic. Roy. Astron. Soc.* 145:271-295, 1969.

Pedlosky, J., "Stability of Currents in the Atmosphere and the Ocean", *J. Atmos. Sci.* 21:201-219 (Part 1), 21:342-353 (Part 2), 1964.

Pedlosky, J., "Spin Up of a Stratified Fluid", *J. Fluid Mech.* 28:463, 1967.

Prandtl, L. J., "Fluid Mechanics", *Essentials of Fluid Dynamics,* Blackie & Sons, Ltd., London, 1952.

Rayleigh, Lord, in *The Theory of Hydrodynamic Stability,* by C. C. Lin, Cambridge University Press, 1955.

Sakurai, T., "Spin-Down of Boussinesq Fluid in a Circular Cylinder", *J. Phys. Soc. Jap.* 26:840, 1969.

Schubert, G., "The Stability of Toroidal Magnetic Fields in Stellar Interiors", *Astrophys. J.* 151:1099, 1968.

Spiegel, E. A., "The Solar Atmospheric Convection Zone: Theory of Turbulent Convection", *Proc. IAU Symp.* 28:347, 1967.

Spiegel, E. A., in *Highlights of Astronomy,* L. Perke, ed., IAU General Assembly XIII, 1968, p. 261.

Spiegel, E. A., and Zahn, J. P., "Instabilities of Differential Rotation, September-October 1970", *Comments on Astrophysics and Space Science* 2(5):178-183, 1970.

Temesvary, von S., "Der Rotationszustand der Sonne", *Z. Naturforsch.* 7(a):103-120, 1952.

Walin, in *Tellus,* 1968.

Page Intentionally Left Blank

CHAPTER 4
DYNAMICS OF THE OUTER SOLAR ATMOSPHERE*

A. J. Hundhausen**
University of California
Los Alamos Scientific Laboratory
Los Alamos, New Mexico

I. INTRODUCTION

The study of the outer solar atmosphere has traditionally involved the chromosphere, the corona, and the transition zone between these two regions. Within the past 15 years, it has been recognized that the interplanetary region, pervaded by solar plasma and magnetic fields, is a direct extension of the solar atmosphere. The present discussion will concentrate on the dynamical processes which produce the corona and its expansion into interplanetary space, the resulting relationship between solar and interplanetary phenomena, and the use of interplanetary observations to infer coronal properties.

II. PHYSICAL CONDITIONS IN THE OUTER SOLAR ATMOSPHERE

Solar and interplanetary observations are reviewed elsewhere in this volume.† Thus, only a brief statement of some pertinent coronal, chromospheric, and interplanetary properties need be given here.

A. The Corona

The distinguishing characteristic of the corona is its high temperature; several independent techniques give values in the range 1.5×10^6 to 2.0×10^6 K. Much of the following discussion will concern the coronal temperature, both as the effect of

*Research performed under the auspices of the United States Atomic Energy Commission.
**Present address: High Altitude Observatory, Boulder, Colorado.
†See Chapter 1, "Introduction to Solar Physics", by Brandt, and Chapter 5, "The Interplanetary Plasma", by Ogilvie.

basic physical processes and as the cause of the coronal expansion into interplanetary space. Observations indicate that 10^6-K temperatures are attained within $\approx$ 5000 km above the photosphere and extend outward to a heliocentric distance of several solar radii. The electron density at the base of the corona is 10^8 to 10^9 cm^{-3}, diminishing outward with a scale height of approximately 0.1 solar radius.

B. The Chromosphere and Transition Region

The increase in temperature from the 4500-K value at the photospheric boundary (defined as the altitude at which the optical depth in the continuum is 1) to the 10^6-K value characteristic of the corona occurs in a thin layer of considerable complexity. The lower part of this layer gives rise to the strong line emission observed as the "flash spectrum" just before and after totality of a solar eclipse. The traditional term "chromosphere" is now generally applied to this lowest region, extending $\approx$ 1000 km above the photosphere. The temperature in the chromosphere rises slowly with height to $\approx$ 5500 K. The electron density falls off rapidly (a scale height at such temperatures is $\approx$ 100 km) from $\approx 10^{16}$ cm^{-3} at the photospheric boundary to $\approx 10^{12}$ cm^{-3} at the top of the chromosphere. The region between 1000 and 5000 km above the photosphere, in which the major part of the rise to coronal temperatures occurs, is generally referred to as the "transition zone"; the electron density must decrease by three to four more orders of magnitude therein. Both hydrogen and helium become ionized in the transition zone.

The approximate nature of the chromospheric and transition zone properties cited above stems partly from the inhomogeneous nature of this part of the solar atmosphere; division into distinct regions where physical parameters vary only with altitude is a gross idealization. The magnetic field is observed to be high at the boundaries of well-defined cells (supergranules) with dimensions near 3×10^4 km. Spicules, or jets of material moving upward at $\approx$ 20 $km\text{-}s^{-1}$, extend from these boundaries along magnetic field lines, from the top of the chromosphere, through the transition zone, and into the lower corona. The temperature and density within these transient structures (lasting $\approx$ 15 minutes) are $\approx 5 \times 10^4$ K and $\approx 10^{11}$ cm^{-3}.

C. The Interplanetary Medium

The direct accessibility of the interplanetary region to spacecraft-borne instruments has allowed rather precise determination of the physical properties in

this region of the solar atmosphere. Table 1 lists a number of these properties under quiet or undisturbed conditions. The characteristic feature of the ionized plasma which pervades interplanetary space is its rapid (in fact, supersonic) motion away from the Sun. This flow of material implies an energy transport away from the Sun; because this transport will be of interest later in the discussion, some relevant flux and energy densities are given in Table 2. The electron and proton heat conduction fluxes are based on actual computations from observed interplanetary distribution functions.

Table 1.—Observed properties in the quiet solar wind at 1 AU.*

Property	Observed Value
Flow speed (nearly radial from Sun)	320 $km\text{-}s^{-1}$
Proton or electron density	8 cm^{-3}
Proton temperature	4×10^4 K
Electron temperature	1.5×10^5 K
Magnetic field intensity	5×10^{-5} gauss

*Particle properties are based on Hundhausen et al. (1970) and Montgomery et al. (1968).

Table 2.—Flux and energy densities in the quiet solar wind at 1 AU.*

Parameter	Value
Proton flux density	2.4×10^8 $cm^{-2}\text{-}s^{-1}$
Kinetic energy flux density	0.22 $erg\text{-}cm^{-2}\text{-}s^{-1}$
Electron heat conduction flux density	≈ 0.01 $erg\text{-}cm^{-2}\text{-}s^{-1}$
Proton heat conduction flux density	$\approx 10^{-5}$ $erg\text{-}cm^{-2}\text{-}s^{-1}$
Kinetic energy density	7×10^{-9} $erg\text{-}cm^{-3}$
Proton thermal energy density	6×10^{-11} $erg\text{-}cm^{-3}$
Electron thermal energy density	1.5×10^{-10} $erg\text{-}cm^{-3}$
Magnetic field energy density	10^{-10} $erg\text{-}cm^{-3}$

*The heat conduction fluxes are directly observed; see Montgomery et al. (1968) and Hundhausen et al. (1967).

The interplanetary observations summarized in the tables were performed near 1 AU. Very little detailed information has been obtained regarding the variation of interplanetary plasma properties with heliocentric distance. The flow speed observed on the Mariner 2 probe to Venus was nearly constant, while the density varied approximately as r^{-2} (Neugebauer and Snyder, 1966) between 1 AU and the orbit of Venus (0.7 AU).

III. PHYSICAL PROCESSES IN THE OUTER SOLAR ATMOSPHERE

The conventional theoretical treatment of solar structure (e.g., in the solar interior or photosphere) is based on the assumption of hydrostatic equilibrium. Our treatment of the outer solar atmosphere will also be based on fluid precepts but must be hydrodynamic in order to account for the flow of material observed in the interplanetary region. To simplify the discussion, the atmosphere will be assumed to be spherically symmetric, in steady motion, and composed of completely ionized hydrogen (relaxation of these assumptions will be considered later). Our results thus apply to a Sun with no spatial structure or temporal change (i.e., in the absence of any solar activity). If viscous and magnetic forces are neglected, the hydrodynamic equations are the equation of mass conservation

$$\frac{1}{r^2}\frac{d}{dr}(r^2 nmu) = 0\,, \tag{1}$$

the equation of momentum conservation

$$nmu\frac{du}{dr} = -\frac{dP}{dr} - nm\frac{GM_\odot}{r^2}\,, \tag{2}$$

the equation of energy conservation

$$\frac{1}{r^2}\frac{d}{dr}\left[r^2 nu\left(\frac{1}{2}mu^2 + \frac{3P}{2n}\right)\right] = -\frac{1}{r^2}\frac{d}{dr}(r^2 uP) - nmu\frac{GM_\odot}{r^2} + S(r)\,, \tag{3}$$

and the equation of state

$$P = nk(T_i + T_e)\,, \tag{4}$$

where

r = heliocentric distance,
n = the number density of the electrons and protons (essentially equal because the Debye length in the solar atmosphere is very small),
u = the radial flow speed of the electrons and protons (again, essentially equal to avoid building up a net electrical charge on the Sun),
m = the proton mass,
P = the total pressure,
G = the gravitational constant,
$M_\odot$ = the solar mass,
k = the Boltzmann constant,
T_i = the proton temperature,

and

T_e = the electron temperature.

The term $S(r)$ in the energy equation (Equation 3) represents any source of energy; $S(r) = 0$ implies an adiabatic flow.

Equations 1 and 3 have simple first integrals, i.e.,

$$4\pi nur^2 = f \tag{5}$$

and

$$4\pi\left\{nur^2\left[\frac{1}{2}mu^2 + \frac{5}{2}k(T_i + T_e) - \frac{GM_\odot m}{r}\right] - \int_{r_0}^{r} r^2 S(r)\,dr\right\} = F\,, \tag{6}$$

where r_0 is some reference heliocentric distance. The constants f and F are the mass and energy flux, respectively, through any spherical surface centered at the Sun.

A. Energy Source Terms in the Outer Solar Atmosphere

Solution of the system of equations given above requires specification of the energy source term $S(r)$. Two contributions to $S(r)$ should obviously be included.

1. Heat Conduction

A hot ionized gas is an extremely efficient heat conductor. We thus expect a large heat conduction flux away from the hot corona into the neighboring, lower

temperature, chromospheric and interplanetary regions:

$$F_c = -4\pi r^2 \kappa \frac{dT}{dr},$$

where κ is the thermal conductivity. The source term (for the plasma) in Equation 3 is the negative of the divergence of this flux:

$$S_c = \frac{1}{r^2}\frac{d}{dr}\left(r^2 \kappa \frac{dT}{dr}\right).$$

Heat conduction in an ionized hydrogen plasma is dominated by the fast-moving electrons. If the electrons interact by Coulomb collisions, the electron conductivity is given by

$$\kappa_e = \kappa_s T_e^{5/2} \text{ erg-cm}^{-1}\text{-s}^{-1}\text{-K}^{-1},$$

where $\kappa_s = (1.84 \times 10^{-5})/\ln \Lambda$, $\ln \Lambda$ being the usual "Coulomb logarithm" (see Spitzer, 1962). Under coronal and interplanetary conditions, $\ln \Lambda \approx 23$ and $\kappa_s \approx 8 \times 10^{-7}$ erg-cm^{-1}-s^{-1}-K$^{-7/2}$. At a given temperature, the proton conductivity κ_p is smaller by the square root of the ratio of electron and ion masses.

The thermal conductivity law given above has been extensively used in models of the outer solar atmosphere. However, we note the following limitations on its applicability:

(1) In the presence of a magnetic field, the usual conductivity law applies only parallel to the field lines. The conductivity transverse to the field lines is diminished by a factor $(\omega\tau)^2$, where ω is the gyrofrequency and τ is the mean collision time of the particles. For electrons in the corona, $\omega\tau \approx 10^5$, and thus the transverse conductivity $\kappa_\perp$ is only $\approx 10^{-10}$ of the parallel conductivity $\kappa_\parallel$. A similar situation is found in the interplanetary region.

(2) Heat conduction results from the skewing of particle distribution functions in the presence of a temperature gradient. The usual conductivity law applies only when Coulomb collisions occur frequently enough to keep the skewness small. This requirement can be stated in terms of a dimensionless parameter (see Spitzer and Harm, 1953)

$$B_T = \frac{2T_e}{\pi n q^4 \ln \Lambda}\frac{dT_e}{dr},$$

where q is the electronic charge. B_T is, in fact, the ratio of the mean free path for Coulomb collisions to the scale length of the temperature gradient. The law $\kappa = \kappa_s T^{5/2}$ applies for $|B_T| \ll 1$, or when there are many collisions in the distance over which the temperature changes. An upper limit to the condition flux density can also be given by the situation in which all electrons move in the direction of the heat flow at the mean thermal speed v_e (Parker, 1964); i.e.,

$$F_c/4\pi r^2 < nv_e \frac{1}{2} mv_e^2 \approx \frac{1}{2} nm(3kT_e/m)^{3/2} .$$

In fact, the $|B_T| \ll 1$ restriction applies long before this limiting conduction flux density can be attained.

2. Radiation

Evaluation of radiative losses generally requires solution of a transfer equation.* Fortunately, the corona and interplanetary region are optically thin in the wavelengths at which most energy is radiated, and a simple radiative loss function can be given for this part of the solar atmosphere. For temperatures above a few times 10^4 K (where hydrogen ceases to be an efficient radiator), the loss rate is nearly independent of temperature. Then,

$$S_r(r) \approx -2 \times 10^{-23} n_e^2 \text{ erg-cm}^{-3}\text{-s}^{-1}$$

(see Kuperus, 1969, or Brandt, 1970). The radiative flux at any heliocentric distance r is

$$F_r(r) = -4\pi \int_{r_\odot}^{r} r^2 S_r(r)\, dr ,$$

where $r_\odot$ denotes the photospheric boundary. Now, $F_r(r_\odot)$ can be set to zero because the optically thin outer atmospheric layers absorb almost none of the photospheric radiation, which is therefore effectively decoupled from the energy equation.

Note that both of the energy source mechanisms given above can transfer energy only from hot regions to cold regions. Insertion of these sources into Equation 3 or Equation 6 and integration outward from r cannot produce the increase in temperature to 10^6 K known to occur between the photosphere and the

*See Chapter 1, "Introduction to Solar Physics", by Brandt.

corona. *In order to explain the high coronal temperature, we must make the* ad hoc *assumption that an additional energy source exists in the outer solar atmosphere.* This heating process, to be discussed in Section III.C, is generally thought to involve the dissipation of mechanical energy transported from underlying layers of the Sun by some type of waves. The energy source term for this process will be denoted by $S_m(r)$. The total flux of energy in this mysterious form is

$$F_m = F_{m\odot} - 4\pi \int_{r_\odot}^{r} r^2 S_m(r)\, dr ,$$

where $F_{m\odot}$ is the flux through the photospheric boundary at $r_\odot$.

With the inclusion of the heat conduction, radiation, and the *ad hoc* mechanical heating term, the energy conservation equation becomes

$$\frac{1}{r^2}\frac{d}{dr}\left[r^2 nu\left(\frac{1}{2}mu^2 + \frac{5}{2}\frac{P}{n} - \frac{GM_\odot m}{r}\right)\right] = \frac{1}{r^2}\frac{d}{dr}\left(r^2 \kappa_e \frac{dT_e}{dr}\right) + S_r(r) + S_m(r) . \quad (7)$$

The energy conservation integral becomes (with the reference level now fixed at $r_\odot$)

$$F = 4\pi\left\{nur^2\left[\frac{1}{2}mu^2 + \frac{5}{2}k(T_i + T_e) - \frac{GM_\odot m}{r}\right] - r^2\kappa_e\frac{dT_e}{dr}\right.$$
$$\left. - \int_{r_\odot}^{r} r^2 S_r(r)\, dr + \frac{F_{m\odot}}{4\pi} - \int_{r_\odot}^{r} r^2 S_m(r)\, dr\right\} . \quad (8)$$

B. Classification of Atmospheric Regions by the Influence of Energy Source Terms

The solution of the mass, momentum, and energy conservation equations with the three energy source terms described in Section III.A is a formidable task. Fortunately, not all of the terms in the energy equation are of equal importance in various regions of the solar atmosphere, and considerable simplification can be achieved by proper neglect of some terms. As the foundation for such approxima-

tions and as an aid in understanding the physical nature of various atmospheric regions, we will now evaluate the significance of the terms in the energy equation (Equation 7 or Equation 8) throughout the outer solar atmosphere. This evaluation and the resulting classification scheme will closely follow those given in a recent review by Kuperus (1969).

The physical bases for this discussion are (1) the continuous decrease in electron density n_e from $\approx 10^{16}$ cm^{-3} at the photospheric boundary to ≈ 10 cm^{-3} at 1 AU, presumably approaching zero as $r \to \infty$, (2) the qualitatively established dependence of electron temperature on heliocentric distance, as shown in Figure 1, and (3) the small outward velocities (in the range 1 to 20 km-s^{-1}) of material observed in the chromosphere and low corona. The qualitative variations with heliocentric distance r of the terms in the energy equation follow directly. Convective transport of energy (given by the terms in square brackets in Equation 7 or Equation 8) must be small below the level where the coronal expansion becomes rapid. Heat conduction will transport energy away from the temperature maximum in the corona, both downward into the chromosphere and outward into interplanetary space. Radiation will be important only in the lower layers, decreasing rapidly with r because of its proportionality to n_e^2.

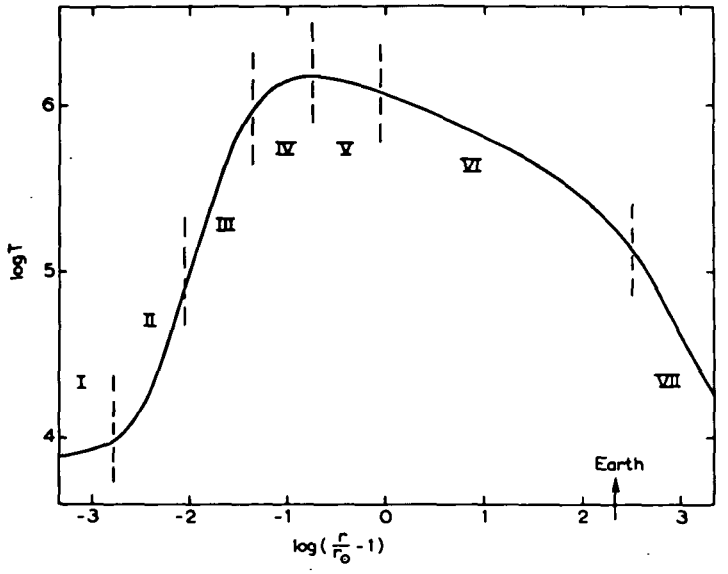


Figure 1.—Qualitative variation of temperature with position in the outer solar atmosphere (Kuperus, 1969). Regions I to VII are defined in text.

On the basis of these qualitative properties of the energy terms, the outer solar atmosphere has been divided by Kopp (1968) and Kuperus (1969) into the seven regions shown in Figure 1. Region I is the chromosphere, from which the radiation loss is large (see Section III.C). Heat conduction into the region is probably negligible because both the temperature and the temperature gradient are small. We are thus forced to postulate that the mechanical dissipation term $S_m(r)$ maintains the chromosphere by balancing the radiation loss. Region II is the lowest part of the transition zone, in which the temperature begins to rise rapidly with r toward the coronal value. The radiation loss is still appreciable. Heat conduction becomes important in this region because of the large temperature gradient. The relative importance of heat conduction and mechanical dissipation in balancing radiation is not clear. Region III is that part of the transition region in which the heat conduction flux has become so large as to dominate both radiation and mechanical heating. The heat conduction flux is nearly constant and the structure of the region is thereby determined (see Section IV).

Region IV marks the beginning of the corona, with the temperature reaching 10^6 K. The dissipation of mechanical energy must maintain the high temperature against a diminishing radiation loss and heat conduction down into the transition zone and chromosphere. The maximum temperature is attained at the boundary between Regions IV and V. In Region V, mechanical dissipation must maintain the high temperature against a still smaller radiation loss, heat conduction outward into interplanetary space, and the convection of energy into the latter region. Whereas Regions I to IV could be treated as in approximate hydrostatic equilibrium, the growing importance of the convection term indicates that a hydrodynamic treatment must be applied in and beyond Region V. Region VI is taken to begin where mechanical dissipation becomes negligible. Heat conduction from the lower, high-temperature regions continues to provide energy for an increasing outward convection. One might divide this region into two parts, coronal and interplanetary, at the heliocentric distance where the hydrodynamic expansion becomes supersonic (see Section V.A). Finally, the arguments of Section III.A indicate that as the electron density approaches zero, so must the heat conduction. Thus, a distant interplanetary zone, Region VII, should ultimately be reached in which the energy flux is entirely due to convection. The reasonable condition that $T \to 0$ as $r \to \infty$ thus results in an energy equation

$$F \to 4\pi nur^2 \frac{mu^2}{2}.$$

The flow becomes adiabatic in this region, so that the temperature should be proportional to $r^{-4/3}$ (see Section V.C).

This classification scheme is summarized in Table 3. The mechanical heating term $S_m(r)$ is seen to be a dominant or significant energy source in all parts of the outer solar atmosphere except for Region III, the transition zone, and Regions VI and VII, the expanding corona and interplanetary medium. In all of the regions where $S_m(r)$ is important, the theoretical determination of atmospheric structure depends directly upon detailed knowledge of the form of this term.

Table 3.–Energy source terms in the outer solar atmosphere.

Region	n (cm^{-3})	T (K)	Convection*	Conduction*	Radiation*	Heating*
I–Chromosphere	10^{13}	5×10^3	N	N	D	D
II–Lower transition zone	10^{12}	10^4	N	D	D	D (?)
III–Transition zone			N	D	S	S (?)
IV–Inner corona	10^8	10^6	N	D	S	D
V–Corona			S	D	S (?)	D
VI–Corona and interplanetary medium	10 (at 1 AU)	10^5 (at 1 AU)	D	D to N	N	N
VII–Distant interplanetary medium	($\sim 1/r^2$)	($\sim 1/r^{4/3}$)	D	N	N	N

*D, S, and N signify dominant, significant, and negligible processes, respectively.

C. Observational Information Regarding the Mechanical Dissipation Term

The mechanical heating term $S_m(r)$ was postulated in Section II.A to explain the existence of a high coronal temperature. We have found in Section II.B that this term may play a significant role in determining the structure of the chromosphere, lower transition zone, and much of the corona. Observations of the structure of these regions could be used to derive the source function $S_m(r)$. Unfortunately, sufficiently detailed observations are not available to permit such an empirical determination. One can, in fact, only roughly determine the total rate of mechanical energy dissipation in the outer solar atmosphere. This required energy input can then serve as a constraint upon theoretical models of the energy source, transport, and dissipation processes.

The energy flux equation (Equation 8) at the photospheric boundary is dominated by the gravitational and mechanical terms:

$$F \approx -4\pi nur^2 \frac{GM_\odot m}{r} + F_{m\odot} . \tag{9}$$

At the heliocentric distance r_e of the Earth's orbit,

$$F = 4\pi \left\{ nur^2 \left[\frac{1}{2} mu^2 + \frac{5}{2} k(T_i + T_e) - \frac{GM_\odot m}{r} \right] - r^2 \kappa_e \frac{dT_e}{dr} \right\}_{r=r_e}$$

$$- 4\pi \int_{r_\odot}^{r_e} S_{\text{rad}}(r)\, dr + F_{m\odot} - 4\pi \int_{r_\odot}^{r_e} r^2 S_m(r)\, dr .$$

The convection and heat conduction fluxes can be evaluated using Tables 1 and 2. Actually, the flux of kinetic energy is found to dominate completely these terms (note in particular that the heat conduction flux is only ≈ 5 percent of the total) so that to a good approximation the total flux at r_e is

$$F \approx 4\pi \left(\frac{1}{2}\right) mn_e u_e^3 r_e^2 - 4\pi \int_{r_\odot}^{r_e} S_{\text{rad}}(r)\, dr + F_{m\odot} - 4\pi \int_{r_\odot}^{r_e} r^2 S_m(r)\, dr . \tag{10}$$

Combining Equations 9 and 10 gives the total dissipation of mechanical energy between $r_\odot$ and r_e in terms of observable or calculable quantities (recalling that nur^2 is constant):

$$4\pi \int_{r_\odot}^{r_e} r^2 S_m(r)\, dr = 4\pi nur^2 \left(\frac{1}{2} mu_e^2 + \frac{GM_\odot m}{r_\odot} \right) - 4\pi \int_{r_\odot}^{r_e} r^2 S_{\text{rad}}(r)\, dr ; \tag{11}$$

i.e., the mechanical dissipation ultimately balances the energy loss in the coronal expansion (including the correction for work done against solar gravity and radiation to interplanetary space).

The terms on the right-hand side of Equation 9 can be evaluated with little difficulty using Table 2:

$$4\pi nur^2\left(\frac{1}{2}mu_e^2\right) = 6.2 \times 10^{26}\ \text{erg-s}^{-1}$$

and

$$4\pi nur^2 \frac{GM_\odot m}{r_\odot} = 21.7 \times 10^{26}\ \text{erg-s}^{-1}\,.$$

The total rate of energy loss in the coronal expansion is then 2.8×10^{27} erg-s^{-1}; about 75 percent of this energy is consumed in lifting the expanding material out of the solar gravitational field. The radiation loss from the corona can be estimated using the approximate $S_r(r)$ given in Section II.A and the electron density

$$n_e = n_{e\odot} \exp\left[-\frac{GM_\odot m}{2kT_\odot}\left(\frac{1}{r} - \frac{1}{r_\odot}\right)\right]$$

derived by assuming hydrostatic equilibrium at constant temperature $T_\odot$.* Then (Kuperus, 1969),

$$4\pi \int_{r_\odot}^{r_e} r^2 S_r(r)\, dr \approx 4\pi r_\odot^2 S_r(r_\odot) \frac{kT_\odot r_\odot^2}{GM_\odot m}$$

(i.e., most of the radiation loss occurs within the first scale height of $r_\odot$). At $T_\odot = 1.5 \times 10^6$ K,

$$4\pi \int_{r_\odot}^{r_e} r^2 S_r(r)\, dr \approx 3.3 \times 10^{27}\ \text{erg-s}^{-1}\,.$$

The corona is thus seen to lose energy to the interplanetary region by radiation and by expansion (including outward heat conduction) at nearly equal rates.

The radiation loss from the chromosphere and transition zone cannot be obtained from the simple form of $S_r(r)$ used above. Athay (1966) has estimated this loss to be 3.4×10^{29} erg-s^{-1} (mostly as H^- continuum radiation from the lowest

*See Chapter 1, "Introduction to Solar Physics", by Brandt.

500 km of the chromosphere). This loss is balanced both by local mechanical dissipation and by heat conduction from the overlying coronal layers. The heat conduction flux downward from the corona is approximately 2×10^{28} erg-s^{-1} (Kuperus, 1969).

The resulting energy balance of the outer solar atmosphere is shown in Figure 2. The total energy loss from the entire region is about 3.5×10^{29} erg-s^{-1}. This loss

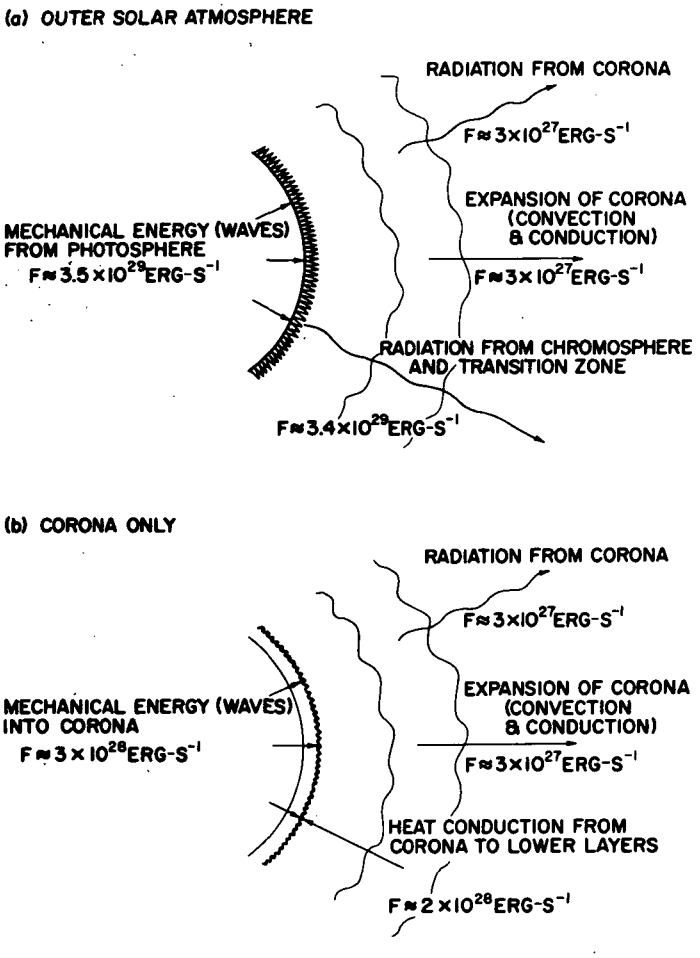


Figure 2.—Estimated energy balance for (a) the entire outer solar atmosphere, and (b) the corona only.

must be balanced by the total mechanical dissipation between the photospheric boundary and 1 AU;

$$4\pi \int_{r_\odot}^{r_e} r^2 S_m(r)\, dr \approx 3.5 \times 10^{29} \text{ erg-s}^{-1}.$$

A corresponding flux of mechanical energy (presumably in waves) must pass through the photospheric boundary (the flux density is $F/4\pi r_\odot{}^2 = 5.7 \times 10^6 \text{ erg-cm}^{-2}\text{-s}^{-1}$). This is a minute fraction of the 4×10^{33} erg-s^{-1} radiated from the photosphere; it is of significance in the outer atmosphere because the latter is so effectively decoupled from the photospheric radiation. The corona loses about 6×10^{27} erg-s^{-1} directly to interplanetary space and about 2×10^{28} erg-s^{-1} by heat conduction to lower layers. The corona must then gain energy by local mechanical dissipation at the rate of $\approx 3 \times 10^{28}$ erg-s^{-1} (Figure 2). The chromosphere and transition zone lose 3.4×10^{29} erg-s^{-1} by radiation and gain 2×10^{28} erg-s^{-1} by heat conduction from the corona. The difference, 3.2×10^{29} erg-s^{-1}, must be gained by local mechanical dissipation.

The energy balance given above illustrates the key role played by radiation in determining the gross structure of the outer solar atmosphere. Although approximately 90 percent of the total dissipation of mechanical energy takes place in the chromosphere and transition zone, the temperature of these layers remains low because the density is high enough to permit efficient removal of energy by radiation. The 10 percent of energy dissipation that occurs in the corona cannot be a balanced by radiation loss because of the low coronal density. The corona must then attain a high temperature in order for the other available energy loss mechanisms (heat conduction and convection) to remove energy at the rate required to maintain a steady state. In fact, about 70 percent of the energy loss from the corona is by heat conduction to the cooler, lower layers, which in turn radiate this energy into interplanetary space.

D. Heating of the Outer Solar Atmosphere by Acoustic Waves

The basic nature of the mechanism which heats the outer solar atmosphere—the dissipation of mechanical energy transported from lower solar layers by waves—is

almost universally accepted. The specific nature of the mechanism—the wave mode involved and the actual physical dissipation process—are, however, still poorly understood. For a review of the various possibilities and the arguments for and against each, see Kuperus (1969). Our discussion will be confined to one possible mechanism based on ordinary sound waves. This mechanism has received much attention and has been chosen for presentation both because of strong plausibility arguments for its applicability to the outer solar atmosphere and because of the general familiarity of most readers with sound waves. The discussion will, in any case, illustrate the features of (and difficulties encountered in) any such theory of mechanical energy transport and dissipation.

1. Generation of Acoustic Noise in the Solar Convection Zone

It is generally accepted that a zone of convective energy transport develops in the outer reaches of the solar interior ($r \approx 0.85 r_\odot$) and extends essentially into the photosphere.* The flow in the upper part of this convection zone is expected to become turbulent, and some energy will be converted into random fluctuations (i.e., noise). The convection zone is then a possible source of a mechanical energy flux which can hopefully be transmitted to and dissipated in the outer atmospheric layers. Figure 3 shows the acoustic power which would be generated by isotropic turbulence in the model convection zone of Vitense (1953). Most of the acoustic noise is produced in a thin layer (≈ 60 km thick) near the top of the photosphere. The flux density of acoustic energy from the entire layer is predicted to be $F_a/4\pi r_\odot^2 \approx 3 \times 10^7$ erg-cm^{-2}-s^{-1} (Kuperus, 1969); this flux is clearly sufficient to maintain the losses from the outer solar layers ($F_m/4\pi r_\odot^2 \approx 6 \times 10^6$ erg-cm^{-2}-s^{-1} from Section III.C). Although this is only an order of magnitude estimate, sound waves generated in the convection zone must be considered as a plausible source of mechanical energy.

2. Propagation of Sound Waves in the Outer Solar Atmosphere

Sound waves are generally treated as small linear perturbations of existing fluid parameters, and as such propagate with very little dissipation. It is clear, however, that sound waves propagating outward through the outer solar atmosphere cannot retain small amplitudes. If there were no dissipation, the flux of acoustic energy F_a would remain constant. The flux density is given by

*See Chapter 1, "Introduction to Solar Physics", by Brandt.

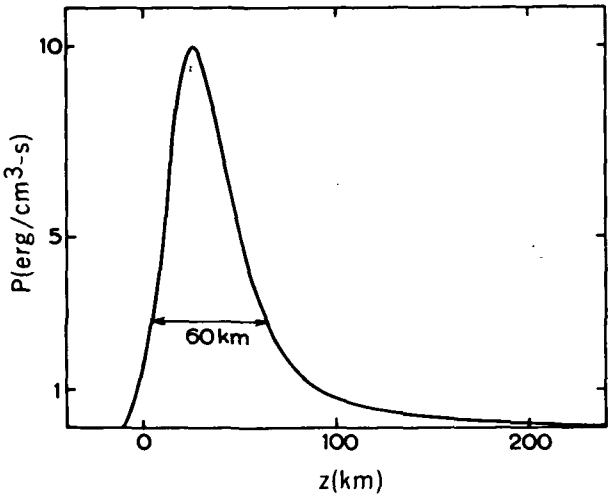


Figure 3.—Rate of generation of acoustical noise by isotropic turbulence in the hydrogen convection zone (Kuperus, 1969). Zero altitude is defined at the level where the optical depth in the continuum is 1.

$$\frac{F_a}{4\pi r_\odot^2} = \frac{1}{2}\rho(\delta u)^2 u_s ,$$

where ρ is the mass density, δu the amplitude of the velocity perturbation in the wave, and u_s the sound speed. In the chromosphere and lower transition zone, the density scale height is ≈ 100 km; thus, in propagating through this region, ρ will decrease by a factor of $\approx 10^5$. A constant flux can be maintained only if δu increases rapidly (u_s also increases, but only by an order of magnitude between the chromosphere and corona). The waves must grow in amplitude and will ultimately become nonlinear. The nonlinear evolution of the waves should lead to the well-known steepening of the wave fronts into shocks.

This growth of wave amplitudes and development of shock waves is quantitatively illustrated in Figure 4, based on a nonlinear computation by Kuperus (1969). A low-amplitude sinusoidal sound wave is introduced into an isothermal atmosphere with $T = 5000$ K ($u_s = 8$ km-s^{-1} for ionized hydrogen). The wave form is shown at the initial altitude and at two higher altitudes. The growth of the wave

amplitude and the steepening of the wave fronts is clearly demonstrated. After propagation through about 5 scale heights, the wave has a sawtooth shape with a series of shock fronts separated by linear relaxations.

Other changes in sound waves will occur in propagation upward from the photosphere. When the rising temperature gradient of the transition zone is encountered, some of the energy in the waves will be refracted or reflected downward. The fraction transmitted is small for ordinary linear sound waves. However, for the steepened, nonlinear wave forms shown to evolve in the chromosphere, the transmission is much larger. A series of shock waves, carrying a sizable flux of acoustic energy, would be expected to reach the transition zone and corona.

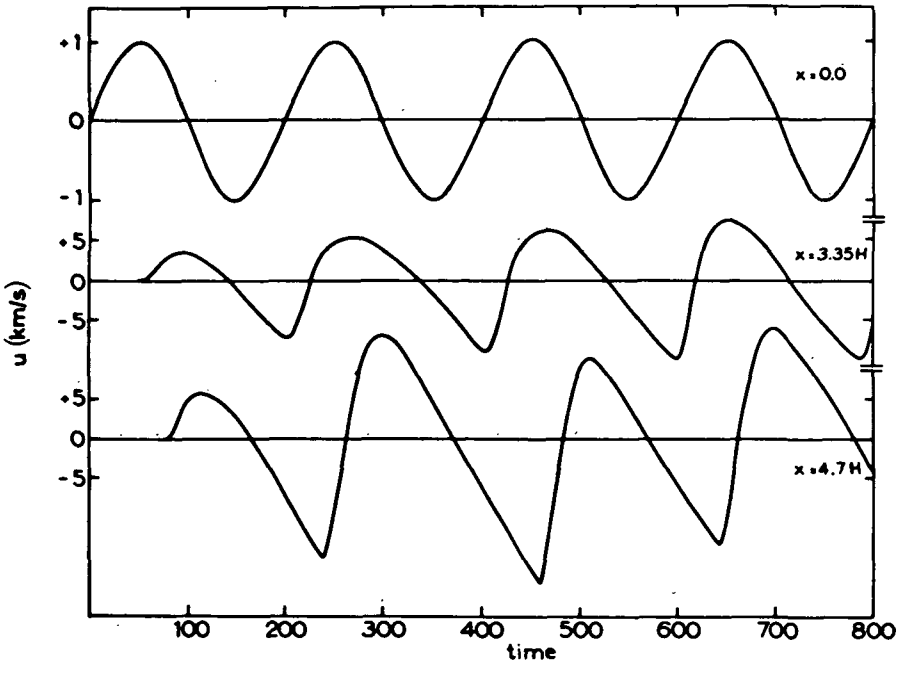


Figure 4.—Development of an ordinary sound wave propagating through an isothermal atmosphere. The wave form is shown at its assumed origin, $x = 0$, and at two higher altitudes, $x = 3.35H$ and $x = 4.7H$, where H is the density scale height. The wave amplitude increases and the wave fronts steepen as the disturbance propagates into lower density material (Kuperus, 1969).

3. *Dissipation of Shock Waves in the Outer Solar Atmosphere*

Some of the kinetic energy of material flowing into a shock front is continuously and irreversibly converted into internal energy of the "shocked" gas. For a series of shocks propagating through an atmosphere, this results in conversion of wave energy into internal energy [i.e., a dissipation of energy as required by the term $S_m(r)$ discussed in Section II.C]. If the arguments given above for the existence of a sizable flux of energy in sound waves and the evolution of these waves into a series of shocks are accepted, a plausible mechanism for energy dissipation results. For a series of weak shock waves, the Rankine-Hugoniot relations for the changes in physical parameters at the shock give an average dissipation rate of

$$S_m(r) = \frac{16}{3} \frac{\rho u_s^2}{\Pi(\gamma + 1)^2} (M_s - 1)^3 , \tag{12}$$

where Π is the period of the shocks, γ is the ratio of specific heats, and M_s is the Mach number of the material flowing into each shock (viewed in the shock frame of reference). For stronger shocks the dissipation rate is higher.

4. *Model Atmospheres Based on Shock Wave Dissipation*

The introduction of an energy source function, even of such seeming simplicity as Equation 12, into the energy Equation 7 or 8 defines a problem of considerable computational difficulty. To illustrate this difficulty, Figure 5 shows the temperature as a function of heliocentric distance predicted by several models, all of which assume heating by shock waves, but each of which makes different simplifying assumptions to facilitate solution of the coupled system of mass, momentum, and energy conservation equations. The models of Bird (1965), which neglect both heat conduction and radiation, predict a maximum coronal temperature at $r = 2r_\odot$, too far away from the Sun to agree with observations. None of the remaining models (see Kuperus, 1969, for references) can be said to be in significantly better or worse agreement with observations than any other. This illustrates the limitations of present coronal observations and our consequent inability to compare the validity of different atmospheric models. Similar difficulties could arise if models based on entirely different theories of energy dissipation were constructed and compared. It should be clear that firm conclusions regarding the nature of the all-important dissipation process cannot be drawn and that the structure of any region of the outer solar atmosphere where this dissipation is significant must remain uncertain.

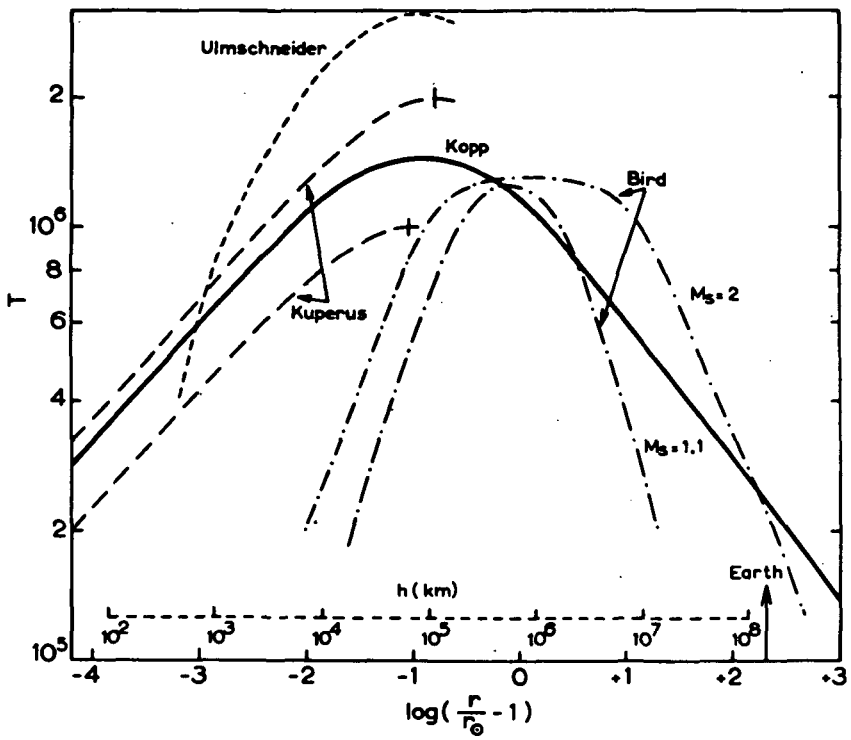


Figure 5.–Model solar atmospheres derived by various authors, using the energy source term due to weak shock waves (Equation 12) but with different treatment of the radiation, conduction, and convection loss terms (Kuperus, 1969).

IV. STRUCTURE OF THE TRANSITION ZONE

In Section III.B it was suggested that the downward heat conduction flux in the transition zone (Region III of the classification scheme) is nearly constant. Kuperus (1969) has estimated this flux to be

$$F_c = -4\pi r^2 \kappa \frac{dT_e}{dr}$$

$$\approx -4\pi r_\odot^2 (2 \times 10^5 \text{ erg-cm}^{-2}\text{-s}^{-1}) . \qquad (13)$$

This is comparable to loss of energy by radiation and presumably to the gain of energy from mechanical dissipation (the integrals of S_r and S_m in Equation 8) in the entire region between the top of the chromosphere and the corona. As the change from chromospheric to coronal temperatures is believed to occur in a small altitude range within this region, Equation 8 implies near constancy of F_c. The structure of the transition zone is then determined by Equation 13. The variation of r^2 can be neglected in a thin layer, and use of the thermal conductivity from Section III.A gives

$$T_e^{5/2} \frac{dT_e}{dr} \approx 2.5 \times 10^{11}\ \mathrm{K^{7/2}\text{-}cm^{-1}}\,. \tag{14}$$

The near constancy of $T^{5/2}dT/dr$ is confirmed by the observed intensities of UV line emission from the transition zone (Athay, 1966).

Equation 14 is easily integrated to give

$$T_e^{7/2} = T_{e0}^{7/2} + 8.7 \times 10^{11}(r - r_0)\,,$$

where the zero subscript denotes parameters at some reference level. The implied temperature gradients are extremely large, especially at the lower temperatures where the conductivity is low. For example, at the altitude where $T = 10^5$ K (sometimes taken to be the "top of the chromosphere"), Equation 14 gives

$$\frac{dT_e}{dr} \approx 10^4\ \mathrm{K\text{-}km^{-1}}\,.$$

Thus, the lower boundary of the transition from the chromosphere to the corona must be very sharp.

If the structure of the transition zone is determined by heat conduction from the corona, that structure will be strongly influenced by the magnetic field configuration, as the conductivity across field lines is small (Section III.A). The chromospheric magnetic field is observed to be high at the supergranule boundaries. The concentration of heat conduction flux at these boundaries results in a strongly inhomogeneous transition zone, as shown in Figure 6 (Kuperus, 1969). The high coronal temperature penetrates to lower altitudes over the supergranule boundaries giving steeper temperature gradients at the level where the change to chromospheric temperatures occurs. The development of spicules at the boundaries has been

attributed (Kuperus and Athay, 1967) to the concentrated energy flux from the corona.

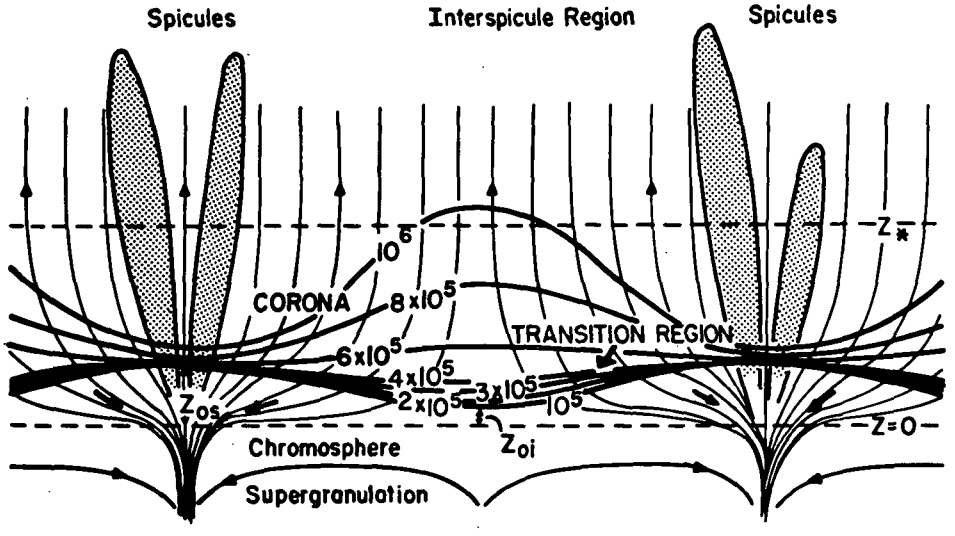


Figure 6.—A model of the transition zone, assuming that heat conduction downward from the corona dominates the energy equation. The concentration of the chromospheric magnetic field at supergranule boundaries leads to a larger conductive flux into the regions above these boundaries and, thus, to a deeper penetration of the high coronal temperature and a steeper temperature gradient (Kuperus, 1969).

V. THE EXPANSION OF THE OUTER SOLAR ATMOSPHERE

The outward flow of material becomes important in the corona and in interplanetary space, i.e., in Regions V, VI, and VII of the classification scheme given in Section III.B. By definition, the dissipation of mechanical energy becomes unimportant in Regions VI and VII. We will concentrate here on quantitative models of the expansion in these two regions; any attempt to treat Region V would involve the dissipation term $S_m(r)$ and encounter uncertainties similar to those already

encountered in Section III.D. The lower boundary of Region VI will be taken as $r = r_\odot$, which is equivalent to assuming that all mechanical dissipation occurs in a thin shell at the base of the corona. The validity of this assumption will be assessed when the resulting models are compared with observations. The loss of energy by radiation can be neglected. Under the assumptions of spherical symmetry, steady flow, and neglect of all magnetic and viscous effects, the fluid equations applicable to Regions VI and VII follow directly from Equations 1 or 5, 2, and 7 or 8.

A. Formal Nature of the Solutions

The fluid equations describing the solar corona have been discussed by Parker (1960, 1963a, 1963b, and 1964), who first demonstrated the existence and relevance of solutions that involve a continuous outward flow of material. Because the basic features of these solutions can be illustrated by the simple situation in which the proton and electron temperatures are taken to be equal and constant, our attention will initially be confined to this case.

If the mass conservation equation (Equation 5) is used to eliminate the density from the momentum equation (Equation 2), the latter can be written in the form

$$\frac{1}{u}\frac{du}{dr}\left(u^2 - \frac{2kT}{m}\right) = \frac{4kT}{mr} - \frac{GM_\odot}{r^2}. \tag{15}$$

For any temperature $T < GM_\odot/mkr_\odot$, the right-hand side of Equation 15 is negative for small r, zero for

$$r_c = GM_\odot m/rkT,$$

and positive for larger r. At $r = r_c$, the left-hand side of the equation can be zero either because

$$u_c^2 = 2kT/m,$$

where the subscript c denotes the value of a parameter at r_c, or because

$$\left.\frac{1}{u}\frac{du}{dr}\right|_{r=r_c} = 0.$$

If attention is limited to continuous solutions with continuous derivatives, the first of these possibilities implies that du/dr has the same sign for all r, so that $u(r)$ is monotonically increasing or decreasing. The second possibility implies that $u^2 - 2kT/m$ has the same sign at all r, so that $u(r)$ has a maximum at r_c if $u^2 < 2kT/m$ or a minimum at r_c if $u^2 > 2kT/m$.

The observation of small material velocities in the low corona eliminates the monotonically decreasing solution and those solutions with a minimum at r_c from applicability to the corona (all such solutions have $u^2 > 2kT/m \approx 100$ km-s^{-1} for $r < r_c$). The remaining solutions are of two types:

(1) A unique solution for which the monotonically increasing $u(r)$ has the value $u_c = (2kT/m)^{1/2}$ at $r = r_c$ (i.e., passes through the so-called "critical point"). Since u_c is the characteristic thermal speed in the fluid, this leads to a supersonic expansion at large heliocentric distances; such an expansion is now generally referred to as the "solar wind". Actual solutions, obtained by numerical integration, are shown in Figure 7 for several coronal temperatures (Parker, 1963a). Flow speeds of several hundred kilometers per second are predicted in interplanetary space.

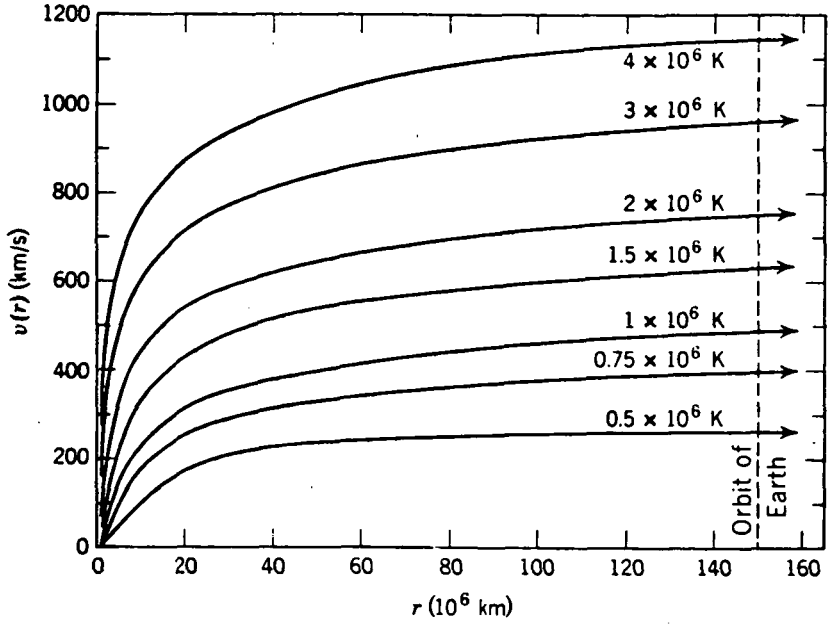


Figure 7.—Coronal expansion speed as a function of heliocentric distance, given by the solar wind model of Parker (1963a) for an isothermal corona.

(2) A family of solutions for which $u(r)$ increases with r until a maximum value smaller than u_c is attained at $r = r_c$, and then decreases, approaching zero as r approaches infinity. The expansion is subsonic at all r. These solutions are similar to the "solar breeze" solutions obtained by Chamberlain (1960) from a more realistic treatment of the energy equation. Flow speeds of tens of kilometers per second are predicted near 1 AU for reasonable coronal temperatures.

Either the solar wind or the solar breeze solutions can exist for a given coronal temperature, but at different coronal densities. The criterion that determines the nature of the expansion for specific coronal conditions can be derived only when the energy equation is actually integrated (see Section V.B). It is clear, however, that the flow speeds observed in interplanetary space (Section II.C) are close to those predicted by the solar wind solutions and much higher than those predicted by the solar breeze solutions. This gives a sound empirical justification for concentrating on the former class in the remainder of this discussion. Parker (1964) has shown that solar wind solutions exist whenever $T(r)$ declines less rapidly than $1/r$ for $r < r_c$. At large heliocentric distances, $u(r)$ is nearly constant, and $n(r)$ must then decline as $1/r^2$ to conserve mass.

B. Analytic Integrations of the Energy Equation

A self-consistent treatment of coronal structure requires simultaneous integration of the mass, momentum, and energy equations. Assuming equal electron and proton temperatures, the pertinent equations become

$$f = 4\pi nur^2 , \tag{16}$$

$$u\frac{du}{dr} = -\frac{1}{nm}\frac{d}{dr}(2nkT) - \frac{GM_\odot}{r^2} , \tag{17}$$

and

$$F = 4\pi nur^2\left(\frac{1}{2}mu^2 + 5kT - \frac{GM_\odot m}{r}\right) - 4\pi r^2\kappa\frac{dT}{dr} . \tag{18}$$

Analytic integrations can be performed only if further (and rather extreme) simplifications are made. Two such approximations, which will help to clarify properties of the numerical solutions to be described later and which illustrate the criterion for solar wind or solar breeze expansions, will be presented here.

1. Energy Transport Determined by Heat Conduction

If the particle flux f from the expanding corona is small enough, the energy flux equation (Equation 18) is dominated by the heat conduction term:

$$F \approx -4\pi r^2 \kappa \frac{dT}{dr} = -4\pi r^2 \kappa_s T^{5/2} \frac{dT}{dr} . \tag{19}$$

This differs from the situation in the transition zone (Section IV) only in that the energy flux is outward (or positive). In Equation 19, assuming that $T \to 0$ as $r \to \infty$ gives

$$T = T_\odot (r_\odot / r)^{2/7} .$$

This is the exact solution for a static corona, as given by Chapman (1957). For $T_\odot = 2 \times 10^6$ K, the energy flux carried by heat conduction would be 2×10^{27} erg-s^{-1} (compare with the values in Section III.C).

2. Energy Transport by Heat Conduction and Convection of Gravitational Energy

Consider next a particle flux f large enough that the convective transport of energy cannot be neglected. For a reasonable coronal temperature, the gravitational term $GM_\odot m/r$ is larger than either the internal energy term $5kT$ or the kinetic energy term $1/2\ mu^2$ in much of the corona (certainly for $r < r_c$). The energy equation (Equation 18) can then be approximated by

$$F \approx -4\pi nur^2 \frac{GM_\odot m}{r} - 4\pi r^2 \kappa \frac{dT}{dr} .$$

This equation can be integrated to give

$$T = T_\odot \left[\frac{r_\odot}{r} \frac{1 + r_2/r}{1 + r_2/r_\odot} \right]^{2/7} ,$$

where

$$r_2 = r_\odot \frac{1}{2F} 4\pi n u r^2 \frac{GM_\odot m}{r_\odot}$$

and the energy flux is related to $T_\odot$ by

$$F = \frac{2}{7} 4\pi r_\odot \kappa_\odot T_\odot - \frac{1}{2} 4\pi n u r^2 \frac{GM_\odot m}{r} . \tag{20}$$

The temperature is proportional to $r^{-4/7}$ for $r \ll r_2$ and to $r^{-2/7}$ for $r \gg r_2$. This solution differs from that in which convection of energy was neglected only if r_2 is comparable to or larger than $r_\odot$. This occurs when $F \geq 1/7(4\pi r_\odot K_\odot T_\odot)$, or when half or more of the conduction flux is cancelled by the gravitational term (which gives the energy required to transport the particle flux f to infinity in the gravitational field).

A nonzero flow speed at infinity implies an outward energy flux, or $F > 0$. Equation 19 then gives a necessary condition for the existence of a solar wind type of solution in the approximation under consideration. Derivation of a criterion in terms of $n_\odot$ and $T_\odot$ (eliminating the flow speed in Equation 19) requires further analysis. Parker (1964) has shown that, as might be expected, solar wind solutions occur for small $n_\odot$, whereas the solar breeze solutions occur for large $n_\odot$.

C. Numerical Integrations of the "One-Fluid" Equations

Simultaneous integrations of the mass, momentum, and energy equations without resort to the simplifications made above generally require numerical techniques. We will now discuss numerical solutions of Equations 16, 17, and 18; because of the assumption that the proton and electron temperatures are equal, these are generally referred to as "one-fluid" solar wind models. Noble and Scarf (1963) adopted typical fluid parameters at 1 AU and integrated inward toward the Sun, obtaining a reasonable coronal temperature but a rather low coronal density. Our approach will be to adopt reasonable coronal parameters, find solutions in the range $r_\odot \leq r < \infty$, and treat the values obtained at 1 AU as predictions to be compared with observation.

1. Boundary Conditions

The system under discussion consists of one algebraic equation and two first-order differential equations, with two arbitrary constants f and F already present. Two more constants will result from integration of the system; thus, four boundary conditions must be stated to determine a unique solution. Specification of the coronal density $n_\odot$ and temperature $T_\odot$ and the requirement that the solution pass through the critical point (i.e., be a solar wind solution) provide three conditions. The fourth has usually been stated as a restriction on the behavior of $T(r)$ as $r \to \infty$. However, different authors have given solutions with different behaviors; e.g., $T \sim r^{-2/7}$ in the analytic solution of Parker (1964) described in Section V.B and in the numerical solution of Scarf and Noble (1965), but $T \sim r^{-2/5}$ in the numerical solution of Whang and Chang (1965). Some discussion of this boundary condition and its physical implications is in order.

It is convenient to rewrite the energy equation by substituting Equations 1 and 2 into Equation 8 to obtain

$$2nu\left(\frac{3}{2}k\frac{dT}{dr} - \frac{kT}{n}\frac{dn}{dr}\right) = \frac{1}{r^2}\frac{d}{dr}r^2\kappa\frac{dT}{dr} \tag{21}$$

Dimensionless variables can be defined in terms of the parameters at some reference position r_0:

$$\eta = n/n_0,$$

$$\mu = u/u_0,$$

$$\tau = T/T_0,$$

$$\chi = r/r_0,$$

and

$$K = \kappa/\kappa_0.$$

In terms of these variables, Equation 21 becomes

$$\epsilon\left(\frac{3}{2}\frac{d\tau}{d\chi} - \frac{\tau}{\eta}\frac{d\eta}{d\chi}\right) = \frac{d}{d\chi}\left(\chi^2 K\frac{d\tau}{d\chi}\right), \tag{22}$$

where

$$\epsilon = fkT_0/4\pi r_0\kappa_0 T_0$$

is a dimensionless constant proportional to the ratio of internal energy flux and heat conduction flux at $r = r_0$. Consider solutions that approach the form $\tau \sim \chi^{-\beta}$ as $\chi \to \infty$. Three possibilities exist:

(1) The first is

$$\frac{d}{d\chi}\left(\chi^2 K \frac{d\tau}{d\chi}\right) \to 0 \quad \text{and} \quad \epsilon \to 0 .$$

This gives $\beta = 2/7$ and corresponds to the $T \sim r^{-2/7}$ solutions of Parker (1964) and Scarf and Noble (1965). The fact that $\epsilon \to 0$ implies that the conduction flux is much greater than the internal energy flux. If such a solution applies for $r > r_a$ (an arbitrary position), the conduction flux approaches a constant nonzero value F_c as $r \to \infty$, and the flow speed can increase only slightly for $r > r_a$. Examination of Equation 18 shows that the flow speed at infinity will take the value

$$u_\infty^2 = \frac{2(F - F_c)}{mf} .$$

(2) Equation 22 is satisfied by some β with ϵ a finite constant. Whang and Chang (1965) demonstrated that such a solution exists as $r \to \infty$ if $\beta = 2/5$ and $\epsilon = 1.26265$ at $r = r_c$. Both the internal energy flux and the heat conduction energy flux approach zero as $r \to \infty$, so that the flow speed at infinity is $u_\infty{}^2 = 2F/mf$.

(3) The third is

$$\frac{3}{2}\frac{d\tau}{d\chi} - \frac{\tau}{\eta}\frac{d\eta}{d\chi} \to 0 \quad \text{and} \quad \frac{1}{\epsilon} \to 0 .$$

This gives $\beta = 4/3$, corresponding to an adiabatic expansion ($n \sim 1/r^2$ and the constancy of P/ρ^γ in an adiabatic process gives $T \sim r^{-4/3}$). The fact that $1/\epsilon \to 0$ implies that the heat conduction flux is much smaller than the internal energy flux. Both approach zero as $r \to \infty$, and $u_\infty{}^2 = 2F/mf$.

We conclude that three different types of solutions to the one-fluid equations can exist, each with a different behavior of $T(r)$ and a different partition of the total energy flux at large heliocentric distances. The statement of a boundary condition in the form of a restriction on $T(r)$ as $r \to \infty$ must be made with care, since the partition of the total energy flux can thereby be implicitly determined.

A *density-dependent* limit on the heat conduction flux was given in Section III.A:

$$F_c \lesssim 4\pi r^2 (nm/2)(3kT/m)^{3/2} .$$

Solar wind solutions give a density proportional to r^{-2} at large r. If the temperature is assumed to approach zero as $r \to \infty$, then F_c must also approach zero. This physical constraint on the heat conduction is not contained in the fluid equations (Equations 16, 17, and 18), which contain a *density-independent* thermal conductivity. In fact, solutions that obey the first of the possible boundary conditions described above violate this constraint on the heat conduction flux. Only the second and third boundary conditions remain physically plausible.

2. *The One-Fluid Solar Wind Model of Whang and Chang*

Whang and Chang (1965) numerically integrated the fluid equations, applying the second of the boundary conditions given above. A unique solution in dimensionless variables (scaled in terms of physical conditions at the critical point) results. Conversion to physical variables gives solutions for a restricted set of coronal parameters (i.e., at a single density $n_\odot$ for a given coronal temperature $T_\odot$). The choice of $r_c = 7.5 r_\odot$ was judged to produce the best agreement between the model and observed coronal and interplanetary properties; the predicted parameters at $r = r_\odot$ and $r = r_e$ (the orbit of the Earth) are given in Table 4. Moving the critical radius closer to the base of the corona gives solutions corresponding to a lower $n_\odot$ and a higher $T_\odot$; the density, flow speed, and temperature at 1 AU all increase.

Table 4.—One-fluid solar wind model with $r_c = 7.5r$.

Region	Density (cm^{-3})	Flow Speed ($km\text{-}s^{-1}$)	Temperature (K)
$r = r_\odot$	7×10^7	1.2	1.6×10^6
$r = r_e$	0.8	260	1.6×10^5

D. The Two-Fluid Model of the Solar Wind

The transfer of energy from electrons to protons by Coulomb collisions is relatively inefficient because of the large difference in mass between the particles. Sturrock and Hartle (1966) argued that the time scale for energy equipartition in the low-density interplanetary plasma is so great that electrons and protons need not have the same temperature. Subsequent observations (Table 1) have borne out this conjecture. Incorporation of this nonequilibrium effect into the fluid equations for the coronal expansion leaves the mass equation (Equation 16) unchanged and requires only substitution of the pressure $P = nk(T_e + T_i)$ for the one-fluid expression $P = 2nkT$ in the momentum equation (Equation 17). However, separate energy equations must be written for electrons and protons. In their formulation of this "two-fluid" model of the solar wind, Sturrock and Hartle (1966; Hartle and Sturrock, 1968) used energy equations in the form of Equation 21:

$$nu\left(\frac{3}{2}k\frac{dT_e}{dr} - \frac{kT_e}{n}\frac{dn}{dr}\right) = \frac{1}{r^2}\frac{d}{dr}\left(r^2\kappa_e\frac{dT_e}{dr}\right) - \frac{3}{2}\nu nk(T_e - T_p) \tag{23}$$

and

$$nu\left(\frac{3}{2}k\frac{dT_p}{dr} - \frac{kT_p}{n}\frac{dn}{dr}\right) = \frac{1}{r^2}\frac{d}{dr}\left(r^2\kappa_e\frac{dT_p}{dr}\right) + \frac{3}{2}\nu nk(T_e - T_p)\,, \tag{24}$$

with the conventional thermal conductivities κ_e and κ_p for electrons and protons. The two equations are coupled by an energy exchange term. This exchange was assumed to occur by Coulomb collisions, for which

$$\nu = \nu_0 n T_e^{-3/2}\,,$$

where ν_0 is a constant (see Hartle and Sturrock, 1968, p. 1158).

Solar wind solutions to Equations 16, 17, 23, and 24 were obtained by numerical integration for varied coronal conditions and with the boundary condition "that both $T_e \to 0$ and $T_p \to 0$ as $r \to \infty$" (Hartle and Sturrock, 1968, p. 1159). The coronal parameters $n_\odot = 3 \times 10^7$ cm^{-3} and $T_{e\odot} = T_{p\odot} = 2 \times 10^6$ K resulted in good agreement between predicted and observed densities in the range $2r_\odot \leqslant r \leqslant 20r_\odot$. This basic two-fluid model is summarized in Table 5.

Table 5.–Basic two-fluid solar wind model.

Region	Density (cm^{-3})	Flow Speed ($km\text{-}s^{-1}$)	Electron Temperature (K)	Proton Temperature (K)
$r = r_\odot$	3×10^7	5.8	2×10^6	2×10^6
$r = r_e$	15	250	3.4×10^5	4.4×10^3

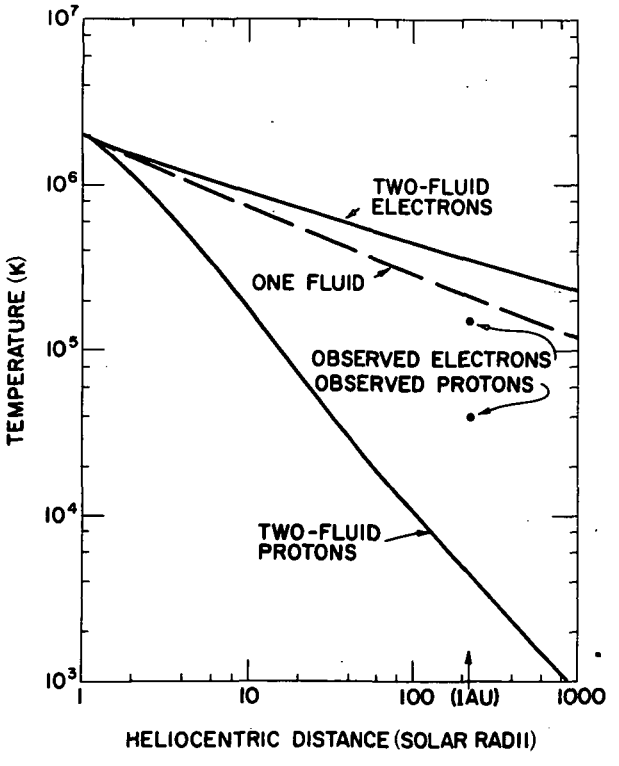


Figure 8.–Particle temperatures as a function of heliocentric distance, given by the two-fluid model (solid lines) and the one-fluid model (dashed-line). A coronal temperature $T_\odot = 2 \times 10^6$ K has been assumed in each model. The electron and proton temperatures observed at 1 AU (from Table 1) are also shown (Hundhausen, 1970).

Figure 8 shows $T_e(r)$ and $T_p(r)$ from the basic two-fluid model and $T(r)$ from the one-fluid model with the critical radius changed also to give $T_\odot = 2 \times 10^6$ K. The coupling term in the energy equations (Equations 23 and 24) is proportional to n^2 and is, therefore, a rapidly decreasing function of r. Both source terms in the proton energy equation ($\kappa_p \ll \kappa_e$) become small, allowing the protons to cool nearly as rapidly as in an adiabatic expansion. In contrast, the heat conduction source term in the electron energy equation remains large, keeping the electron temperature high. The one-fluid model, which is the limiting case of a two-fluid model with ν large, predicts a $T(r)$ between $T_p(r)$ and $T_e(r)$, as would be expected. Note, however, that the electron temperature predicted by the two-fluid model varies as $r^{-2/7}$ for $r \gtrsim 10 r_\odot$ (Holzer and Axford, 1970; Hundhausen, 1970; Hartle and Barnes, 1970), whereas the temperature in the one-fluid model varies as $r^{-2/5}$ as $r \to \infty$. The two solutions obey different boundary conditions as $r \to \infty$: in the two-fluid model, the heat conduction flux F_c will approach some constant value, whereas in the one-fluid model, F_c will approach zero as $r \to \infty$. The two-fluid model does not exhibit the physically required transition to a distant interplanetary region with a density-limited heat conduction.

E. Comparison of Models and Observations

Comparison of the coronal and interplanetary properties predicted by one-fluid (Table 4) and two-fluid (Table 5) models of the coronal expansion with observed coronal (Section II.A) and interplanetary (Table 1) properties reveals basically good agreement. If the coronal density and temperature are given reasonable values, the resulting solar wind models reproduce with fair accuracy the characteristics of the interplanetary plasma near 1 AU. A detailed comparison at 1 AU leads to the following specific conclusions:

(1) Predicted densities are slightly high (or a rather low coronal density must be chosen).

(2) Predicted flow speeds are slightly lower than *usually* observed. However, flow speeds as low as 250 to 260 km-s^{-1} are observed on rare occasions.

(3) The temperature predicted by the one-fluid model agrees with the observed electron temperature and is about four times higher than the observed proton temperature.

(4) The electron temperature predicted by the two-fluid model is two or three times higher than observed, whereas the predicted proton temperature is an order of magnitude higher than observed.

Much of the recent work on models of the coronal expansion has concentrated on the refinement of the basic models described above in order to improve the detailed agreement of predictions and observations. Two such refinements will be described in the remainder of this section because of their close relationship to the discussions of mechanical energy dissipation and the energy balance of the outer solar atmosphere in Section III. However, it should be recalled that the models basic to this work contain many idealizations (e.g., the assumptions of spherical symmetry and steady flow). The effects of these simplifying assumptions on the models may be as large as those produced by the refinements under discussion. Any conclusions that may be drawn regarding physical processes in the corona and interplanetary region are inherently subject to this basic uncertainty and lack of uniqueness.

F. Extension of Mechanical Energy Dissipation Into the Interplanetary Region

The tendency for the two-fluid model to predict flow speeds and proton temperatures lower than those actually observed near 1 AU has been interpreted as evidence for an energy source in addition to heat conduction; i.e., ". . . heating of the corona, and its extension to form the solar wind, continues far out into interplanetary space" (Hartle and Sturrock, 1968, p. 1168). Integration of Equations 16, 17, 23, and 24, then, could not begin near $r = r_\odot$ without inclusion of the source term $S_m(r)$ in the energy equations. Sturrock and Hartle (1966) and Hartle and Sturrock (1968) have hypothesized that the heating term affects only the proton component of the coronal and interplanetary plasma.

An extended mechanical dissipation term has been quantitatively incorporated into the two-fluid model by Hartle and Barnes (1970). The source term

$$S_m(r) = D_0(n/n_0) \exp\left[-(r/r_\odot - a)^2/b^2\right]$$

is added to the proton energy equation (Equation 24). Figure 9 shows the modification of the basic two-fluid solution for $T_e(r)$ and $T_p(r)$, in the range $r > 2r_\odot$, produced by a source with $a = 2$ and $b = 22$. As might be expected, the

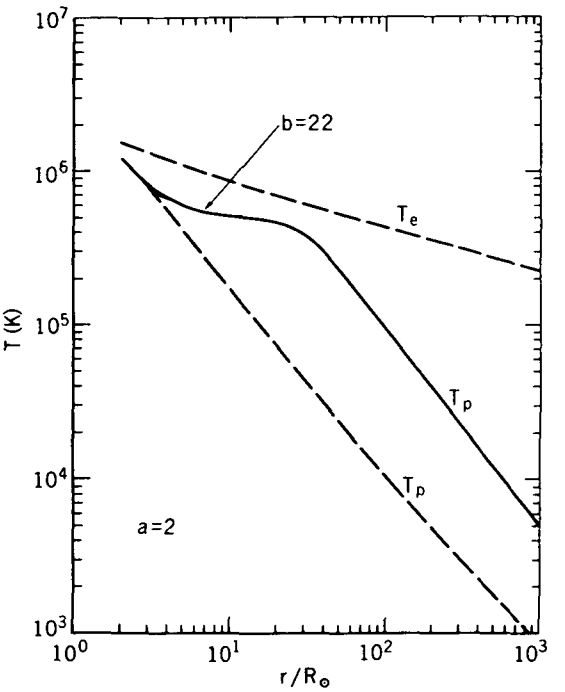


Figure 9.–Electron and proton temperatures as a function of heliocentric distance as given by the two-fluid model (dashed lines) and the modified proton temperature (solid line) produced by the extended mechanical dissipation term of Hartle and Barnes (1970).

proton temperature is elevated in the region of additional heating; when this added energy is ultimately converted into kinetic energy, the flow speed rises. Table 6 summarizes the resulting model, with mechanical dissipation extending as far as $r \approx 20r_\odot$. The proton temperature and flow speed at 1 AU have both been raised above the values given by the basic two-fluid model and are in good agreement with observations (Table 1). The density remains significantly higher than observed, even though a rather low coronal density $n_\odot = 3 \times 10^7$ cm^{-3} has been assumed.

Despite the improved agreement between the observations and the two-fluid model with extended heating, the latter has some striking deficiencies which cast doubt on the conclusion that mechanical heating extends into interplanetary space. Addition of the two-fluid electron and proton energy equations with the added

source term $S_m(r)$ leads to the energy conservation integral

$$F = 4\pi\left\{nur^2\left[\frac{1}{2}mu^2 + \frac{5}{2}k(T_e + T_i) - \frac{GM_\odot m}{r}\right] - r^2\kappa_e\frac{dT_e}{dr}\right.$$

$$\left. - r^2\kappa_p\frac{dT_p}{dr} + \frac{F_{m0}}{4\pi} - \int_{r_0}^{r} r^2 S_m(r)\,dr\right\}. \tag{25}$$

A valid discussion of energy addition to the expanding corona must be based on this equation. Table 7 lists the observed convective and electron heat conduction flux densities (proton heat conduction is far smaller than any of these and can be

Table 6.—Two-fluid model with extended mechanical dissipation.

Region	Density (cm^{-3})	Flow Speed ($km\text{-}s^{-1}$)	Electron Temperature (K)	Proton Temperature (K)
$r = r_\odot$	3×10^7	5.8	2×10^6	2×10^6
$r = r_e$	13	320	3.4×10^5	3.7×10^4

Table 7.—Observed and predicted energy flux densities at 1 AU.

Flux Density	Observed ($erg\text{-}cm^{-2}\text{-}s^{-1}$)	Two-Fluid Model ($erg\text{-}cm^{-2}\text{-}s^{-1}$)
Convection of kinetic energy	0.22	0.20
Convection of enthalpy (electrons)	0.011	0.045
Convection of enthalpy (protons)	0.005	0.0006
Convection of gravitational energy	-0.004	-0.005
Heat conduction by electrons	≈0.01	0.29
Total	0.24	0.53

neglected) and those predicted by the basic two-fluid model. Comparison of these energy flux densities leads to a conclusion rather different from that drawn by the comparison of predicted and observed flow speeds and proton temperatures. The two-fluid model does not predict a total energy flux smaller than is observed; on the contrary, it predicts a flux almost twice as large as the observed value. However, over half of the total flux is in heat conduction. Adding more energy to the protons does increase the flow speed and proton temperature, but it can only make the discrepancy in total energy flux still larger.

We have already noted that the two-fluid solutions of Hartle and Sturrock (1968) give a finite heat conduction flux as $r \to \infty$. The extended heating models of Hartle and Barnes (1970) share this property. Both models thus require an unphysical conductivity at large r. Table 7 shows that both models predict a heat conduction flux at 1 AU which is more than an order of magnitude larger than the observed value. It is clear that a very important energy transport term is not well handled by present two-fluid models. Any conclusions regarding energy in the flow must be accepted with caution. It is not clear whether this problem is fundamental to two-fluid models or whether the problem is caused by the boundary condition satisfied by the present solutions.

G. Noncollisional Processes in the Interplanetary Plasma

The important role played by heat conduction in solar wind models is clearly demonstrated by the above discussion. The widely used conventional plasma thermal conductivity $\kappa = \kappa_s T^{5/2}$ is actually valid only when the dimensionless parameter B_T (see Section III.A) has a magnitude much less than 1. Figure 10 shows $|B_T|$ as a function of heliocentric distance for both the two-fluid and one-fluid models. The condition $|B_T| \ll 1$ is satisfied only in the dense plasma near the Sun. In the former model, $|B_T|$ becomes as large as ½ at $r \approx 15 r_\odot$, and in the latter model, at $r \approx 50 r_\odot$ (even though the upper limit on heat conduction flux is not exceeded as $r \to \infty$). Thus, the actual heat conduction flux will be smaller throughout most of interplanetary space than would be obtained using $\kappa = \kappa_s T^{5/2}$.

Such a "cutoff" of the heat conduction would have two important effects on the coronal expansion. The first of these, demonstrated by Parker (1964) and easily seen from the energy conservation equation, is an increase in the predicted flow speed for heliocentric distances greater than that at which the cutoff occurs

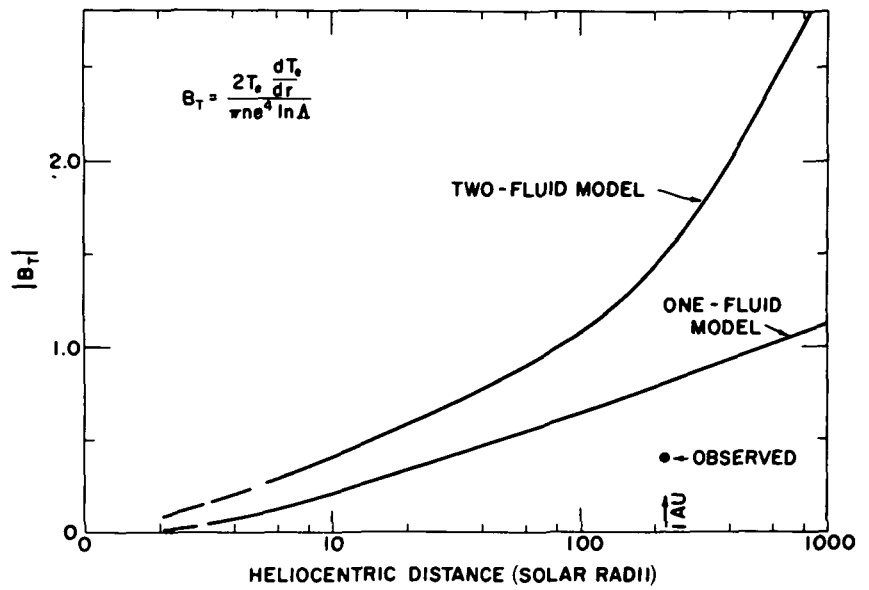


Figure 10.—Dimensionless parameter B_T as a function of heliocentric distance, as given by the two-fluid and one-fluid models. The conventional plasma thermal conductivity $\kappa = \kappa_s T^{5/2}$ applied only for $|B_T| \ll 1$.

(with $u_\infty{}^2 = 2F/mf$ as an upper limit). The second effect, discussed by Forslund (1970), stems from the development of several plasma instabilities when B_T is larger than a few tenths. These instabilities not only lead to noncollisional particle interactions that modify the conductivity, but they can transfer thermal energy from hot electrons to cooler protons. Both of the "diffusion" coefficients in the two-fluid energy equations (Equations 23 and 24), namely, the conductivity κ and the energy exchange rate ν, would be modified. One expects these modifications to lead to higher flow speeds and higher electron temperatures [i.e., the same effects that were produced by the extended proton heating source in the model of Hartle and Barnes (1970)]. One also expects a drastic limitation of the electron heat conduction flux and a small (less than a factor of 2) but significant decrease in the electron temperature. Both of these effects (which are not produced by extended coronal heating) should improve still further the agreement between the two-fluid model and observations. The incorporation of these noncollisional processes into quantitative models is necessary to determine how good an agreement can be obtained.

H. The Magnetic Inhibition of Heat Conduction at Large Heliocentric Distances

The existence of coronal and interplanetary magnetic fields has been neglected in the expansion models so far discussed. Weber and Davis (1967) have included the magnetic force on the plasma in a model with $T(r)$ assumed to be known. This force produces only a small change in the radial flow speed plus a small nonradial velocity component. For $r \gtrsim r_c$, the kinetic energy density of the plasma is much larger than the energy density in the magnetic field, and the latter can be treated as entirely passive, "frozen" into the plasma by the high electrical conductivity and carried with the plasma flow. If the Sun did not rotate, the radial expansion of the corona would draw the solar magnetic field lines out in the radial direction. The actual rotation of the Sun distorts this simple pattern into a spiral. The field line from a given point on a "source surface" rotating with the Sun will be drawn out along the locus of plasma elements flowing from the same point. For a steady, azimuthally symmetric flow, the field configuration is as shown in Figure 11, with a radial component at any distance proportional to the radial expansion speed u and an azimuthal component proportional to the speed of a frame rotating with the Sun, ωr. At large heliocentric distances, $u(r)$ is nearly constant and the field lines are Archimedes spirals.

Although the magnetic force does not produce any significant modification of the plasma flow, the field configuration does affect heat conduction. For reasonable interplanetary parameters, the thermal conductivity transverse to the magnetic field is negligible (Section III.A). The heat conduction flux through a Sun-centered sphere with radius r is

$$F_c = 4\pi r^2 \kappa_{\parallel} \frac{dT}{ds} \cos\phi \,,$$

where $\kappa_{\parallel}$ and s are the conductivity and distance along the field line and ϕ is the angle between the field and the radial direction (see Figure 11). As $dr = ds \cos\phi$ and $\kappa_{\parallel} = \kappa$,

$$F_c = 4\pi r^2 \cos^2\phi \, \kappa \frac{dT}{dr} \,. \tag{26}$$

The radial conduction flux is "inhibited" by the factor $\cos^2\phi$. Inspection of Figure 11 gives the inhibition factor as an explicit function of r:

$$\cos^2\phi = 1/[1 + (\omega r/u)^2] .$$

1. The Effect of Magnetic Inhibition on Boundary Conditions

As $r \to \infty$, the inhibition factor $\cos^2\phi \to (u/\omega r)^2$. The conduction flux is then

$$F_c = 4\pi(u/\omega)^2\kappa \frac{dT}{dr}$$

$$= 4\pi(u/\omega)^2\kappa_s T^{5/2} \frac{dT}{dr} .$$

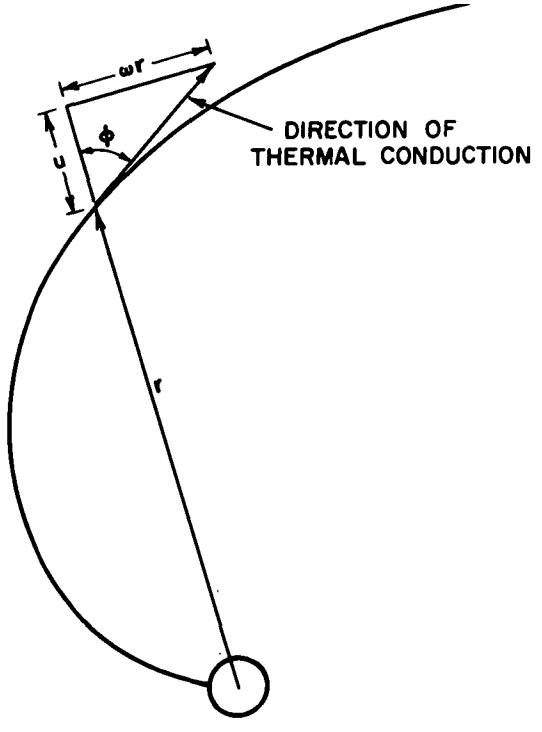


Figure 11.—Spiral configuration of the interplanetary magnetic field. Heat conduction transverse to the magnetic field will be negligible.

For any well-behaved temperature such that $T \to 0$ as $r \to \infty$, F_c must also approach zero. The magnetic inhibition factor thus removes the possibility of solutions (shown to be nonphysical in part C of this section) with $T \to 0$ but a finite heat conduction at infinity. Use of the flux equation (Equation 26) in the one-fluid energy equation gives, in dimensionless form,

$$\epsilon\left(\frac{3}{2}\frac{d\tau}{d\chi} - \frac{\tau}{\eta}\frac{d\eta}{d\chi}\right) = \frac{d}{d\chi}\left(K\frac{d\tau}{d\chi}\right),$$

where

$$\epsilon = \frac{fkT_0}{4\pi\kappa_0 T_0/r_0} = \frac{fk}{4\pi\kappa_0} r_0 .$$

As $r \to \infty$, $\epsilon \to \infty$, and the only plausible solution is that in which

$$\frac{3}{2}\frac{d\tau}{d\chi} - \frac{\tau}{\eta}\frac{dv}{d\chi} \to 0 ;$$

that is, the ambiguity demonstrated in part C is no longer present. The admissible solution is that for an adiabatic expansion, $T \sim r^{-4/3}$ and $u_\infty = 2F/mf$ at large r. Note also that for the adiabatic expansion $|B_T| \sim T^2$ and will approach zero as $r \to \infty$. The inclusion of the magnetic inhibition factor in the one-fluid equations results in a "well-posed" problem, with the conventional conductivity remaining valid at large heliocentric distances.

2. *Quantitative Models With Magnetic Inhibition*

Several authors (Brandt et al., 1969; Urch, 1969; Grzedzielski, 1969; Gentry and Hundhausen, 1969) have obtained quantitative, one-fluid, coronal expansion models including the effects of the magnetic field on heat conduction. Figure 12 shows $T(r)$ from one of the solutions of Gentry and Hundhausen (1969) for which $T_\odot = 1.6 \times 10^6$ K, as in the ordinary Whang and Chang (1965) one-fluid model. The effect of the inhibition factor at large distances is clearly demonstrated. The solution follows that for the ordinary model at small r but merges into the expected adiabatic solution at $r \approx 5r_e$. The transition to a distant interplanetary solution, as discussed in Section III.B, occurs near this distance (compare Figures 1 and 12).

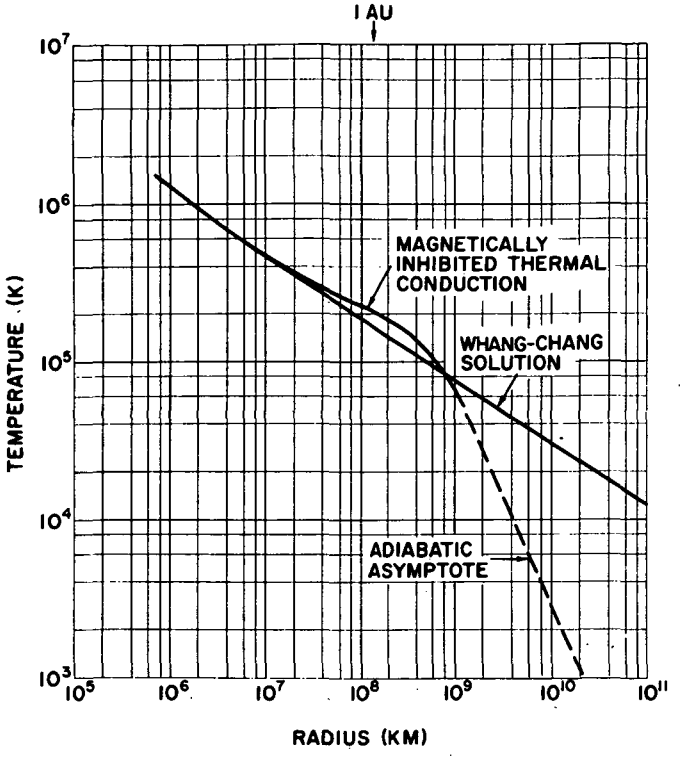


Figure 12.—Temperature as a function of heliocentric distance given by both the one-fluid model of Whang and Chang (1965) and the one-fluid model with magnetic reduction of transverse heat conduction of Gentry and Hundhausen (1969). The latter approaches an adiabatic expansion at large heliocentric distances.

Table 8.—One-fluid solar wind model with magnetically inhibited heat conduction.

Region	Density (cm^{-3})	Flow Speed ($km\text{-}s^{-1}$)	Temperature (K)
$r = r_\odot$	3×10^7	4.3	1.8×10^6
$r = r_e$	8.4	324	2.8×10^5

Table 8 summarizes one of the solutions of Gentry and Hundhausen. Coronal conditions have been assumed which are very close to those in the basic two-fluid model. The predicted density and flow speed at 1 AU are very close to those actually observed (Table 1). The predicted heat conduction flux, 0.05 erg-cm^{-2}-s^{-1}, is larger than the observed value, indicating that some additional inhibiting effect (such as is discussed in part G) should be incorporated into the model. However, the heat conduction flux is predicted to be only 20 percent of the total energy flux; the large discrepancy displayed by the two-fluid model is avoided. The one-fluid model with magnetic inhibition of the heat conduction is thus found to give a more accurate description of the gross flow properties (i.e., the density, flow speed, proton flux, total energy flux, and partition of energy flux) than does the two-fluid model, with or without the extended heating term. We are again led to question the conclusion that coronal heating must extend into the interplanetary region.

VI. CHEMICAL COMPOSITION IN THE EXPANDING SOLAR ATMOSPHERE

The solar atmosphere has been assumed to consist of completely ionized hydrogen in the entire preceding discussion. The relative rarity of all other elements makes this a good approximation in dealing with gross dynamics of the atmosphere. However, ions of the other elements are observed in the interplanetary plasma, so theoretical models that relate interplanetary, coronal, and photospheric ionic abundances are of interest. Two such models will be discussed here; for a review of the pertinent interplanetary observations see Hundhausen (1970).

A. The Coronal and Interplanetary Abundance of Helium

The ions $^1H^+$ and $^4He^{++}$ are routinely observed by many interplanetary probes. From such measurements made near 1 AU, the average abundance of He relative to H (by number) is found to be 4 to 5 percent. Large fluctuations are also observed, with relative abundances as high as 20 percent occurring under disturbed conditions. The solar abundance of helium can be measured only by indirect means

and appears to be 6 to 10 percent by number. The difference between these values may reflect observational uncertainties but could be explained by a separation of ions of different charge-to-mass ratios in the coronal expansion.

Consider a spherically symmetric corona composed only of $^1H^+$, $^4He^{++}$, and electrons. The momentum equations for the three constituents are

$$n_p m_p u_p \frac{du_p}{dr} = -\frac{d}{dr}(n_p k T_p) - \frac{n_p m_p G M_\odot}{r^2} + n_p q E - C , \tag{27}$$

$$4n_\alpha m_p u_\alpha \frac{du_\alpha}{dr} = -\frac{d}{dr}(n_\alpha k T_\alpha) - \frac{4n_\alpha m_p G M_\odot}{r^2} + 2n_\alpha q E + C , \tag{28}$$

and

$$n_e m_e u_e \frac{du_e}{dr} = -\frac{d}{dr}(n_e k T_e) - \frac{n_e m_e G M_\odot}{r^2} - n_e q E , \tag{29}$$

where q = magnitude of the electronic charge and the subscripts p, α, and e refer to $^1H^+$, $^4He^{++}$, and electrons, respectively. The term C in the $^1H^+$, $^4He^{++}$ equations (Equations 27 and 28) represents the momentum transfer by such processes as collisions between the two ion species. It is sufficient for our immediate purposes to note that C must go to zero as $u_p - u_\alpha$ approaches zero. The electric field E is produced by charge separation and is thus related to the charge density $n_p + 2n_\alpha - n_e$ by Poisson's equation.

Steady-state solutions of Equations 27 through 29, but neglecting the frictional term C, have been considered by Yeh (1970). This analysis shows a supersonic flow for both ion species at large heliocentric distances and a helium-hydrogen density ratio that decreases with increasing r. For our purposes, it is more instructive to assume that $n_\alpha \ll n_p$ so that the electric field is determined by the proton and electron densities only. This permits independent treatment of Equations 27 and 29 and neglect of the frictional term in the former. The expectation of near charge neutrality again gives $u_e \approx u_p$ and $n_e \approx n_p$. Under these approximations, addition of Equations 27 and 29 gives

$$n_p(m_p + m_e)u_p \frac{du_p}{dr} = -\frac{d}{dr}[n_p k(T_p + T_e)] - \frac{n_p(m_p + m_e)GM_\odot}{r^2} .$$

We will henceforth neglect m_e relative to m_p and assume for purposes of algebraic simplicity that $T_e = T_\alpha = T_p$ (this latter assumption is probably valid throughout most of the corona but not in interplanetary space). Then,

$$n_p m_p u_p \frac{du_p}{dr} = -2\frac{d}{dr}(n_p k T_p) - \frac{n_p m_p G M_\odot}{r^2} . \tag{30}$$

This is the standard "one-fluid" momentum equation. The now-vanished electric field has merely served the purpose of coupling the protons and electrons so that the total pressure $n_p k(T_e + T_p)$ affects the flow. The field E can be explicitly evaluated by subtracting Equation 27 from Equation 29, giving

$$E = \frac{m_p}{2q}\left(\frac{GM_\odot}{r^2} + u_p \frac{du_p}{dr}\right) . \tag{31}$$

The interested reader may verify by substitution in Equations 27 and 29 that this field gives equal $^1H^+$ and electron accelerations.

With a solution for $u_p(r)$ from Equation 30, the electric field of Equation 31 can be substituted in the $^4He^{++}$ momentum equation (Equation 28) to determine the motion of the latter ions. The nature of the solutions can be understood by considering several special cases.

1. Static Corona

In an isothermal corona, Equation 30 gives the simple solution for proton density:

$$n_p(r) = n_{p\odot} \exp\left(-\frac{m_p G M_\odot}{2kTr_\odot}\frac{r - r_\odot}{r}\right) .$$

The electric field is

$$E = \frac{m_p}{2q}\frac{GM_\odot}{r^2} ,$$

and the solution to Equation 28 for the helium density is

$$n_\alpha(r) = n_{\alpha\odot} \exp\left(-\frac{3m_p G M_\odot}{kTr_\odot}\frac{r - r_\odot}{r}\right) .$$

The helium density thus decreases more rapidly with increasing r (the helium scale height is one-sixth the hydrogen scale height at any r) than does the hydrogen density, giving a decreasing helium abundance with increasing solar altitude.

More general solutions of the static coronal case are given in Parker (1963a). However, this simplified example illustrates the basic interplay of the charge separation field and gravity in producing a vertical stratification of ions with different charge-to-mass ratios. In this particular case, the electric field cancels half of the gravitational force on an $^1H^+$ ion but only one-fourth of the gravitational force on a $^4He^{++}$ ion. The difference in scale heights is larger than would occur in a neutral atmosphere.

2. *Expanding Corona*

Our discussion will be largely based on the work of Geiss et al. (1970). Steady-state solutions of the one-fluid momentum equation (Equation 27) are obtained assuming a polytropic law relating the temperature T to n_p. It is also assumed that no ionizations or recombinations occur. The resulting $u_p(r)$ is used to evaluate the electric field

$$E = \frac{m_p}{2q}\left(\frac{GM_\odot}{r^2} + u_p \frac{du_p}{dr}\right)$$

and thereby permit solution of the helium momentum equation (Equation 28). The frictional term C is taken to be due to Coulomb collisions (including the cumulative effects of multiple small-angle scatterings). For a sufficiently high proton flux, the frictional term tends to bring the $^4He^{++}$ expansion speed to near equality with the $^1H^+$ expansion speed at large heliocentric distances. Observations of interplanetary $^4He^{++}$ and $^1H^+$ indicate that the speeds of the two ions are, in fact, nearly equal in the solar wind. Processes other than Coulomb collisions (e.g., streaming instabilities or magnetic forces) may be operative in the solar wind to maintain this equality. However, the presence of a term in the Geiss et al. models which produces this result facilitates comparison of results with observations at 1 AU and motivates its use in this discussion.

Figure 13 illustrates one set of solutions obtained by Geiss et al. for the $^1H^+$ and $^4He^{++}$ expansion speeds in the corona. The $^4He^{++}$ ions, feeling a larger effective gravity due to the smaller influence of the electric field, expand at a considerably

slower speed than the $^1H^+$ ions near the Sun, only reaching a comparable speed at large heliocentric distances. The particular solution for $u_\alpha(r)$ shown here assumes a hydrogen flux at 1 AU of 6×10^8 $cm^{-2}\text{-}s^{-1}$. For lower fluxes the difference in helium and hydrogen speeds is still more pronounced. The expansion speed of $^{56}Fe^{+14}$ is also shown (assuming a hydrogen flux of 3×10^8 $cm^{-2}\text{-}s^{-1}$); such heavier ions also experience smaller accelerations than $^1H^+$ ions but are swept along by collisions with hydrogen more efficiently than is $^4He^{++}$.

In the spherically symmetric expansion assumed here, the helium and hydrogen fluxes obey the conservation laws

$$n_\alpha u_\alpha(r) r^2 = f_\alpha$$

and

$$n_p u_p(r) r^2 = f_p ,$$

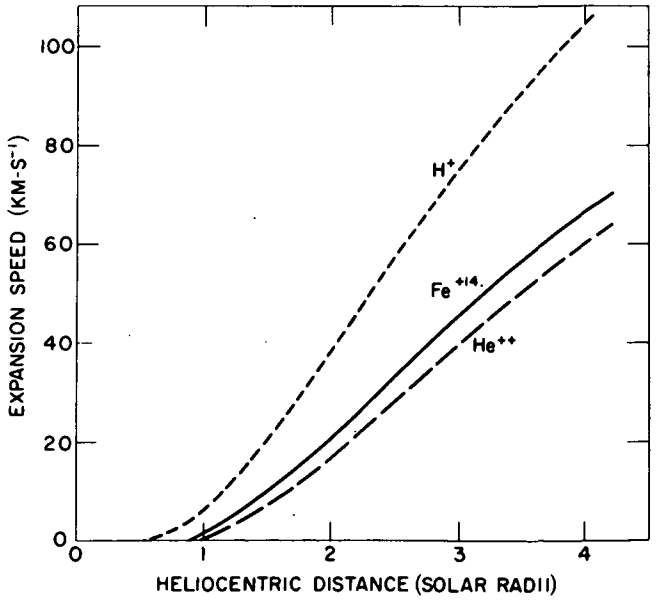


Figure 13.–Coronal expansion speeds of the ions $^1H^+$, $^4He^{++}$, and Fe^{+14} given by the model of Geiss et al. (1970).

where f_α and f_p are constants. Dividing one equation into the other and using the equality of u_p and u_α at 1 AU, which is both observed and predicted in these solutions, gives at any heliocentric distance,

$$\frac{n_\alpha}{n_p} = \frac{u_p}{u_\alpha} \frac{n_\alpha(1\ \mathrm{AU})}{n_p(1\ \mathrm{AU})}.$$

Figure 14 shows the helium-hydrogen density in the corona implied by the flow speeds shown in Figure 13, using the observed $n_\alpha/n_p = 0.05$ at 1 AU. An appreciable concentration of helium is predicted in the corona, with the density ratio reaching values above 0.2 for $r < 1.2r_\odot$. The validity of these solutions near $r = r_\odot$ is questionable. The large temperature gradients in the chromosphere will lead to diffusion of the ions, which is not included in the Geiss et al. models. Further, the high concentration would perturb the electric field based on the assumption that

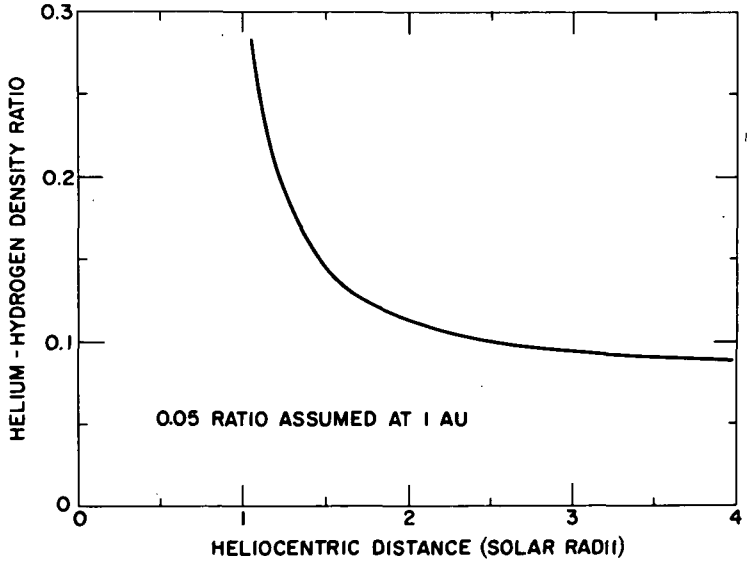


Figure 14.—Relative helium-hydrogen abundance in the corona implied by Figure 13. An abundance ratio of 0.05 was assumed at 1 AU, in accord with direct observations (Hundhausen, 1970).

$n_\alpha \ll n_p$. Despite these limitations, the effect of gravitational settling is explicitly demonstrated by these solutions, and a concentration of helium in the corona is implied. A similar concentration of other heavy elements such as Fe would be expected and consistent with spectroscopic evidence (Brandt, 1966).

The relationship between the *photospheric* and coronal helium abundances can be established only by models that are valid in the chromosphere and lower corona, where diffusion terms are important (or even dominant) in the transport equations. For highly idealized treatments of these equations, see Jokipii (1966), Delache (1965, 1967), and Nakada (1969). Each of these authors predicts an enhancement in the abundance of elements heavier than hydrogen as photospheric material rises through the chromosphere into the corona. Some of the difficulties precluding a realistic treatment of this problem can be grasped by reexamination of Figure 6. The effects of convection, local inhomogeneities, and magnetic fields make the quantitative results of the present models uncertain (which explains our decision above to relate the coronal helium abundance to that observed in interplanetary space rather than to a photospheric value). However, an enhanced helium abundance near $r = r_\odot$, as in Figure 14, is qualitatively consistent with the predicted effects of diffusion in the lower corona and chromosphere. The observation of extremely high solar wind helium abundances in association with such transient phenomena as interplanetary shock waves suggests a transient ejection of this helium-rich material into interplanetary space.

B. The Ionization State of the Expanding Solar Atmosphere

Several different ions of three elements, O, Si, and Fe (Bame et al., 1968 and 1970), have been observed in the interplanetary plasma. The large temperature variation encountered in the outer solar atmosphere implies different ionization states of a given element at different altitudes. We will concentrate our attention on the ionization state of the expanding corona.

The dominant ionization and recombination processes affecting the coronal ion X^i (i.e., the ith ionization state of element X) are–

(1) Collisional ionization by electron impact:

$$X^i + e \rightarrow X^{i+1} + e + e\,.$$

(2) Radiative (including dielectronic) recombination:

$$X^i + e \rightarrow X^{i-1} + h\nu \, .$$

The inverse processes, namely three-body recombination and photo ionization, would balance (1) and (2) in a gas in thermodynamic equilibrium. They fail to do so in the corona because of the low density (three-body recombination is proportional to $n_i n_e^2$) and low flux of radiation capable of ionizing the states present at the high temperatures existing in the corona. The conservation law for the ion X^i in an expanding atmosphere is obtained by equating the divergence of the flux $n_i u_i$ to the rate at which the ion is produced by atomic processes. Assuming spherical symmetry,

$$\frac{1}{r^2}\frac{d}{dr}(r^2 n_i u_i) = C_{i-1} n_{i-1} n_e - (C_i + R_i) n_i n_e + R_{i+1} n_{i+1} n_e \, , \tag{32}$$

where the subscript i pertains to the ith ion and the subscript e pertains to electrons. Here, C_i is the collisional ionization rate $\langle \sigma v \rangle$ for X^i (in the units $cm^3\text{-}s^{-1}$), and R_i is the radiative recombination rate for X^i. Equation 32 should be written for each ionization state of X (or at least each state important in the physical system under consideration) and the resulting system of coupled equations, plus the momentum and energy equations, solved. For our purposes, consider only two ions, X^i and X^{i+1}. Equation 32 becomes

$$\frac{1}{r^2}\frac{d}{dr}(r^2 n_i u_i) = -C_i n_i n_e + R_{i+1} n_{i+1} n_e \, ,$$

whereas the equation for X^{i+1} has a right-hand side which is the negative of that written above. In dimensionless form, this equation becomes

$$\frac{u_0}{r_0}\frac{1}{C_{i0} n_{e0}}\frac{1}{\chi^2}\frac{d}{d\chi}(\chi^2 \eta_i \mu_i) = -\gamma_i \eta_i \eta_e + \frac{R_{i+1,0} n_{i+1,0}}{C_{i0} n_{i0}} \rho_{i+1} \eta_{i+1} \eta_e \, ,$$

where γ_i and ρ_i are the ionization and recombination coefficients normalized in terms of C_{i0} and R_{i0}. Since r_0/u_0 is essentially the time scale τ_{exp} for expansion through a characteristic length r_0 and $1/C_{i0} n_e$ is the time scale τ_a for the atomic

ionization process, the dimensionless constant multiplying the left side of the equation is of the form $\tau_a/\tau_{\exp}$. If $\tau_{\exp} \gg \tau_a$, the right-hand side of the ionization equation can be set to zero, and one has the standard static coronal ionization equation

$$0 = -C_i n_i n_e + R_{i+1} n_{i+1} n_e ,$$

which gives $n_i/n_{i+1} = R_{i+1}/C_i$, a function of the local temperature only. If $\tau_{\exp} \ll \tau_a$, the left-hand derivative in the ionization equation must be set to zero, and one has the single conservation of flux equation

$$\frac{1}{r^2}\frac{d}{dr}(r^2 n_i u_i) = 0 ,$$

or $r^2 n_i u_i$ is constant. Thus, the ionization and recombination processes are no longer important.

At the low densities which prevail throughout interplanetary space, the scale times for atomic processes such as ionization or recombination are extremely long. A comparison of these scale times with the time required for the solar wind to flow through a density scale height leads to the conclusion (e.g., Brandt and Hodge, 1964; Delache, 1965; Brandt and Hunten, 1966; Cloutier, 1966) that the ionization state of the expanding solar plasma will not change in interplanetary space but is entirely determined by initial conditions in the corona. This conclusion can be made more specific by considering the pertinent scale times as functions of position in the corona. Consider oxygen as a particular example. Tucker and Gould (1966) have shown that the most abundant oxygen ions at corona temperatures will be O^{+6} and O^{+7}. Figure 15 shows τ_6, the time for collisional ionization of O^{+6}; τ_7, the time for radiative (including dielectronic) recombination of O^{+7}; and τ_e, the time for expansion through a density scale height [densities, flow speeds, and temperatures from the one-fluid model of Whang and Chang (1965) have been used]. The scale times τ_6 and τ_7 are much less than τ_e near the base of the corona, indicating that the ionization state of oxygen will adjust to local conditions in this region. However, these scale times increase very rapidly with increasing heliocentric distance (due to their inverse dependence on density, which falls off rapidly as a function of heliocentric distance), becoming equal to the expansion time at $r \approx 1.2 r_\odot$ and much larger shortly above this level. One thus expects the ionization state of the expanding coronal material to remain constant beyond a heliocentric distance of $\approx 1.5 r_\odot$.

A momentum equation and an energy equation should also be written for each ion. The solution of this rather formidable collection of equations has not been undertaken; however, the oxygen flux conservation equations (Equation 32) have been integrated for the expanding corona by Hundhausen et al. (1968a and 1968b), using densities, flow speeds, and temperatures from the one-fluid coronal model, and by Kozlovsky (1968), using a constant expansion speed. The solid lines in Figure 16 show the solution obtained in the former integration, and the dashed lines show the

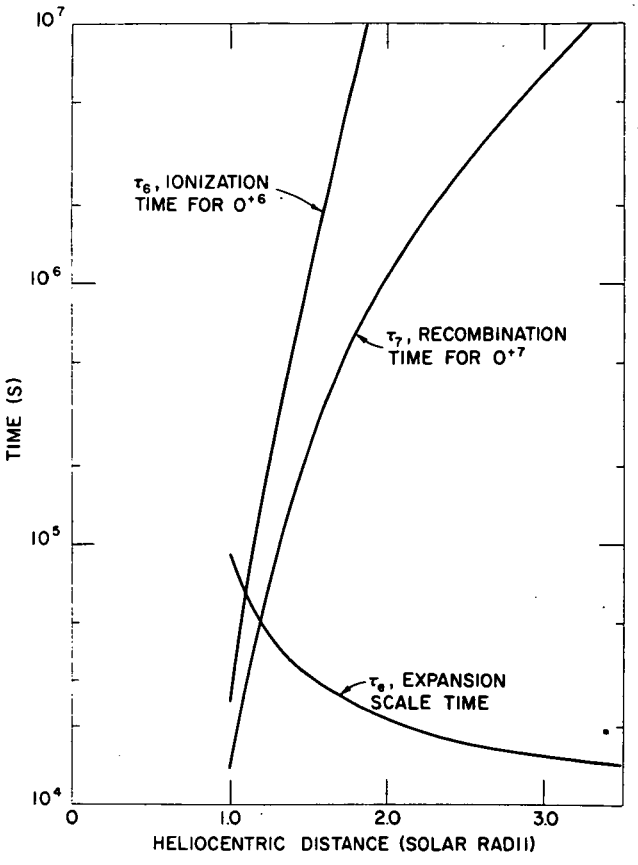


Figure 15.—Scale times for the collisional ionization of O^{+6}, the radiative recombination of O^{+7}, and the flow of material through a density scale height, as functions of heliocentric distance. Coronal parameters from the one-fluid model of Whang and Chang were used (Hundhausen et al., 1968a, 1968b).

"static" solution ($u_i \equiv 0$) which depends on the local temperature only. As expected, little change occurs in the ionization state beyond $r \approx 1.5r_\odot$; the relative abundance of O^{+6} and O^{+7} at large distances is equal to that given by the static solution at a heliocentric distance of $1.2r_\odot$ and agrees with the interplanetary observation that $^{16}O^{+6}$ is the most common oxygen ion in the solar wind.

Thus, the temperature implied by the observed ionization state of a given element in the solar wind can be directly related to the temperature deep in the corona. This relationship opens the interesting possibility of using a measured "ionization temperature" as a tracer of coronal source regions of the solar wind. As an example, VELA 3 positive-ion spectra obtained on March 22 and 23, 1966, indicate that $^{16}O^{+7}$ is several times more abundant than $^{16}O^{+6}$. The calculations of

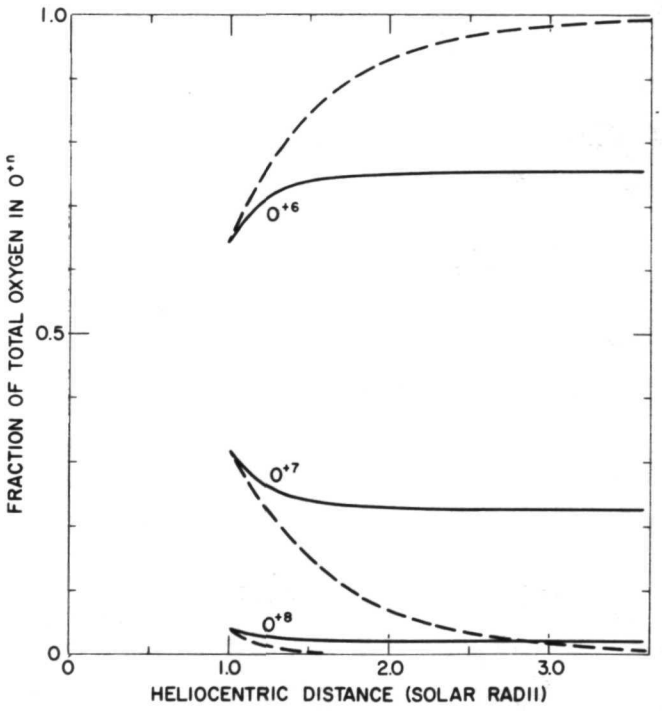


Figure 16.–Ionization state of oxygen as a function of heliocentric distance (solid lines) predicted by the model of Hundhausen et al. (1968a, 1968b). The dashed lines show the ionization state for a static corona at the appropriate local temperature.

Hundhausen et al. (1968b) and Kozlovsky (1968) show that this implies a coronal source temperature of 3×10^6 to 4×10^6 K, well above the accepted normal coronal temperature. Figure 17 shows an H_α spectroheliogram from March 22, 1966. Only one large active region, above which such high coronal temperatures would be expected, is present. The 9.1-cm radio observations also identify this region as the only large hot spot in the visible corona. It is thus highly plausible to regard this active region as the source of the observed $^{16}O^{+7}$-rich solar wind.

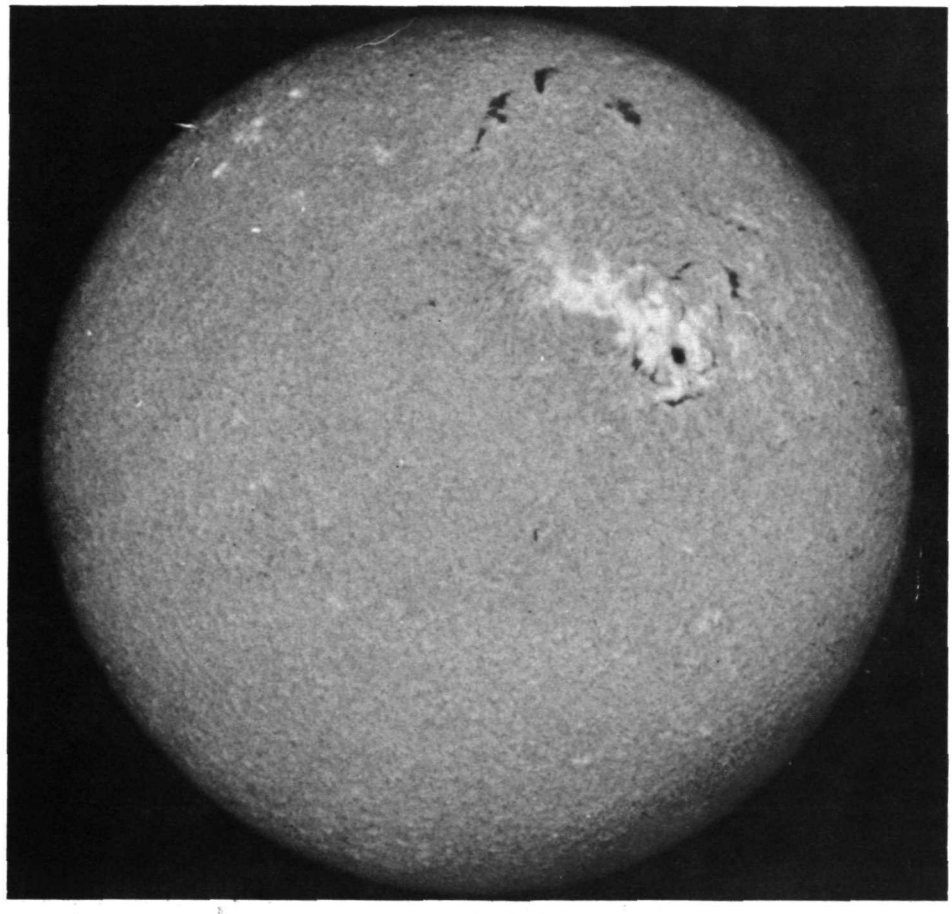


Figure 17.—An H_α spectroheliogram from March 22, 1966.

VII. PROPAGATION OF FLARE-ASSOCIATED SHOCK WAVES IN INTERPLANETARY SPACE

The discussion of the coronal expansion given in Sections V and VI was based on the assumptions of spherical symmetry and steady flow. Even with these simplifying assumptions, solving the fluid equations is no mean task. Analysis of nonspherical, nonsteady systems introduces new dependent (i.e., nonradial flow velocity components) and new independent (i.e., the azimuthal and polar angles) variables, and tremendously complicates the fluid equations.

Nonetheless, observations of both the corona and the interplanetary plasma reveal considerable spatial structure and temporal variation. This fact is beautifully illustrated in Figure 18, a coronal photograph obtained during the 1970 Eclipse Expedition of the High Altitude Observatory. A radially symmetric neutral density filter was used to compensate for the steep decline in coronal brightness with distance from the limb. Even a theorist, highly motivated to resist the introduction of new complications into coronal expansion models, would have considerable difficulty in describing the features of Figure 18 as spherically symmetric.

If one is forced to consider the incorporation of nonspherical, nonsteady features into coronal expansion models, the logical first step is to consider one complication at a time (i.e., either a steady but nonsymmetric flow or a spherically symmetric but nonsteady flow). We will limit our discussion to a phenomenon in the latter class: the propagation through the solar wind of a spherically symmetric disturbance produced by a solar flare. The flare, perhaps the most spectacular manifestation of solar activity, is observed optically as a sudden, localized brightening of a chromospheric region. Indirect evidence has been available for over a century that solar material is ejected into interplanetary space by solar flares. Interplanetary plasma probes observe the passage of an outward propagating shock wave about 2 days after some flares. This shock is presumably produced as the ejected material interacts with the solar wind already present in the interplanetary region. As the energy in such a shock wave is greater than that in the "ambient" plasma and magnetic field, one expects a nearly spherical disturbance at large heliocentric distances; thus, this system can be treated as approximately spherically symmetric.

The theoretical treatment of propagating solar wind disturbances requires integration of the time-dependent equations for mass, momentum, and energy

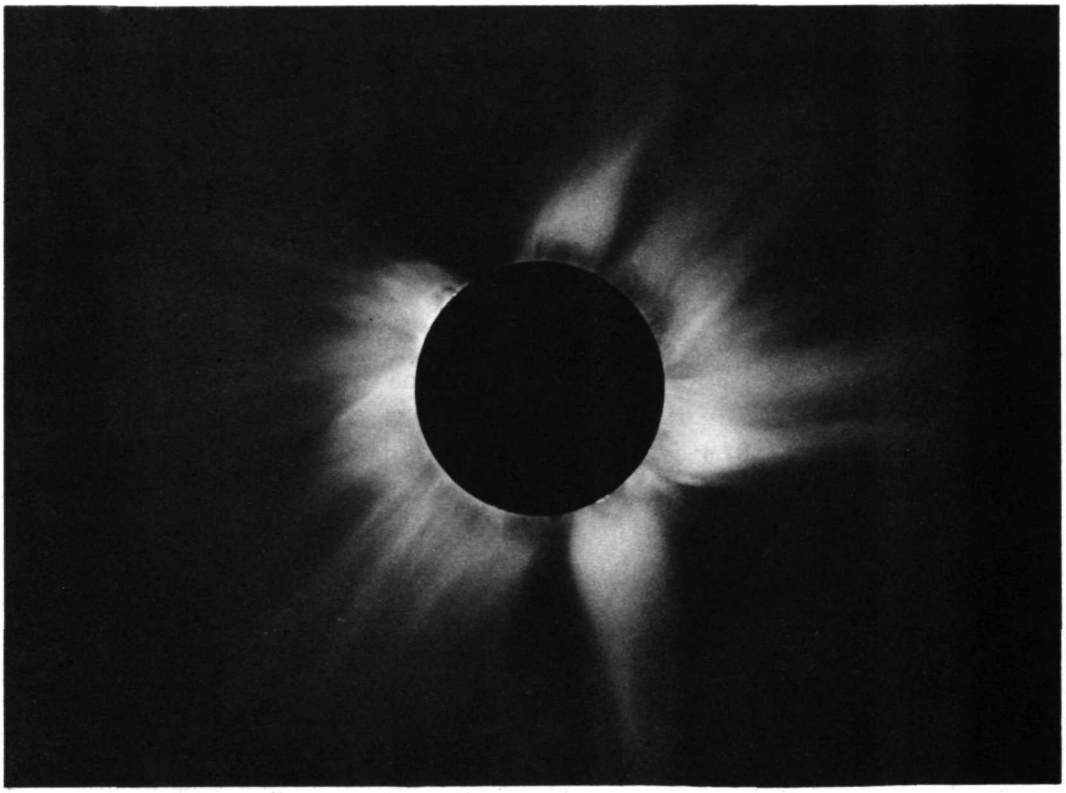


Figure 18.—The corona on March 7, 1970, photographed during the 1970 Eclipse Expedition of the High Altitude Observatory, National Center for Atmospheric Research (courtesy of G. Newkirk, Jr.).

conservation. Even in the one-fluid formulation which has been employed in all models to date, the presence of a second independent variable so complicates the integrations that numerous additional simplifications must be made. Given spherical symmetry, the mass and momentum equations are

$$\frac{\partial n}{\partial t} + \frac{1}{r^2}\frac{\partial}{\partial r}(nur^2) = 0 \tag{33}$$

and

$$nm\left(\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial r}\right) = -2\frac{\partial}{\partial r}(nkT) - \frac{nmGM_\odot}{r^2} \tag{34}$$

If all energy sources, including heat conduction, are neglected beyond the base of the corona, the energy equation is

$$\frac{\partial}{\partial t}(nE) + \frac{1}{2}\frac{\partial}{\partial r}(nEur^2) = -\frac{2}{r^2}\frac{\partial}{\partial r}(nkTur^2) - \frac{nmGM_\odot}{r^2}u\,, \tag{35}$$

where $E = 1/2mu^2 + 3kt$ is the energy per particle. Solutions of this system of equations have been obtained by two different techniques.

A. Similarity Solutions

If the gravitational terms are neglected, Equations 33 through 35 become

$$\frac{\partial n}{\partial t} + \frac{1}{r^2}\frac{\partial}{\partial r}(nur^2) = 0\,,$$

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial r} = -\frac{1}{\rho}\frac{\partial}{\partial r}P\,,$$

and

$$\frac{\partial}{\partial t}\left(\frac{P}{\rho^{5/3}}\right) + u\frac{\partial}{\partial r}\left(\frac{P}{\rho^{5/3}}\right) = 0\,,$$

where $P = 2nkT$ is the pressure. The last of these equations simply gives the well-known relationship between pressure and density for an adiabatic expansion, which was implicitly assumed here when all energy sources were neglected. Progressive wave solutions with a shock at the leading edge of the transient disturbance can be found (by numerical quadrature) as a function of the similarity parameter $\eta = tr^{-\lambda}$ (see Parker, 1963b) if the flow speed and internal energy of the ambient material through which the wave propagates are assumed to be negligible. In the case where the density of the ambient material varies as $1/r^2$, as for a constant (but negligible) solar wind speed, these solutions correspond to waves in which the total energy varies as the $3/\lambda - 2$ power of time. Figure 19 shows the density versus radius (in terms of R_1, the position of the shock) given by solutions with different values of λ. The shock (and any other feature of the solution) moves with time

according to the similarity law $r = R_1 t^{1/\lambda}$; the shock speed thus varies as $1/r^{\lambda-1}$. The case $\lambda = 3/2$ is of special interest, since it implies a constant total energy in the wave. This corresponds to an impulsively generated disturbance, generally referred to as a "blast wave". The cases $\lambda < 3/2$ imply the steady addition of energy to the disturbance and are referred to as "driven disturbances".

For detailed discussions of similarity solutions applied to interplanetary shock propagation, see Parker (1961, 1963a, and 1963b), Simon and Axford (1966), and Lee and Balwanz (1968). For future use, we note the following results:

(1) All properties of the blast wave solutions are determined by the total energy W in the wave. In particular, the transit time T or shock speed V at a given radius r are related to W by

$$T^2 = 2\pi n m_p r^5/3W$$

and (36)

$$V^2 = 2W/3\pi n m_p r^3 .$$

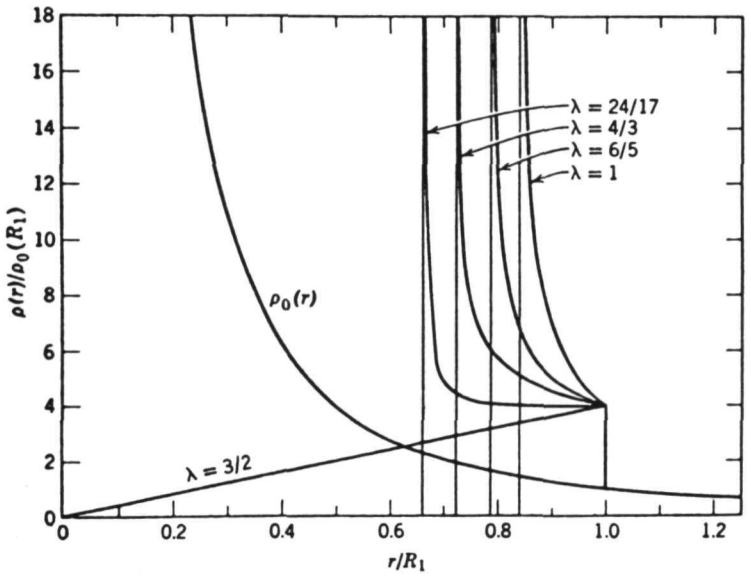


Figure 19.–Density as a function of heliocentric distance in the similarity shock wave solutions of Parker (1963a). The term R_1 is the position of the shock at the leading edge of the disturbance.

Both the density and flow speed observed at any position decrease monotonically with time after passage of the shock.

(2) For the "driven" solutions both the density and flow speed continue to increase with time after the abrupt changes at the shock.

B. Numerical Solutions

Equations 33 through 35 have been integrated numerically, with no further simplifying assumptions, by Hundhausen and Gentry (1969a and 1969b). A shock of arbitrary strength is introduced into an adiabatic, steady, or ambient solar wind by changing fluid parameters at 1/10 AU; post-shock parameters are maintained for an arbitrary duration Δ after which the ambient conditions are resumed. The following results are given for comparison with the similarity solutions and for future use in interpretation of the observations:

(1) For $\Delta \lesssim 10^{-2}\ T$, where T is the transit time of the shock wave to an observer (for example, at 1 AU) , all properties of the disturbance observed at the latter distance are functions only of the total energy W in the wave. These solutions correspond to the blast wave limit $\lambda = 3/2$ in the similarity theory described above. Figure 20 shows the numerical solution (density and flow speed as functions of heliocentric distance) for a blast wave with total energy $W = 1.6 \times 10^{31}$ ergs propagating in an ambient solar wind with a flow speed of 400 km-s^{-1} and a density of 12 cm^{-3} at 1 AU. Note that again the flow speed and density decrease monotonically in a large region behind the shock. Figure 21 shows the relationship between transit time to 1 AU and energy W for the numerical blast wave solutions in an ambient solar wind with the properties given above.

(2) Figure 22 shows the numerical solution obtained when the post-shock conditions have been maintained continuously at 1/10 AU. A new steady-state solar wind has formed for $r \lesssim 0.8$ AU, with a flow speed of 550 km-s^{-1}. Between this steady-state and the original ambient solar wind is a compressed shell with a shock at the leading edge. This corresponds to the driven solution $\lambda = 1$ in the similarity theory described above. Both the flow speed and density again continue to increase behind the abrupt changes at the shock. This type of solution, with post-shock increases in some flow parameters, would be observed at a heliocentric distance r if the duration of the initial disturbance Δ is as long as $\approx 5 \times 10^{-2}\ T$, where T is the transit time of the shock to r. A given shock speed or transit time at r implies a higher energy W for a driven disturbance than for a blast wave.

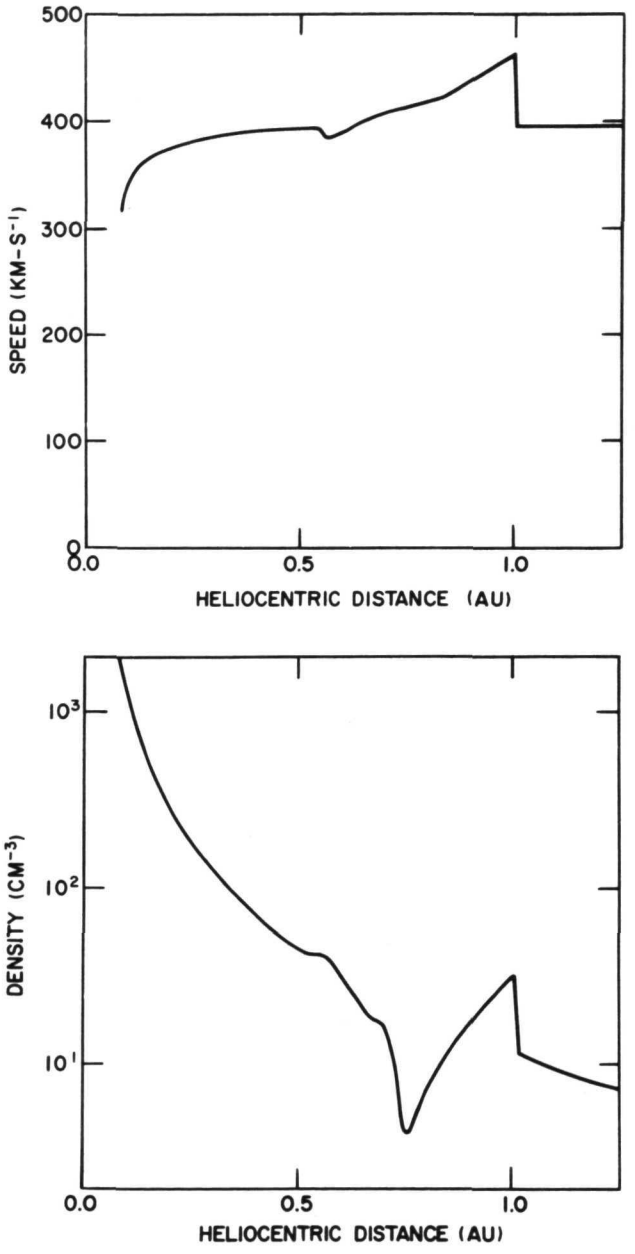


Figure 20.–Flow speed and density as functions of heliocentric distance in a numerical blast wave (i.e., impulsive disturbance) solution of Hundhausen and Gentry (1969a).

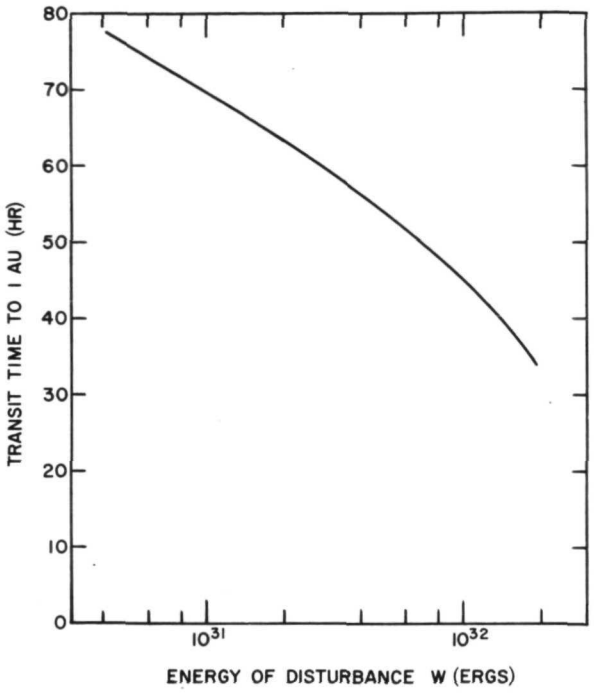


Figure 21.–Transit time to 1 AU for a blast wave with energy W propagating in a solar wind with $n = 12$ cm^{-3} and $u = 400$ km-s^{-1} at 1 AU (Hundhausen and Gentry, 1969a).

C. Energy Release by Solar Flares

An interesting application of these shock propagation models is their use to estimate the energy released into the corona gas (and not radiated) by the flare and the time scale of this process. The unique dependence of blast wave properties on the total energy W predicted by both the similarity and numerical models immediately suggests such an interpretation of observations made near the interplanetary shock (Dryer and Jones, 1968; Hundhausen and Gentry, 1969a; Korobeinikov, 1969). The association of flares with observed shocks gives an average transit time to 1 AU of ≈ 55 hours (Hundhausen, 1970). For an average solar wind density of 8 proton-cm^{-3} at 1 AU, $\langle T \rangle = 55$ hours implies an average flare energy $\langle W \rangle = 5 \times 10^{32}$ ergs, using Equation 36 from similarity theory, or $\langle W \rangle = 3 \times 10^{31}$

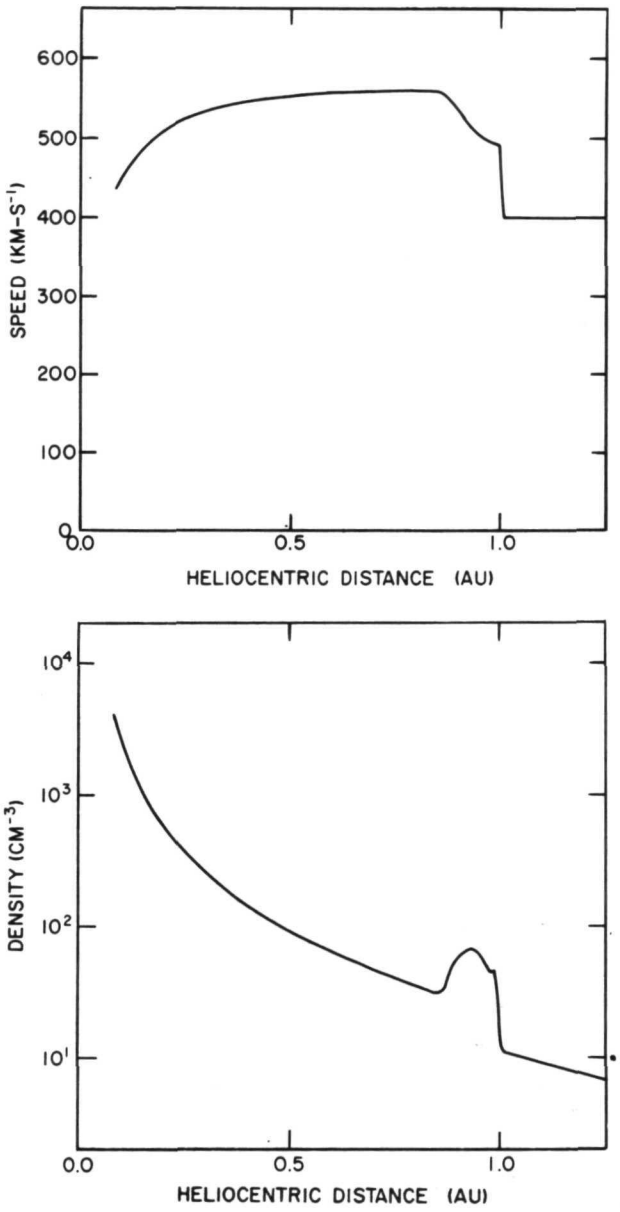


Figure 22.–Flow speed and density as functions of heliocentric distance in a numerical driven (i.e., long-duration) wave solution of Hundhausen and Gentry (1969a).

ergs, using Figure 22 from the numerical integrations. Similar values result from the observed average shock speed $\langle V \rangle = 500$ km-s^{-1}. The difference between the energy estimates based on similarity and numerical methods is largely due to the neglect of the ambient solar wind speed in the former. For driven disturbances, the energy $\langle W \rangle$ implied by the same $\langle T \rangle$ or $\langle V \rangle$ would be several times larger. Such model-dependent estimates of the energy in flare-associated solar wind disturbances were until recently the only source of such information. The recent derivation of independent estimates by integration of observed energy fluxes* (Hundhausen et al., 1970) provides a check on the model-based values. The observed average energy at 1 AU was $\langle W \rangle = 5 \times 10^{31}$ ergs. The estimate based on similarity theory is an order of magnitude too high, again showing the effect of assuming a negligible ambient solar wind speed. The estimate based on the numerical solutions is in good agreement with the observed value. Shock waves which resemble both blast waves and driven waves have been observed. This implies that the energy deposition by solar flares occurs over a time scale of ≈ 15 min (less than $10^{-2} \langle T \rangle$) to several hours.

ACKNOWLEDGMENTS

The author wishes to thank Dr. D. S. De Young, of the National Radio Astronomy Observatory, and Dr. R. A. Gentry, of Las Alamos Scientific Laboratory, for discussions of the material covered in his paper; and Dr. G. Newkirk, Jr., of the High Altitude Observatory, for permission to use the March 7, 1970, eclipse photograph.

REFERENCES

Athay, R. G., *Astrophys. J.* 145:784, 1966.

Bame, S. J., Hundhausen, A. J., Asbridge, J. R., and Strong, I. B., *Phys. Rev. Lett.* 20:393, 1968.

Bame, S. J., Asbridge, J. R., Hundhausen, A. J., and Montgomery, M. D., to be published in *J. Geophys. Res.* 75, 1970.

Bird, G. A., *Astrophys. J.* 141:1455, 1965.

Brandt, J. C., and Hodge, P. W., *Solar System Astrophysics,* McGraw-Hill Publications, New York, 1964.

Brandt, J. C., *Astrophys. J.* 143:265, 1966.

Brandt, J. C., and Hunten, D. M., *Planet. Space Sci.* 14:95, 1966.

*See Chapter 5, "The Interplanetary Plasma", by Ogilvie.

Brandt, J. C., Wolff, C., and Cassinelli, J. P., *Astrophys. J.* 156:1117, 1969.
Brandt, J. C., *Introduction to the Solar Wind,* W. H. Freeman and Co., San Francisco, 1970.
Chamberlain, J. W., *Astrophys. J.* 131:47, 1960.
Chapman, S., *Smithsonian Contrib. Astrophys.* 2:1, 1957.
Cloutier, P., *Planet. Space Sci.* 14:809, 1966.
Delache, P., *Compt. Rend.* 261:643, 1965.
Delache, P., *Ann. Astrophys.* 30:827, 1967.
Dryer, M., and Jones, D. L., *J. Geophys. Res.* 73:4875, 1968.
Forslund, D. W., *J. Geophys. Res.* 75:17, 1970.
Geiss, J., Hirt, P., and Lentwyler, H., *Solar Phys.* 12:458, 1970.
Gentry, R. A., and Hundhausen, A. J., *Trans. Amer. Geophys. Union* 50:302, 1969.
Grzedzielski, S., *Astrophys. Space Sci.* 3:139, 1969.
Hartle, R. E., and Sturrock, P. A., *Astrophys. J.* 151:1155, 1968.
Hartle, R. E., and Barnes, A., *J. Geophys. Res.* 75:6915, 1970.
Holzer, T. E., and Axford, W. I., *Ann. Rev. Astron. Ap.* 8, 1970 (in press).
Hundhausen, A. J., Asbridge, J. R., Bame, S. J., Gilbert, H. E., and Strong, I. B., *J. Geophys. Res.* 72:5265, 1967.
Hundhausen, A. J., Gilbert, H. E., and Bame, S. J., *Astrophys. J.* 152:L3, 1968a.
Hundhausen, A. J., Gilbert, H. E., and Bame, S. J., *J. Geophys. Res.* 73:5485, 1968b.
Hundhausen, A. J., and Gentry, R. A., *J. Geophys. Res.* 74:2908, 1969a.
Hundhausen, A. J., and Gentry, R. A., *J. Geophys. Res.* 74:6229, 1969b.
Hundhausen, A. J., *Rev. Geophys. Space Sci.* (in press), 1970.
Hundhausen, A. J., Bame, S. J., Asbridge, J. R., and Sydoriak, S. J., *J. Geophys. Res.* 75:4643, 1970.
Jokipii, J. R., in *The Solar Wind,* R. J. Mackin and M. Neugebauer, eds., Pergamon Press, New York, 1966, p. 215.
Kopp, R., Harvard University Scientific Report No. 4, AFCRL-68-0312, Cambridge, Massachusetts, 1968.
Korobeinikov, V. P., *Solar Phys.* 7:463, 1969.
Kozlovsky, B. Z., *Solar Phys.* 5:410, 1968.
Kuperus, M., and Athay, R. G., *Solar Phys.* 1:361, 1967.
Kuperus, M., *Space Sci. Rev.* 9:713, 1969.
Lee, T. S., and Balwanz, W. W., *Solar Phys.* 4:240, 1968.
Montgomery, M. D., Bame, S. J., and Hundhausen, A. J., *J. Geophys. Res.* 73:4999, 1968.
Nakada, N. P., *Solar Phys.* 7:302, 1969.
Neugebauer, M., and Snyder, C. W., *J. Geophys. Res.* 71:4469, 1966.
Noble, L. M., and Scarf, F. L., *Astrophys. J.* 138:1169, 1963.
Parker, E. N., *Astrophys. J.* 132:1445, 1960.
Parker, E. N., *Astrophys. J.* 133:1014, 1961.
Parker, E. N., in *The Solar Corona,* J. W. Evans, ed., Academic Press, New York, 1963a, p. 11.
Parker, E. N., *Interplanetary Dynamical Processes,* Interscience Publishers, New York, 1963b.
Parker, E. N., *Astrophys. J.* 139:93, 1964.

Scarf, F. L., and Noble, L. M., *Astrophys. J.* 141:1479, 1965.
Simon, M., and Axford, W. I., *Planet. Space Sci.* 14:901, 1966.
Spitzer, L., and Harm, R., *Phys. Rev.* 89:977, 1953.
Spitzer, L., *Physics of Fully Ionized Gases,* Interscience Publishers, New York, 1962.
Sturrock, P. A., and Hartle, R. E., *Phys. Rev. Lett.* 16:628, 1966.
Tucker, W. H., and Gould, R. J., 1966, *Astrophys. J.* 144:244, 1966.
Urch, I. H., *Solar Phys.* 10:219, 1969.
Vitense, E., *Z. Astrophys.* 32:135, 1953.
Weber, E. J., and Davis, L., *Astrophys. J.* 148:217, 1967.
Whang, Y. C., and Chang, C. C., *J. Geophys. Res.* 70:4175, 1965.
Yeh, T., *Planet. Space Sci.* 18:199, 1970.

Page intentionally left blank

CHAPTER 5
THE INTERPLANETARY PLASMA

K. Ogilvie
Goddard Space Flight Center
Greenbelt, Maryland

I. INTRODUCTION

In the study of the interplanetary medium, as in the study of plasmas in the laboratory, the purpose is to measure the properties of the plasma and compare them with theoretical predictions. If reasonable agreement is reached, knowledge of these properties can be used to make statements of astrophysical interest about the medium and its interaction with macroscopic bodies. The interplanetary medium is of great interest, as the only example of an astrophysical plasma which is at present accessible. In Table 1, we contrast the properties of the interplanetary and interstellar plasmas with those of the solar corona and two examples of laboratory plasmas, the Q-plasma and the thermonuclear plasma. The plasma produced by the Q-machine (D'Angelo, 1969) is a stable quiescent plasma made by contact ionization of cesium with hot tungsten plates. It is probably as close to being in equilibrium as any plasma available for laboratory study. The thermonuclear plasma represents that region of parameter space in which thermonuclear reactions could be self-sustaining; at present this has not yet been realized. From this table, one can readily see how the study of the astrophysical cases complements that of terrestrially generated ones. In particular, we note two points which bear upon the necessary diagnostic methods:

(1) The Debye length

$$\lambda_D = (kT/4\pi ne^2)^{1/2},$$

which characterizes the range of interaction in the plasma, is of the order of several meters in the interplanetary medium. Collective effects are important in a plasma over length scales greater than λ_D. When measurements are carried out using an apparatus much larger than λ_D, the properties of the plasma must be deduced from

Table 1.–Properties of some plasmas.

Plasma	Density n (cm^{-3})	Temperature T (K)	Magnetic Field B (gauss)	Electron Plasma Frequency ω_{pe} (s^{-1})	Debye Length λ_D (cm)	Collision Frequency ν_c (s^{-1})	Electron Cyclotron Frequency Ω_e (s^{-1})	Ion Cyclotron Frequency Ω_i (s^{-1})
Interstellar	1	10^4	10^{-6}	6×10^4	700	6×10^{-4}	20	10^{-2}
Interplanetary (1 AU)	5	4×10^4	5×10^{-5}	12×10^4	600	4×10^{-4}	40	2×10^{-2}
Corona	10^6	10^6	10^{-4}	6×10^7	7	0.6	2×10^3	1
Q-plasma	10^8	10^5	10^2	6×10^8	0.2	10^3	2×10^9	10^6
Thermonuclear	10^{16}	10^8	10^5	6×10^{12}	7×10^{-4}	5×10^6	2×10^{12}	10^9

collective macroscopic effects. This case corresponds, for example, to probe measurements in laboratory plasma diagnostics. In interplanetary space, however, all satellites and measuring instruments are on a scale $<< \lambda_D$, and so the plasma properties are found by determining the velocity distribution function by counting particles in a given direction and velocity interval. Thus, the apparatus used for the study of these plasmas is very similar to that used in atomic scattering studies in the appropriate energy range.

(2) The frequency of collisions between particles in the interplanetary medium is so low that the plasma can be considered to be collisionless, at least at a distance from the Sun greater than about 0.5 AU. This condition is satisfied in the absence of restrictions on the length scale or time duration of the experiment. The interplanetary medium is therefore ideal for the study of the properties of a collisionless plasma.

II. DIAGNOSTIC METHODS

In the laboratory, various plasma diagnostic techniques are used; examples are spectroscopy in various wavelength intervals, probe measurements, and microwave transmission measurements. These observations often have to be made during the short interval of time for which the plasma is available. In the interplanetary medium a knowledge of the properties of the plasma is gained by directly determining the distribution function as a function of time, species, and position. In practice, we determine an approximation to the velocity distribution function for as many species as possible, usually hydrogen and helium, over a limited range of velocity and angle at one or two points in space. We are able to observe changes with time by making observations at regular intervals.

The interactions between particles in a collisionless plasma involve the wave motions which it can support. A knowledge of these is essential to a complete understanding of the plasma, and the particle properties must not be stressed too heavily. However, the present experimental situation is that most of the available knowledge has been gained through particle experiments. Diagnostic methods for wave observations will not be discussed here, but some results obtained using such methods will be referred to below.

The velocity distribution function $f_i(\mathbf{v}, \mathbf{r}, t)$, characteristic of a particular species i, is approximated by counting particles, or measuring currents, over a range of velocity and angle. The completeness with which f may be determined forms one

criterion of the success of an experimental design, although completeness competes with time resolution. The latter must be sufficiently good that significant changes in f do not take place during a determination. The function f_i represents the density of particles of the ith species in the differential interval $d\mathbf{r}\, d\mathbf{v}$ at position $\mathbf{r}$ and velocity $\mathbf{v}$. A complete knowledge of f_i and its time variations as a function of space would allow the plasma properties to be deduced. However, at the present time, due to limitations of experiment and interpretation, the solar wind is described in terms of the macroscopic approximation, in which velocity moments of f define a density, bulk speed, pressure tensor, and heat flux for a given species. Since the plasma is collisionless, the species do not interact with each other by collisions. If the distribution is in equilibrium, or does not change rapidly with time, it can be characterized by a temperature.

The electron temperature, for example, is defined in terms of a distribution function obtained from electron measurements by techniques to be discussed later. Temperatures obtained in this way are tensor quantities; the presence of the interplanetary magnetic field moving with the plasma causes the particle distribution to be anisotropic about the direction defined by the field, and this is incorporated by defining two temperatures, $T_\parallel$ and $T_\perp$.

Bulk speeds V_b^i, defined by the equation

$$nV_b^i = \int f_i \mathbf{v}\, d\mathbf{v}\,,$$

are obtained for each species and have been shown to be the same for electrons, protons, and helium ions, at least for macroscopic time intervals.

This macroscopic fluid theory has been successful in describing the solar wind up to the present time. To go to the next step and incorporate kinetic theory would require a relatively large extension in the power of the experiments. Time resolution would have to be drastically improved, from seconds to milliseconds or better, and more detailed knowledge of the properties of f_i, especially in the high velocity tail region, would be required. Rather than speculate on these matters, we shall discuss how the macroscopic theory is applied to the experimental results and then discuss the nature of these results, which will show where things stand at present.

In order to illustrate the taking of measurements by detectors, we use a vector $\mathbf{U} = (U, \theta, \phi)$ which was introduced by Vasyliunas (1969) and is illustrated in Figure 1. The vector $\mathbf{U}$ represents a detector set to record particles with speed U and aligned so that the normal to its sensitive area makes angles θ and ϕ to the xy- and

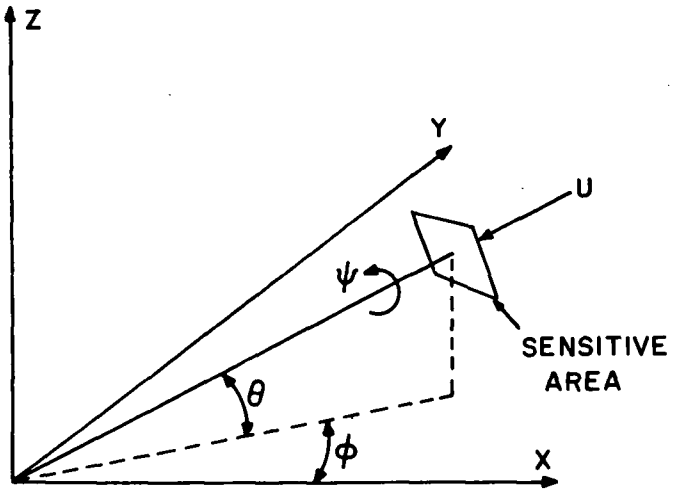


Figure 1.—Definition of the vector $\mathbf{U} = (U, \theta, \phi)$.

xz-plane, respectively, of a fixed coordinate system. We assume that the third angle ψ in Figure 1 is not significant, as it would not be if the sensitive area were round. A common situation is one where the z-axis is the spin axis of a spacecraft and $\theta = 0$. Figure 2, then, shows how the observations are made in four cases: Part A shows three values of |U| sampled at 18 angles during each spin, the value of |U| being changed between spins; B shows a similar arrangement, with observations unequally spaced in angle; and D shows observations where a number of values of |U| are sampled during a small angular rotation of the spacecraft. Having now defined and illustrated U, we consider $C(\mathbf{U})$, the counting rate, or current, obtained during a particular observation.

It is the product of the number of particles having velocity U, the component of U perpendicular to the sensitive area of the detector, the sensitive area, and the transmission function of the instrument. If the species are separated, this resulting quantity refers to a given species; otherwise it is a sum over species:

$$C(\mathbf{U}) = A \sum_i \int d\mathbf{v}\, v_n f_i(\mathbf{v}) G(\mathbf{U},\mathbf{v}) \,. \tag{1}$$

The integral is over velocity space. If $H(\mathbf{U},\mathbf{v}')$ is the transmission function of the instrument in a coordinate system fixed with respect to the instrument, which is assumed to have been previously measured or calculated,

$$G(\mathbf{U},\mathbf{v}) = H(R \cdot \mathbf{v},\mathbf{U}) ,$$

where R is a matrix which carries out the change of coordinates. It is now clear that

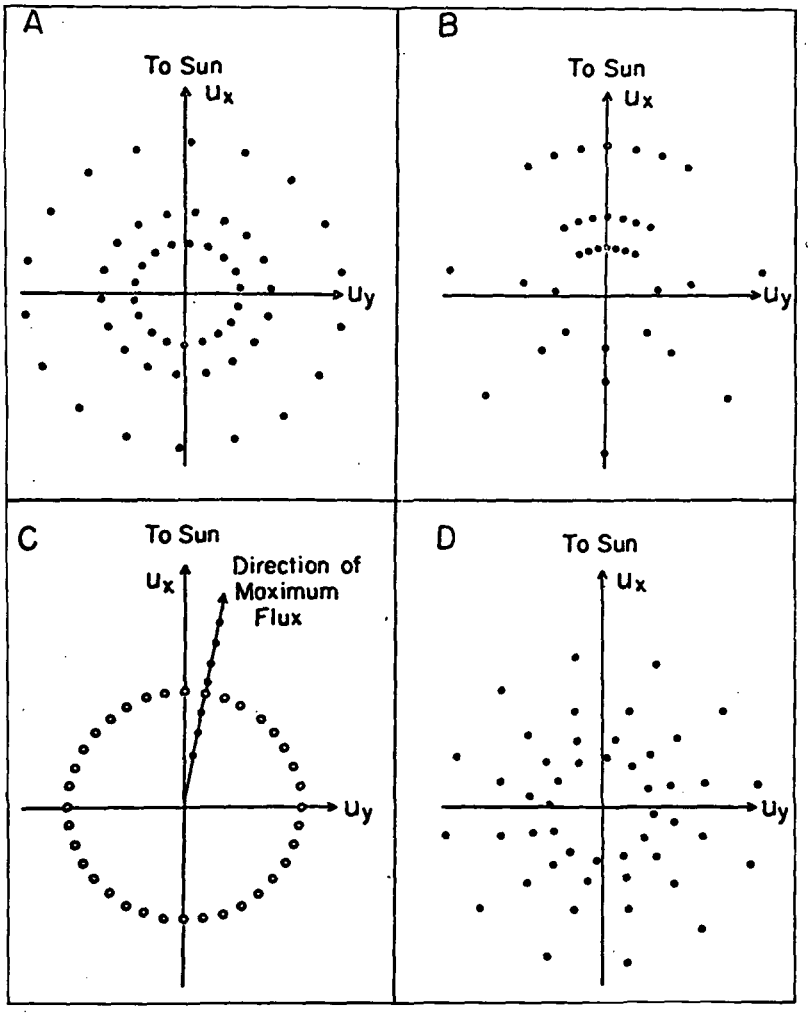


Figure 2.—Illustration of data sampling methods used in space plasma experiments. An explanation is given in the text.

although $f(\mathbf{v})$ is the fundamental quantity we require to know, $C(\mathbf{U})$ is the measured quantity. Going from one to the other is not an entirely trivial problem for several reasons. At normal times in the solar wind, the distribution function is now known to have the gross characteristics of a convected bi-Maxwellian distribution, and it is often the small detailed departures from this form which are of interest. The instrumental response must thus be known to high accuracy. The transformation from the fixed to the moving frame often requires approximations. There exist methods of inverting Equation 1 and obtaining $f(\mathbf{v})$, but in practice two other methods have been used. In the first, a form of $f(\mathbf{v})$, for example, a convected bi-Maxwellian, is assumed and fitted to the observations. The deduced parameters of this distribution are then taken to be characteristic of the distribution function. In the second, suitable for a differential detector, the counting rate of each given velocity interval and angle is expressed in terms of dn/dv, the differential density. Such a procedure requires, in most cases, approximations associated with the coordinate transformation. These differential densities are fitted to a curve which becomes the approximation to the distribution function. The fluid quantities are determined by calculation of the moments of this empirical curve. Another point that cannot be overlooked is that a detector may generate every few seconds a set of data from which a distribution function could be calculated. The reduction procedure must occupy a much shorter time than the recording time if it is to be useful. It may be necessary to have two methods: one to produce approximate results quickly and which is applied to all the data; and a second, accurate method which is applied less frequently or only to selected data. The approximate method may be carried out on the spacecraft by a special-purpose computer using a preset algorithm.

Fluid parameters determined by different instruments in the same region of the interplanetary medium now agree quite well in spite of the differences in method of operation and data reduction. The degree of accuracy to be expected is 3 percent for bulk speed (10 km-s^{-1}), 25 percent for density (absolutely, but somewhat better relatively), and better than 50 percent for temperature. Increasing the accuracy of higher moments will permit the study of details such as high velocity "tails" on the velocity distribution and allow species separation when it is not carried out directly in the instrument. To accomplish this, the relevant velocity range should be broken into as many small differential intervals as is consistent with the needed time resolution. In other words, the width of the instrument response function should be small relative to the width of the expected velocity distribution. It is an advantage if the differential velocity intervals are contiguous and leave no gaps.

III. OBSERVATIONS OF DIFFERENT IONIC SPECIES

A very important problem at present concerns the relative abundances and the relative populations of charge states in the solar wind. The separation of $^4He^{++}$ from H^+ has been carried out both with an instrument (Ogilvie et al., 1968) and analytically (Robbins et al., 1970; Neugebauer and Snyder, 1966), and the results of the studies indicate large variations in the relative abundance of helium which seem to be linked to solar activity. Because of the very long scale times for recombination in the interplanetary medium and their very rapid increase with heliocentric distance as the solar wind is accelerated, the charge states of ions in the medium at 1 AU are characteristic of the corona at a heliocentric distance of 1.5 to 2.0 solar radii. These observations show promise of allowing the study of the variation of relative abundances in the corona and the processes producing this variation (Hundhausen, 1970). Studies at 1 AU in principle can provide information on the temperature of the region in the corona from which individual ions were accelerated. The most suitable constituent to use for this study is O, the next most abundant element after helium. The problem of observation of all the heavy ions other than $^4He^{++}$ is a difficult one, since the intensities of these ions is low relative to that of H and He. In principle, since $\frac{1}{2}mv^2/z \approx$ constant $\times\, m/z$, dispersion in energy per unit charge alone will separate ions with different masses. However, V_{th}/V_b may not always be small enough to prevent the overlapping of several species in a single energy per unit charge observation. This is illustrated in Figure 3, which shows the expected distribution of flux with energy per unit charge for a hypothetical solar wind made up as shown in Table 2.

This medium is assumed to be moving with $V_b = 5 \times 10^7$ cm-s^{-1}, and the curves shown are for proton temperatures of 10^4, 5×10^4, and 2×10^5 K, respectively; all the ions are assumed to have the same velocity distribution function. At the lowest ion temperature, the peaks are separated well enough for abundance measurements to be made. Such measurements have already been carried out with some success (Bame et al., 1968), but a complete solution of the problem reduces to that of making a mass spectrometer without an ionizer, which will operate with sufficient resolution using the solar wind as a source. The mass per charge range required is from 1 (H^+) to approximately 5.0, or from 2.0 to 5.0 if $^4He^{++}$ is used as a standard, and it must accept ions having a velocity range of $\pm V_{th}/V_b$, of the order of 10 percent. Thus, the instrument must have a very different design from that of a

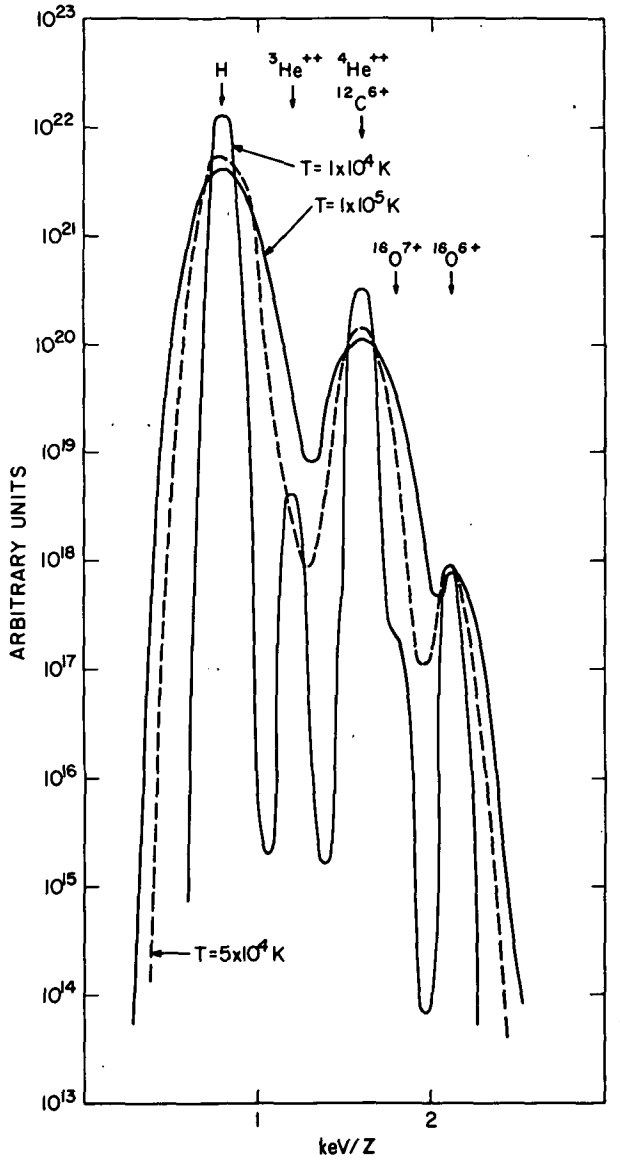


Figure 3.–The distribution of flux as a function of energy per unit charge for a hypothetical solar wind with the same convected Maxwellian velocity distribution function for all species and relative abundances as in Table 2. The bulk speed assumed is 5×10^7 cm-s^{-1}, and the three proton temperatures are 10^4, 5×10^4, and 2×10^5 K.

Table 2.—Assumed abundances to generate Figure 3.

Species	M	Z	Relative Abundance
Hydrogen	1	1	10^{12}
Helium	3	2	5×10^{8}
	4	2	5×10^{10}
Carbon	12	6	1×10^{8}
Oxygen	16	6	5×10^{8}
	16	7	1×10^{8}

conventional mass spectrometer for which the speed of the incoming ions can be controlled at will and has a very small spread. One possibility consists of a collimator, to reduce the angular spread of incoming ions and direct them along the axis of the instrument, followed by an electrostatic analyzer, to select ions with a small range of energy per charge, and a Wien filter, to perform a velocity analysis. This instrument would have intrinsically poor time resolution due to the necessity of scanning in velocity at each energy per charge to compile a mass per charge spectrum. This could be overcome, at the expense of weight and bulk, by using a cycloidal mass spectrometer, which can be made to fulfill the necessary specifications and records ions at each mass per charge simultaneously.

IV. OBSERVATIONS

A. Normal Times

The solar wind is characterized by extreme variability. Great outbursts of plasma, emitted at the time of occurrence of major flares that produce shock waves, occur at a rate which varies with the solar activity cycle. There are cyclic changes, connected with the sector structure of the interplanetary medium (Wilcox, 1968), which also can be traced to changes on the solar surface (Schatten et al., 1969). There is a day to day variability which also must be the result of the average activity of the Sun but which is hard to connect with any particular solar activity parameter.

This variability is seen most easily by inspecting plots of the fluid parameters. The time scale of these plots is of crucial importance for recognizing the phenomenon to be studied. A scheme due to Burlaga (1969) which is appropriate because of its ability to organize data is shown in Figure 4. The kinetic scale, the finest time resolution, is required for the study of instabilities and shock structure, for example. The meso scale has been most useful for the study of discontinuities (Burlaga, 1968). The macro scale is the appropriate one for viewing the flow of material behind shocks on a length scale of a fraction of an AU. The frequency f given in the figure represents the time required for the solar wind to traverse the characteristic length L.

Direct observations have been made only in the ecliptic plane because of the high probe velocity ($\approx$30 km-s^{-1}) required to move an appreciable distance out of this plane in a time of the order of 1 year. Most measurements have been made between 0.7 and 1.3 AU, but this situation will be remedied before 1975 by probes to the vicinity of Jupiter and Mercury, and by Helios to 0.3 AU.

Figure 5 is a histogram of bulk speed, in 20 km-s^{-1} intervals, of observations made by Explorer 34 during 3000 hours in the interplanetary medium. These are

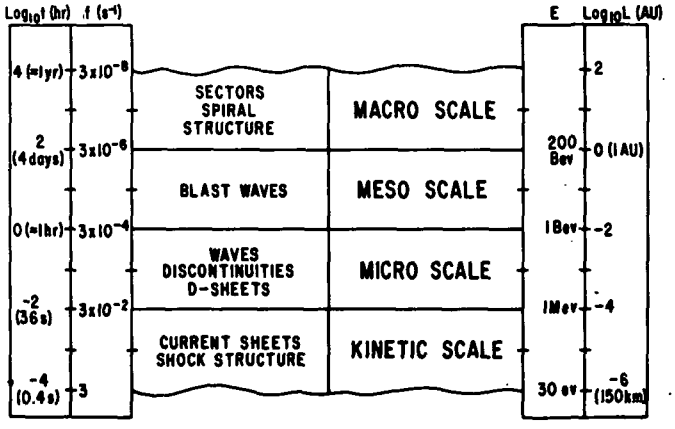


Figure 4.–Four scales are defined in terms of the time intervals on which data can be viewed. The time scale t is converted to a length scale L by multiplication of the time by a typical solar wind speed, 400 km-s^{-1}. Frequency f is the reciprocal of t; this is useful in discussions of power spectra. Also shown is the energy E of a proton for which the gyroradius in a 5γ field is equal to the scale length. Some structures which are seen on the various scales are indicated.

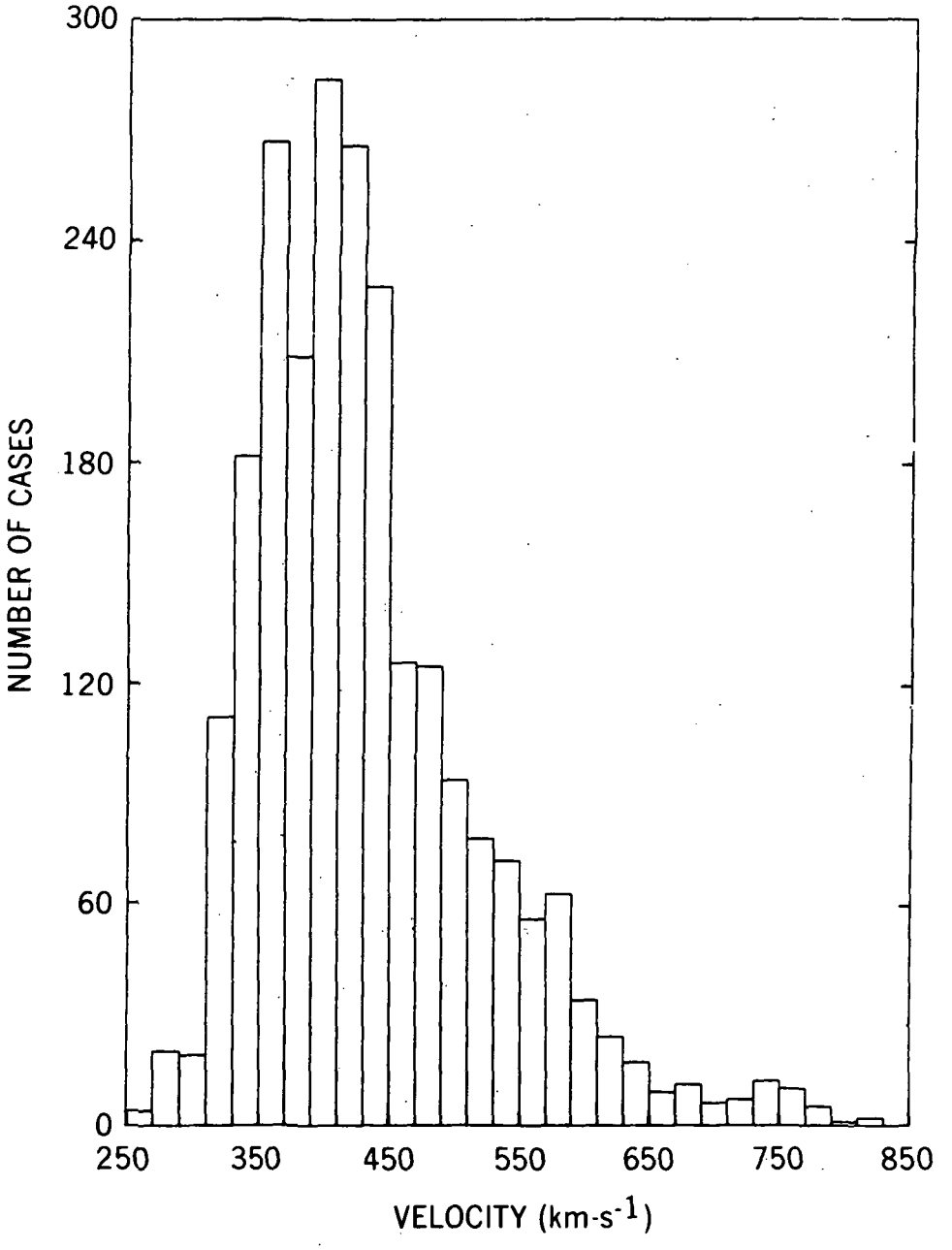


Figure 5.—Histogram of hourly average bulk speeds for Explorer 34 observations, ≈3000 hours in the interplanetary medium, between June 1967 and February 1968.

hourly average values, and the most probable bulk speed is about 385 km-s^{-1}. As pointed out by Hundhausen, there are few values below 300 km-s^{-1}, and about 90 percent of the time, the bulk speed falls between 300 and 600 km-s^{-1}. The bulk speed is closely the same for all species observed (H, He, and electrons), which indicates that although He probably is accelerated less strongly in the corona than H (Hundhausen, 1968), processes in the wind eventually bring the He ions up to the flow speed of the major constituent. This equality of flow speed seems to be true for "disturbed" and "quiet" conditions alike. Other bulk speed observations are summarized in Table 3.

The bulk speed is predicted to vary very little between a heliocentric radius of about 1/3 AU and a region in the outer solar system where appreciable quantities of kinetic energy are converted to thermal motion. There is, however, a hint of a variation of bulk speed with time over the solar activity cycle.

Explorer 34 hourly average values of density are shown as a histogram in Figure 6. The most probable value observed was 3.5 cm^{-3}; this quantity is not as well determined as the bulk speed, and its absolute value may be in error by 25 percent. The density varies with the inverse square of the heliocentric radius and is much more variable than the bulk speed, reaching values of 50 cm^{-3} in association with solar disturbances; there also appears to be an inverse correlation between bulk speed and density, which indicates a tendency for the flux to be constant. The lowest densities observed were less than 1 cm^{-3}.

The hourly average temperature observations for Explorer 34 are shown in Figure 7. They varied over the range of 1×10^4 K $\leqslant T \leqslant 3 \times 10^5$ K, with the

Table 3.–Some measured bulk speeds.

Vehicle	Bulk Speed Range (km-s^{-1})	Average	Year
Mariner 2	319 to 771	504	1962
IMP 1	200 to 675	360	1963
IMP 1	<300 to >700	378	1963
VELA 2	280 to 750		1964 to 1965
Explorer 34	270 to 840		1967 to 1968

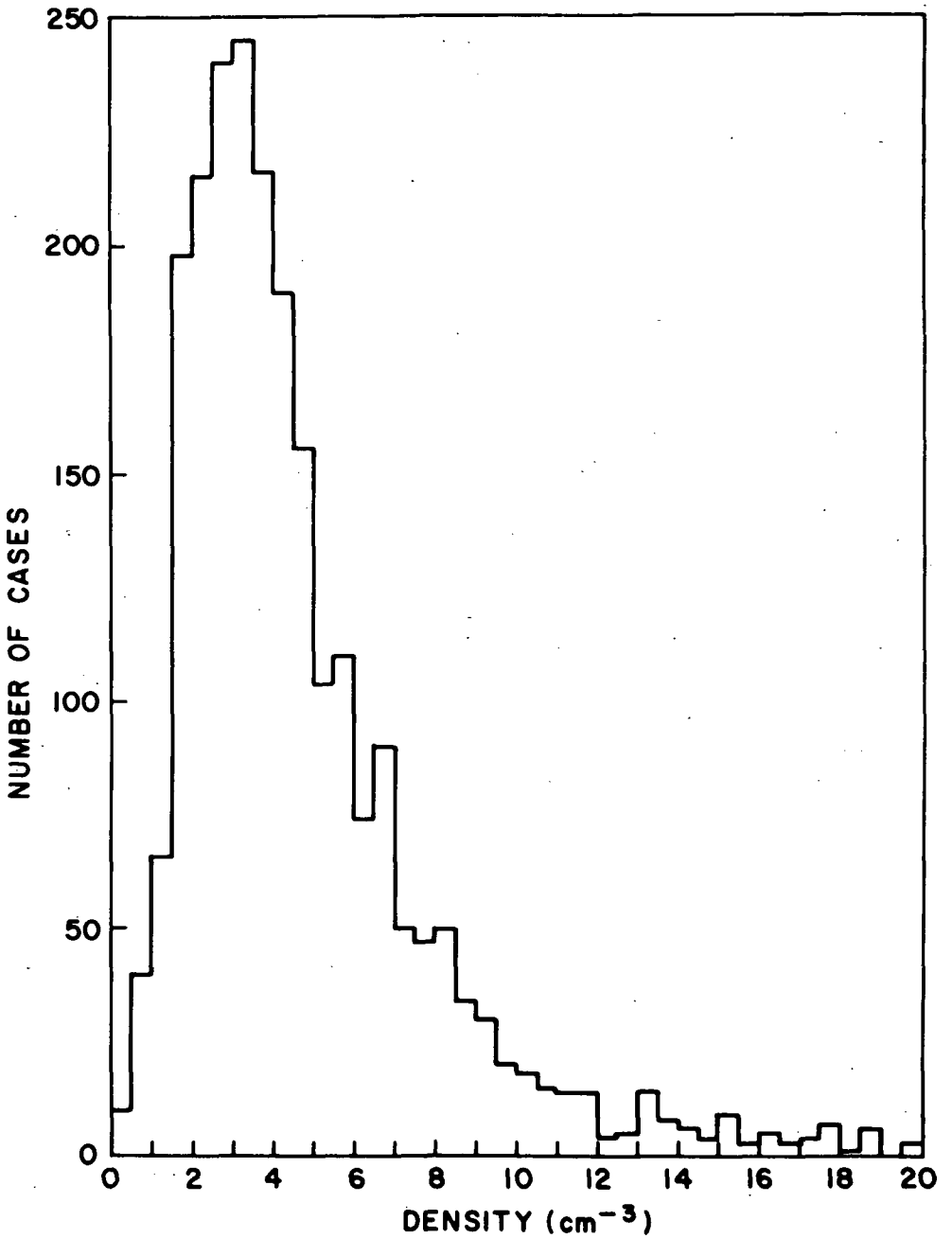


Figure 6.–Histogram of hourly average densities for Explorer 34.

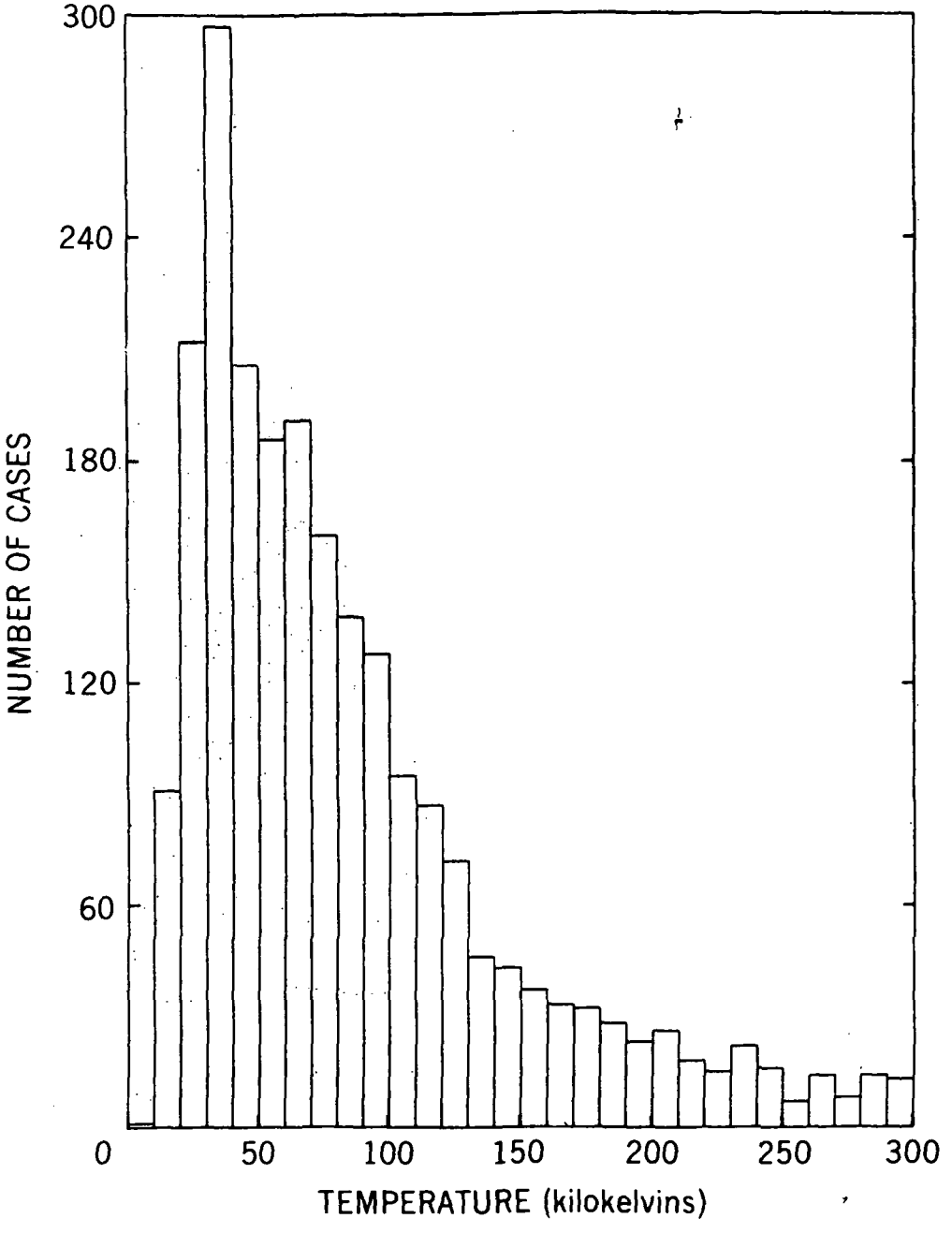


Figure 7.–Histogram of hourly average temperatures for Explorer 34.

greatest relative uncertainty present in the lowest values. The most probable value was 4×10^4 K. A fairly strong correlation observed on the macroscopic scale between temperature and bulk speed will be discussed later.

Figure 8 shows daily average values of bulk speed plotted as a function of time for June to December 1967 and projected in the left to form a histogram. This illustrates the tendency towards a minimum value of about 300 km-s^{-1} and shows the occurrence of high-speed streams separated by slower plasma. Geomagnetic sudden commencements, normally the signature of the passage past the Earth of shocks caused by solar flares, are indicated. As the solar wind flows out radially and $\frac{1}{2}mv_b^2 \gg kT \approx B^2/8\pi$, the high conductivity of the plasma causes the gross structure of the magnetic field to approximate an Archimedean spiral, so that, on the average, the magnetic field makes a small angle with the radial direction close to the Sun, and increases in angle with heliocentric distance. The sense of the field direction, toward or away from the Sun, divides up the medium into more or less well defined sectors (Wilcox, 1968), abruptly changing after a macroscopic time period to the opposite

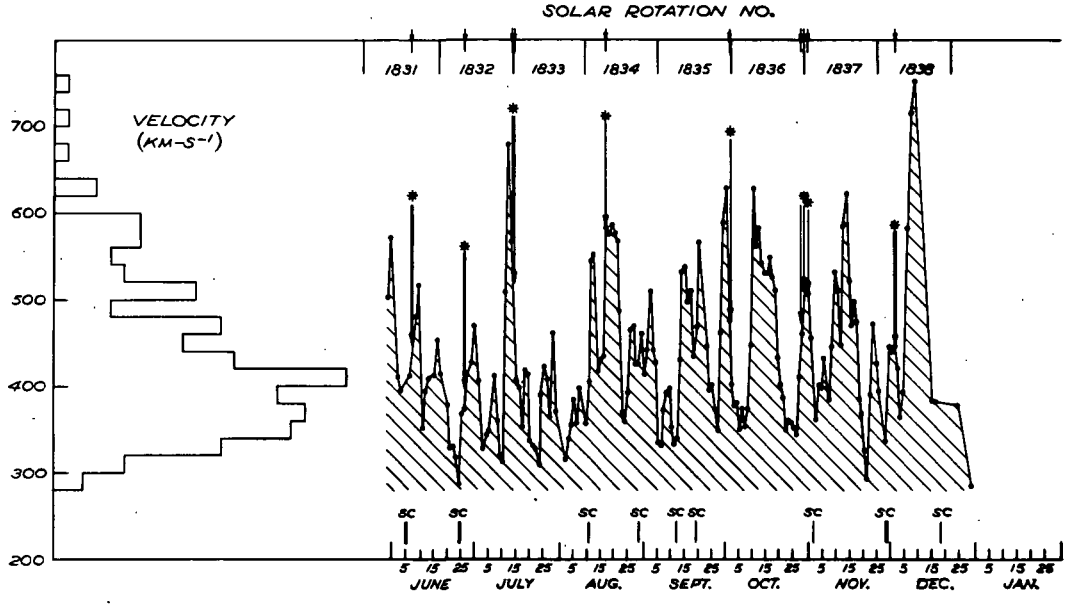


Figure 8.—Daily average values of bulk speed for Explorer 34, showing solar rotation numbers for June to December 1967. Note gaps due to the satellite being inside bow shock.

polarity. The pitch of the spiral formed by a given magnetic field line varies with the bulk speed, with the result that the plasma flux density tends to be greatest near the high speed streams and lower between them. If one neglects times when the state of the solar wind derives from an "event" on the Sun, the remainder of the observations show the normal variability of the solar wind.

Since $V_{th}/V_b \approx 0.1$, angular resolution of a few degrees is required for detailed study of the distribution function of the streaming plasma. When this is carried out (Hundhausen, 1968), a temperature anisotropy is observed. The thermal conductivity of the plasma is reduced in directions perpendicular to the magnetic field, leading to a lower temperature measured perpendicular to the field than along it. Thus, the temperatures illustrated in Figure 7 represent a mixed set, because the direction of the magnetic field was not taken into account. If we now apply the CGL (for Chew, Goldberger, and Lowe) approximation, instead of the hydromagnetic treatment used up to now, two quantities γ and δ will be conserved:

$$\gamma = kT_{\perp}/B$$

and

$$\delta = kT_{\parallel}B^2/n^2 ,$$

where n is density. Thus,

$$T_{\parallel}/T_{\perp} = \text{constant} \times n^2/B^3 .$$

Also $B = (B_r^2 + B_\phi^2)^{1/2}$, where B_r varies as $1/r^2$ and B_ϕ varies as $1/r^x$. Thus, $T_{\parallel}/T_{\perp} \sim r^{(3x-4)}$. There is a range of heliocentric radius r where $x > 1$, so that $T_{\parallel}/T_{\perp}$ should increase with heliocentric radius becoming of the order of 10 or larger at 1 AU. This will be true unless interactions, which cannot be Coulomb interactions, intervene to reduce it. Early measurements (Wolfe et al., 1966), which indicated large anisotropies, were in error because of experimental difficulties (Hundhausen, 1969), but the presently accepted average value for the ion anisotropy is about 2.0 (Hundhausen, 1968). This indicates the uncertainty to be associated with the temperatures in the approximation discussed above. The heat flow in the direction parallel to the magnetic field is not saturated, so this low value indicates wave-particle interaction in the plasma.

The electron component of the plasma was at first inferred from the condition of neutrality, but observations proved difficult because of instrumental problems associated with photoelectrons. These have now been overcome (Montgomery et al., 1968; Ogilvie and Lind, 1968), and the properties of the electrons measured. They show a remarkably constant electron temperature in the range 1×10^5 to 2×10^5 K and a temperature anisotropy which averages 1.2. Because of their high thermal speed, the electrons carry most of the heat in the plasma, and the constant electron temperature shows their conductivity to be sufficient to redistribute heat from the energy producing processes, such as colliding streams. Measurements of the heat flux by Montgomery et al. (1968) and by Ogilvie et al.* during normal times show that the conductivity of the plasma is less than that predicted by Spitzer-Harm formula. This is an expected result, since the Spitzer derivation assumes collisions, and in the present case, conduction is inhibited in the direction perpendicular to the magnetic field, and wave-particle interactions take the place of Coulomb collisions.

For a treatment of the theory of hydromagnetic waves in a multifluid collisionless plasma which is also anisotropic, the reader is referred to Burlaga.** Besides Alfvén waves, magnetoacoustic and fast waves should occur, and possibly also ion-acoustic waves. The existence and importance of various wave-particle coupling mechanisms have been discussed by the authors given in Table 4.

Because of the complexity of the phenomena and observational difficulties, the relative importance of various phenomena is not clear. The mechanisms dominant at 1 AU, for example, depend upon the state of the plasma there, and the value of

$$\beta = \frac{nkT}{B^2/8\pi}$$

is often approximately unity, whereas the theory of the onset of plasma instabilities is best developed for low β conditions. As an example, in a recent paper, Eviatar and Schulz (1970) conclude that conditions favoring the firehose instability occur relatively rarely but that this process sometimes acts to limit the ion temperature anisotropy near 1 AU. Both Coulomb collisions, despite their rarity, and growing electromagnetic waves play a significant role some of the time. Sorting out their effects and exploring the properties of this unusual region of parameter space in plasma physics is a major justification for the space program.

*K. W. Ogilvie, J. Scudder, and M. Sugiura, "Electron Energy Flux in the Solar Wind", submitted to *J. Geophys. Res.*, 1971.

**L. F. Burlaga, "Hydromagnetic Waves and Discontinuities in the Solar Wind", *Space Science Reviews*, to be published, 1971.

Table 4.—Instabilities discussed by various authors.

Mechanism	Authors
Low-frequency whistlers	Scarf, Wolfe, and Silva, 1967
Firehose instability	Kennel and Scarf, 1968
Cyclotron resonance instability	Kennel and Scarf, 1968

In the observational situation, a line is drawn arbitrarily at a frequency of 0.1 Hz. Above this frequency, special methods using search coils for electromagnetic waves and antennas for electrostatic waves are used, and below, conventional plasma detectors and magnetometers are capable of giving usable information. High-frequency wave measurements have been made on the eccentric orbit satellites OGO 1, 3, and 5, but, as pointed out by Scarf (1970), measurements by Fairfield (1969) show waves propagating upstream from the bow shock, and perturbations such as this may render the results unrepresentative of the undisturbed interplanetary medium. VLF electric field measurements were made on the space probes Pioneer 8, Pioneer 9, Zond 3, and Venus 2, but necessary high-rate telemetering was not available. The forthcoming Venus-Mercury flight provides an important opportunity to continue this study in a more comprehensive way. An interesting observation (Scarf et al., 1970) concerns the absence of detectable energy in the frequency range of 20 to 30 kHz away from the Earth, whereas close to the Earth, activity is common in this frequency range.

The most significant results obtained up to the present time at high frequencies have concerned the Earth's bow shock. An example of the observation of electric field noise associated with an interplanetary shock is shown in Figure 9, where $|\overline{\overline{F}} - \overline{\overline{F}}|$ measures the magnetic noise level. This does not increase after the shock on January 26, whereas a prolonged period of increased broadband electromagnetic noise follows the shock (Scarf et al., 1968).

Below 0.1 Hz, more experimental data exist, and the interpretation has progressed further, although there are still problems and disagreements. At the lowest frequencies on the microscopic scale ($\approx$1-hr period), the interplanetary medium, when observed close to the Earth as by Explorer 34, is often extremely quiet as shown by the magnetic field in part (b) of Figure 10 (Burlaga et al., 1969).

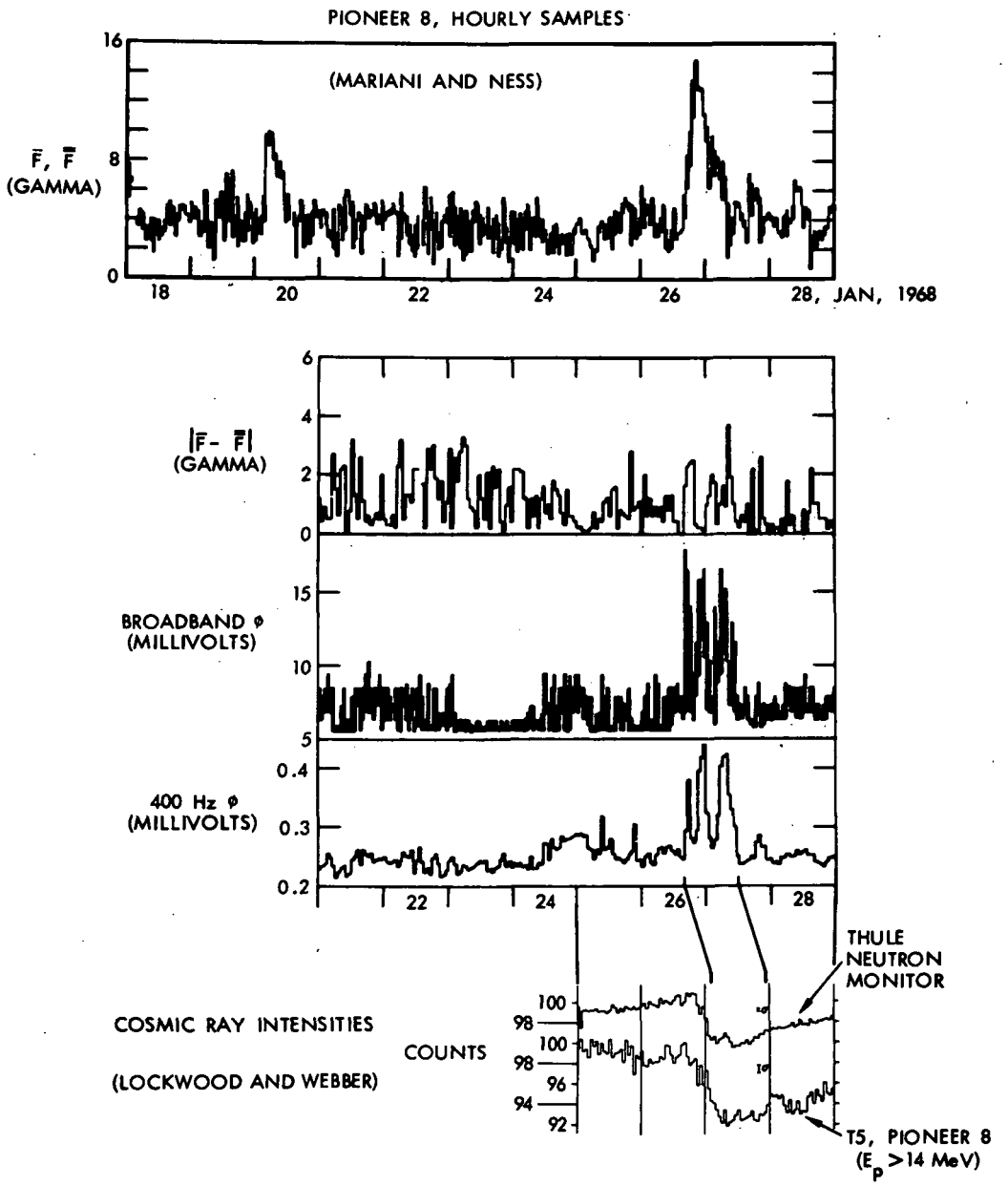


Figure 9.–Pioneer 8 magnetic field and wave observations for an interplanetary event that produced a clear Forbush decrease. $\bar{F}$ = average field magnitude, $\bar{\bar{F}}$ = magnitude derived from component averages.

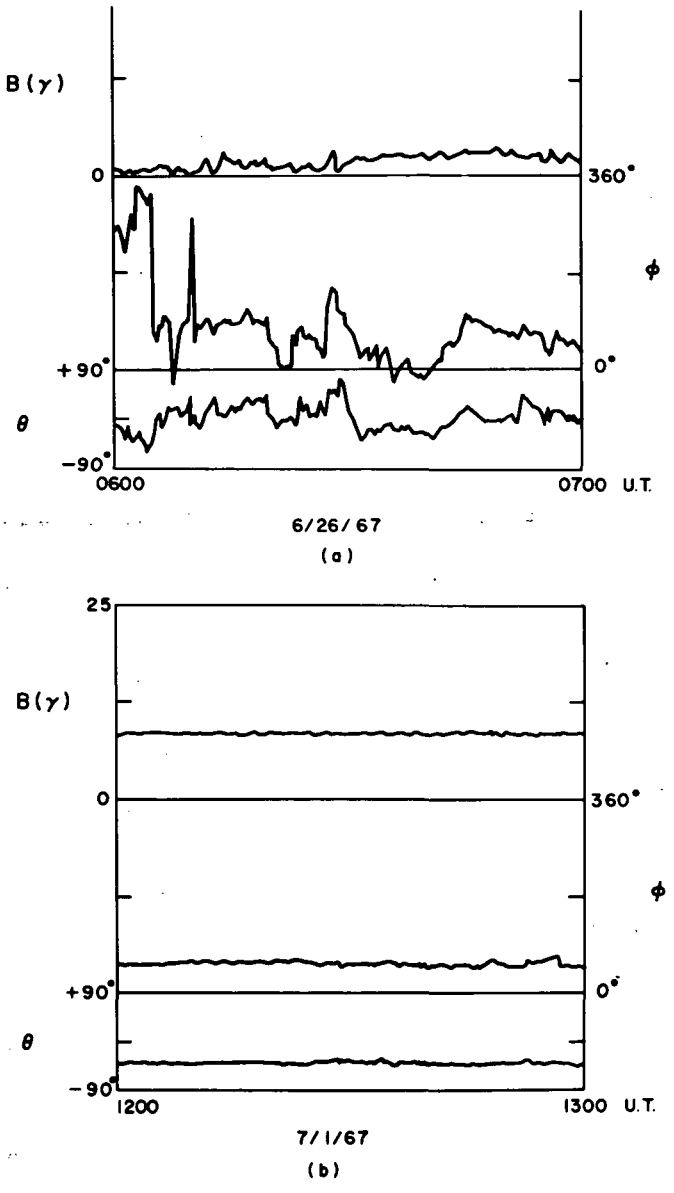


Figure 10.—(a) Example of an hour interval during which the field was judged to be very disturbed (plots show the magnetic field magnitude |**B**| and solar ecliptic latitude θ and longitude ϕ); (b) example of an hour interval during which the interplanetary magnetic field was judged to be very quiet.

At other times, it is often very disturbed, as shown in part (a) of the figure. The frequency of waves from the Earth's bow shock is 0.01 to 0.05 Hz, so there is no question of the contamination of these measurements by the presence of the Earth. In a stretch of 2500 hr of observation, it was striking how the individual hours could be divided as to quiet or disturbed, with very few disputes arising among the authors of this work over the classification! After the degree of disturbance of the magnetic field was correlated with the simultaneously measured value of β for the plasma, the following conclusions were reached:

(1) Very quiet intervals only occur if $\beta < 1$.

(2) When $\beta \geqslant 0.6$, very quiet conditions are exceedingly unlikely.

(3) Very disturbed conditions usually occur when $nkT > \frac{1}{2}(B^2/8\pi)$.

(4) Very disturbed conditions are very unlikely if $\beta \ll 1$.

Figure 11 shows relative probabilities of quiet and disturbed conditions as a function of β for the present definitions of quiet and disturbed. This behavior is consistent with several explanations: First, the fluctuations are caused by local plasma instabilities which are β dependent; second, the fluctuations are produced locally in some other way but are strongly damped, which tends to cause a high local value of β; and third, fluctuations propagating in the medium are strongly damped when $\beta \ll 1$.

The third explanation is eliminated because T is not abnormally high during quiet periods; the other two explanations are open. In particular, no instability was identified, so it can only be said that locally produced fluctuations form a consistent explanation of the observations.

Another low frequency, large scale effect is the observation of correlations between variations in the bulk speed V and the magnetic field. In an Alfvén mode, the density does not fluctuate, and the bulk speed and magnetic field perturbations $\mathbf{b}$ are perpendicular to the plane of $\mathbf{B}_0$ and $\mathbf{k}$. Therefore,

$$\mathbf{b} = (4\pi mn)^{1/2}\mathbf{v}\,.$$

Thus, the conditions to be associated with an Alfvén wave, or a disturbance made up of such waves, are constant density and a high degree of correlation, or anticorrelation, between the bulk speed and the radial component of B. If $\mathbf{b}$ and $\mathbf{v}$ were in phase for negative magnetic sectors, for example, propagation away from the Sun would be indicated. Figure 12 shows results of Belcher, Davis, and Smith (1969) which show correlated variations of B_r and V obtained from observations by Mariner 5. Conditions like this, when the coefficient of correlation was between 0.8

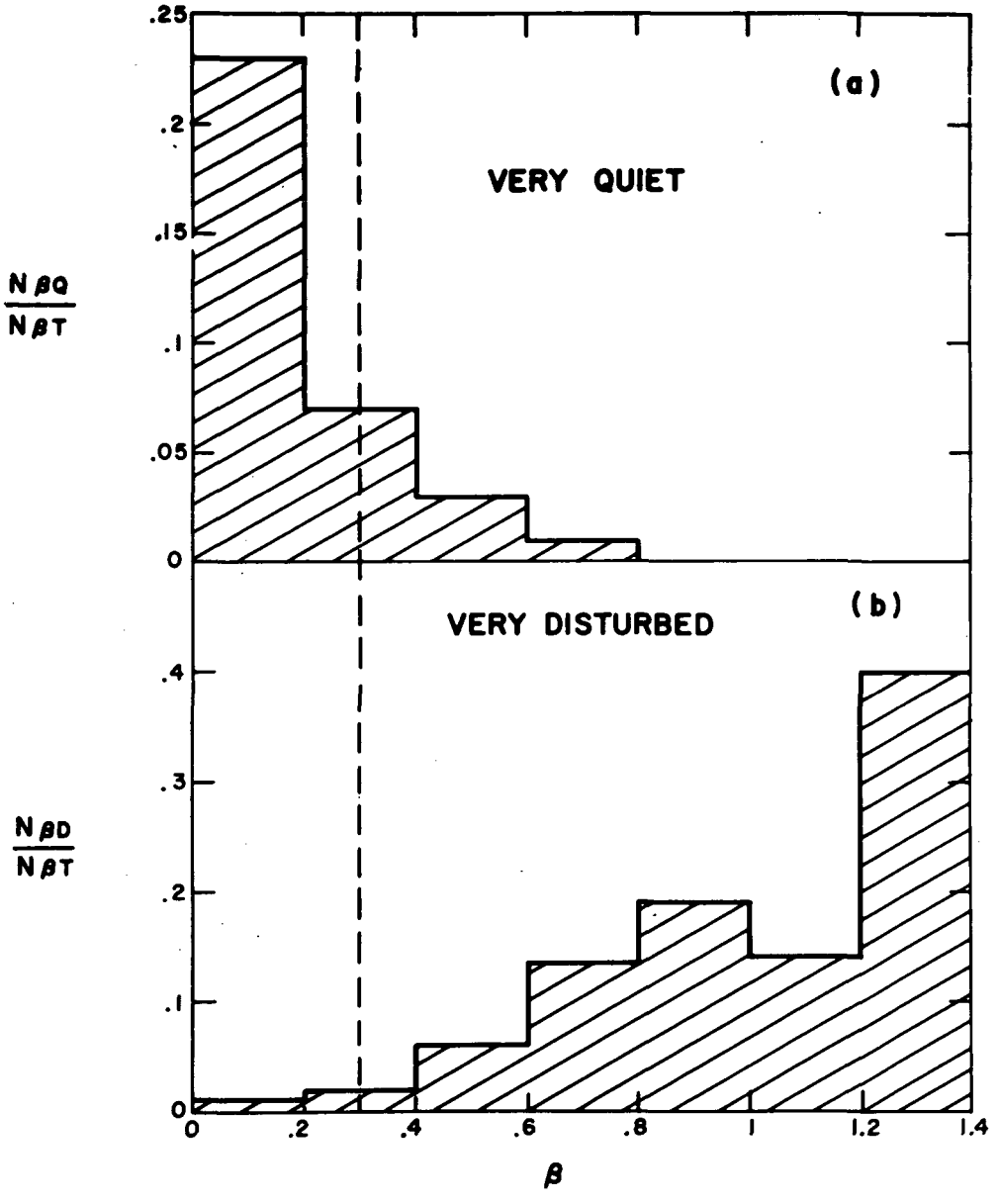


Figure 11.—(a) Computed relative probability that a quiet interval will be associated with a given β; (b) computed relative probability that a very disturbed interval will be associated with a given β.

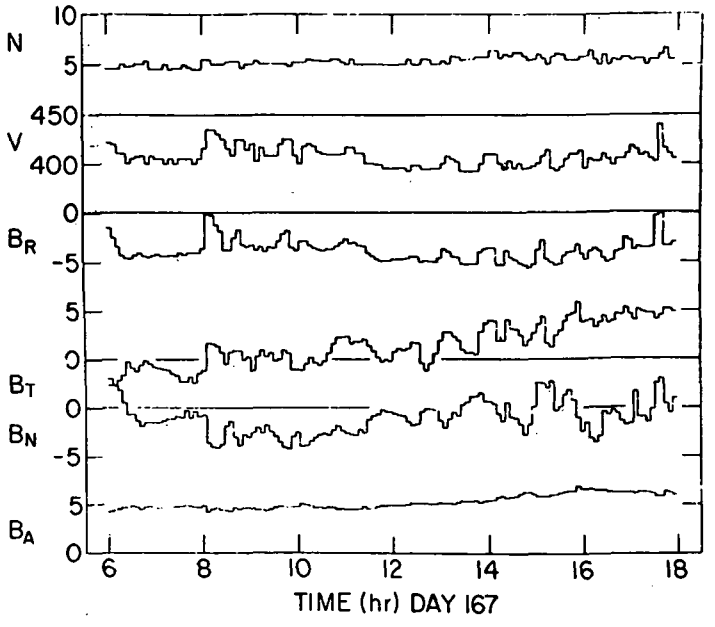


Figure 12.—Correlations between field and bulk speed observed by Belcher, Davis, and Smith (1969).

and 0.9, occurred some 20 percent of the time. They are very suggestive of aperiodic disturbances in the Alfvén mode, an interpretation which is strengthened by the observation that the sense of the correlation is always such as to suggest propagation outwards from the Sun. The sense of the correlation has been observed to change at a sector boundary to preserve this "direction of propagation". The highest correlation and most wavelike behavior are associated with high speed streams in the plasma. The presence of magnetoacoustic waves would lead to correlations between fluctuations in density and magnetic field strength which have not as yet been observed, perhaps because of the high degree of damping to which, in contrast to Alfvén waves, they should be subjected.

Examination of the magnetic field data indicates the presence of many discontinuous changes on the microscopic scale. Colburn and Sonett (1966) have discussed the kinds of discontinuities which can occur in a hydromagnetic fluid, and forward (Ogilvie and Burlaga, 1969) and reverse (Burlaga, 1970) shocks and tangential and rotational discontinuities (Burlaga, 1968) have already been observed in the interplanetary medium.

A tangential discontinuity (Figure 13) is one in which a rotation of the magnetic field vector occurs in a plane so that $(\mathbf{B} \cdot \mathbf{n}) = 0$. There can be an arbitrary speed change across the discontinuity, which is convected with the flow because the magnetic and particle pressures on each side are balanced. A rotational discontinuity (Figure 14), sometimes called an Alfvén shock, has a field with a nonzero normal component. At the discontinuity, the field undergoes a kink which propagates along the field line with the Alfvén speed:

$$[B] = 0 ,$$

$$V_n = \pm B_n/\sqrt{4\pi p} ,$$

$$[V_b] = [B_t]/\sqrt{4\pi p} ,$$

and

$$[p] = 0 .$$

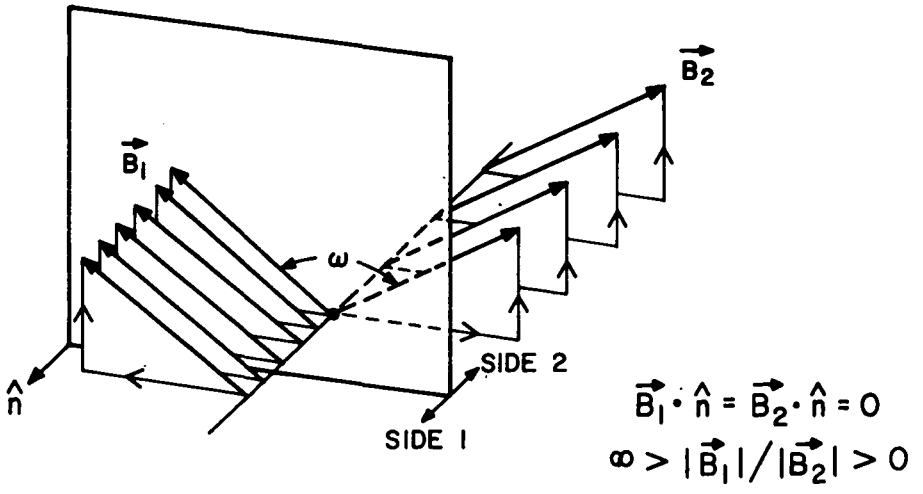


Figure 13.—A diagram of a tangential discontinuity showing relations between the field and normal vectors.

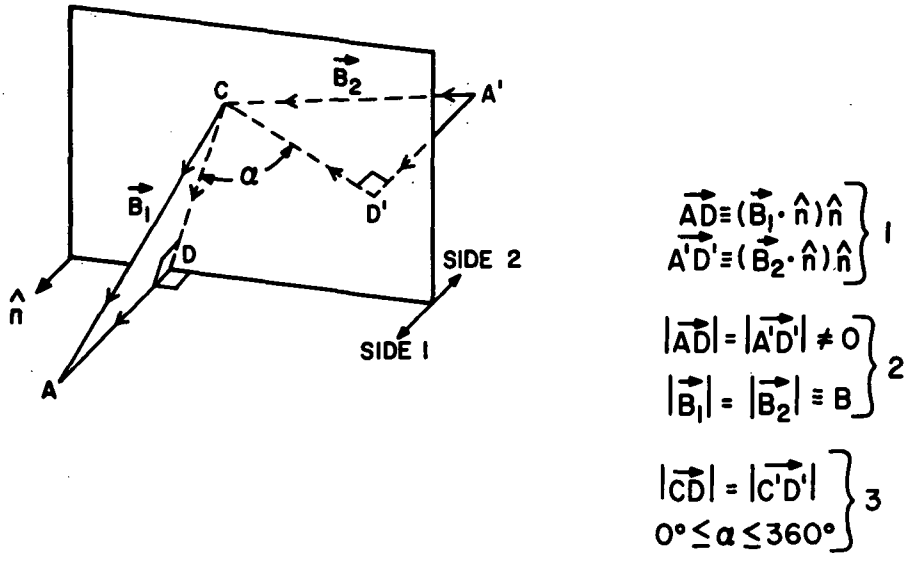


Figure 14.—A diagram of a rotational discontinuity showing relations between the field and normal vectors.

Some observed discontinuities have been proven to be tangential by using the condition $[\mathbf{B} \cdot \eta] = 0$ and showing that the pressure is balanced across them (Burlaga, 1968). Rotational discontinuities certainly must exist, but at the present time their existence has not been proved quantitatively because this requires their speed of propagation relative to the solar wind to be shown to be equal to the Alfvén speed, which is typically 50 to 100 km-s^{-1} in the interplanetary medium. The aperiodic wave trains observed by Belcher probably are not an ensemble of rotational discontinuities, but this interpretation cannot be ruled out entirely without examining the data on a finer scale. It seems almost certain that both Alfvén waves and rotational discontinuities exist and that their source is the Sun, but the details of their formation are not as yet understood.

A method of treating data which has been applied to magnetic field observations is to determine power spectral density as a function of frequency. This is the numerical equivalent of filtering, rectifying, and smoothing the time series

formed by the observations $X(t)$. (Blackman and Tukey, 1958). The power spectrum is given by

$$P(f) = \lim_{T\to\infty} \frac{1}{T}\left|\int_{-T/2}^{T/2} X(t)e^{-i2\pi ft}\,dt\right|^2$$

$$= \int_{-\infty}^{\infty} C(T)e^{-i2\pi fT}\,dT\,,$$

the Fourier transform of the autocovariance function, which is

$$C(T) = \lim_{T\to 0} \frac{1}{T}\int_{-T/2}^{T/2} X(t)X(t+T)\,dt\,.$$

An example of the variation of the power spectral density determined for magnetic field observations in the interplanetary medium is shown in Figure 15. Such plots show peaks at several periodicities and a characteristic slope which is approximated by an f^{-1} to f^{-2} dependence. The discrete periodicities, which must be treated with caution in short data samples, are between 2 and 120 hours. The slope seems to vary somewhat with frequency and also among observers. Siscoe et al. (1968) pointed out that the frequency dependence of the power spectral density of a discontinuity is proportional to f^{-2}, and, thus, the spectral density of an ensemble of discontinuities should also vary as f^{-2}. Sari and Ness (1969), using synthetic data composed of discontinuities alone, have shown that an f^{-2} frequency response, a power level similar to that sometimes observed, and even the occurrence of certain discrete periodicities could be reproduced.

These considerations favor the idea that the major contributor to the power spectrum at these frequencies may sometimes be discontinuities, but such arguments supply only necessary conditions. A spectrum of periodic Alfvénic disturbances could probably be constructed to fit the observed power density also, and it is likely that, in general, both causes play a part. The relative importance of waves and discontinuities in this frequency range is at present an interesting question.

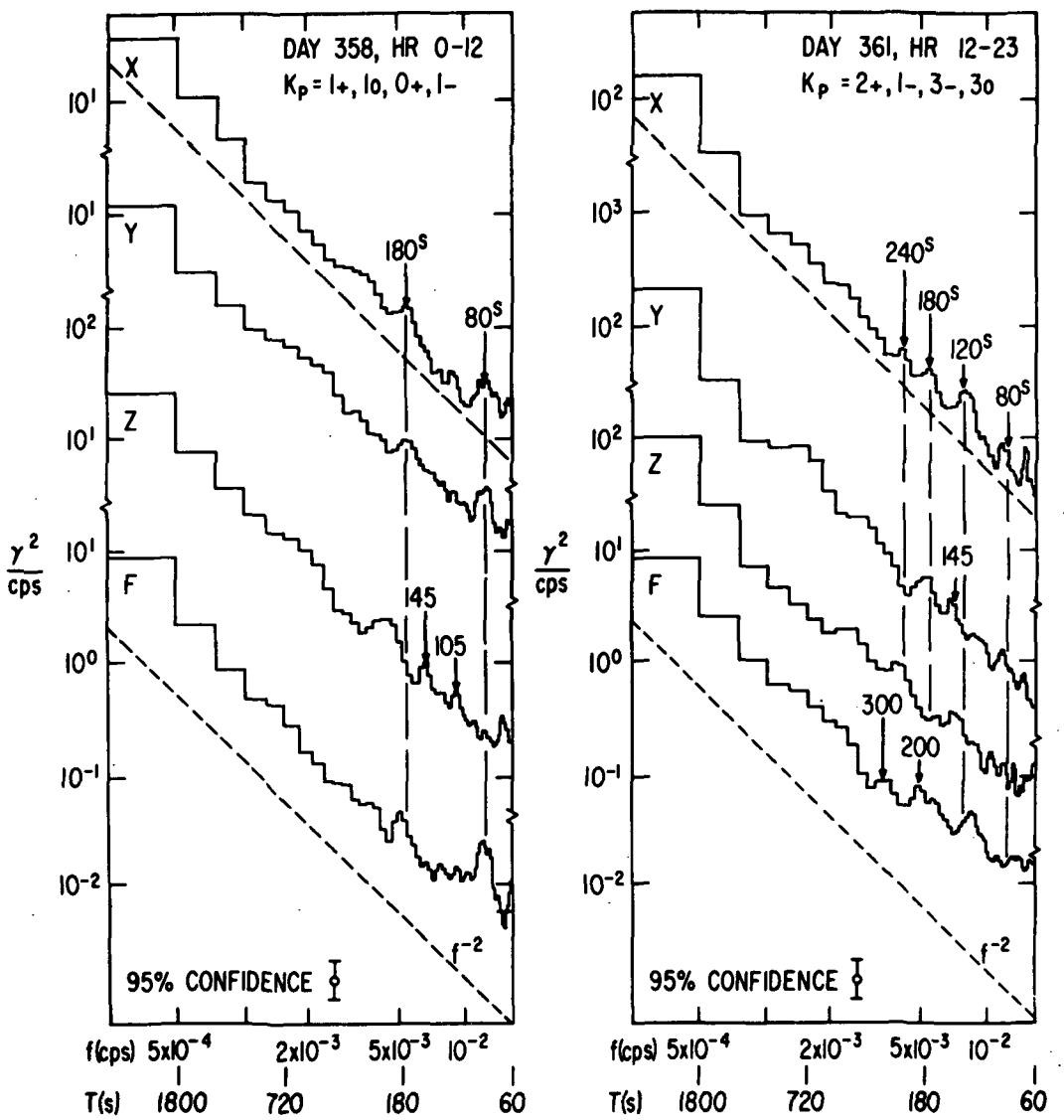


Figure 15.–Power spectra of the interplanetary magnetic field components and magnitude for two 12-hour periods in December 1965. Dotted lines indicate an inverse square frequency dependence. The K_p index for each period is included. At this time, the spacecraft was 2×10^6 km from the Earth at a Sun-Earth-probe angle of 90°.

We now consider correlations observed between the bulk speed and other quantities. In 1963, Snyder, Neugebauer, and Rao (Snyder et al., 1963) published details of a correlation between V_b and ΣK_p, the geomagnetic activity index. In the Explorer 34 data, this shows up as a tendency rather than as a strong correlation, as is illustrated in Figure 16. The upper part is a scatter plot of ΣK_p against V_b, and the lower a histogram of ΣK_p against V_b for all values of V_b and, in the shaded plot, for those V_b's for which 375 km-s^{-1} $\leqslant V_b \leqslant$ 425 km-s^{-1}. These histograms are very similar in shape except for very large values of K_p, which are usually paired with very high values of V_b.

A more fundamental correlation has been shown to exist between $T_p{}^{1/2}$ and V_b (Burlaga and Ogilvie, 1970), illustrated in Figure 17. Each point represents the average of the 3-hourly average values of $T_p{}^{1/2}$ for which similar average values of V_b fell in the 50 km-s^{-1} interval shown. The open circles represent Explorer 34 observations, and the error bar represents the variability of the data. The line corresponds to the relation

$$T_p^{1/2} = (0.036 \pm 0.003)V_b - (5.6 \pm 1.5)\,,$$

with T_p in units of 10^3 K and V_b in km-s^{-1}. The data of other observers at different parts of the solar cycle also fit this relationship quite well. Thus, it appears to describe a general and fundamental characteristic of the solar wind. Note that the spread in values of T_p and the variability of field direction during a 3-hour period make it unnecessary to take the anisotropy into account in discussing this macroscopic relation. The upper line corresponds to Parker's (1963) isothermal calculation, which of course predicts temperatures too high for a given bulk speed. The points V, W, and N' correspond to published theoretical predictions using one-fluid theory (Whang et al., 1965; Whang and Chang, 1965; Noble and Scarf, 1963), and the points X correspond to predictions of the two-fluid theory of Hartle and Sturrock (1968), the lower of these giving an unrealistically low flow speed. At extreme upper values of V_b, which are usually associated with solar "events", the largest discrepancies between the relation above and the observations might be expected. Energy in the form of heat added to the expanding solar wind at a heliocentric distance of a few solar radii increases its bulk speed and farther away ($> 50R_\odot$) increases the plasma temperature. Postulating a source of heat at, for example, $20R_\odot$, Hartle and Barnes (1970) have been able to reproduce the

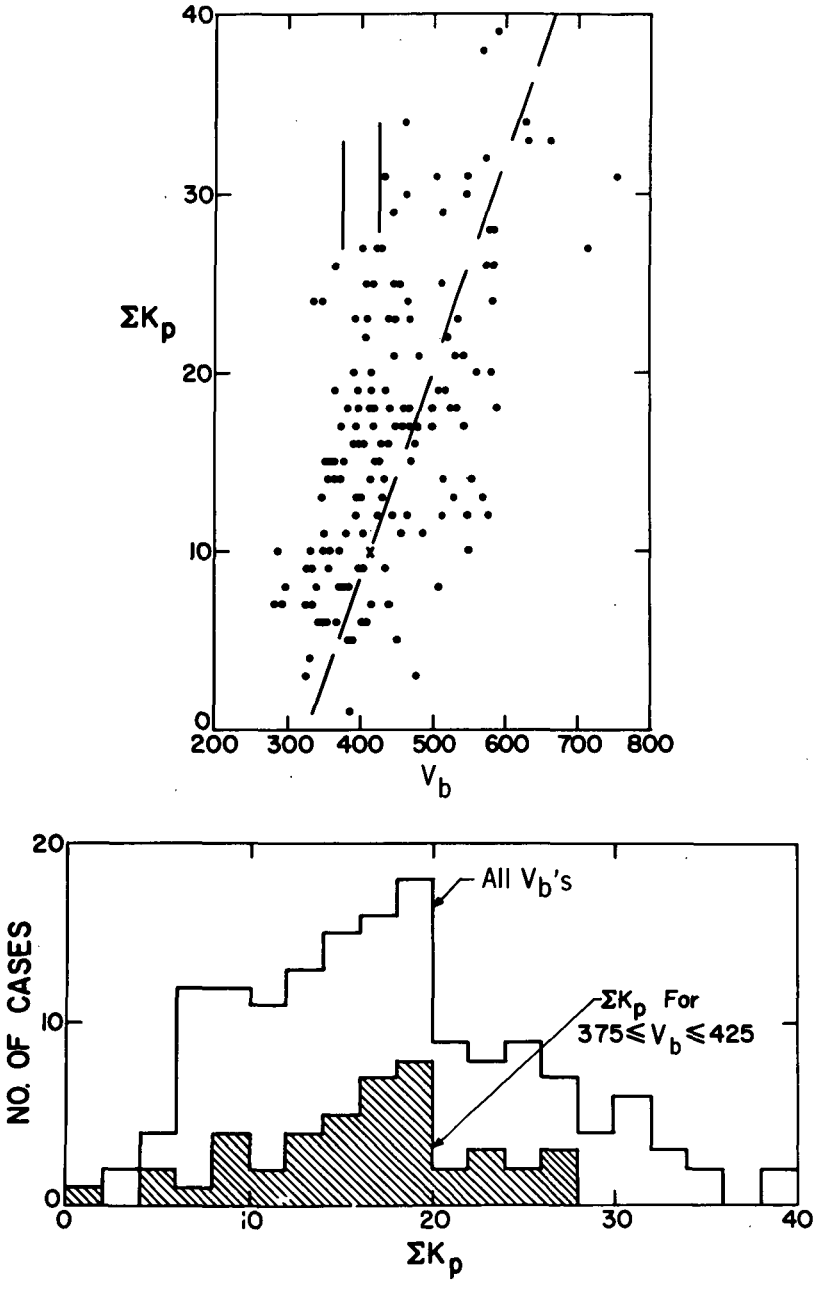


Figure 16.–Illustration of the tendency of ΣK_p to correlate with daily average bulk speed for Explorer 34.

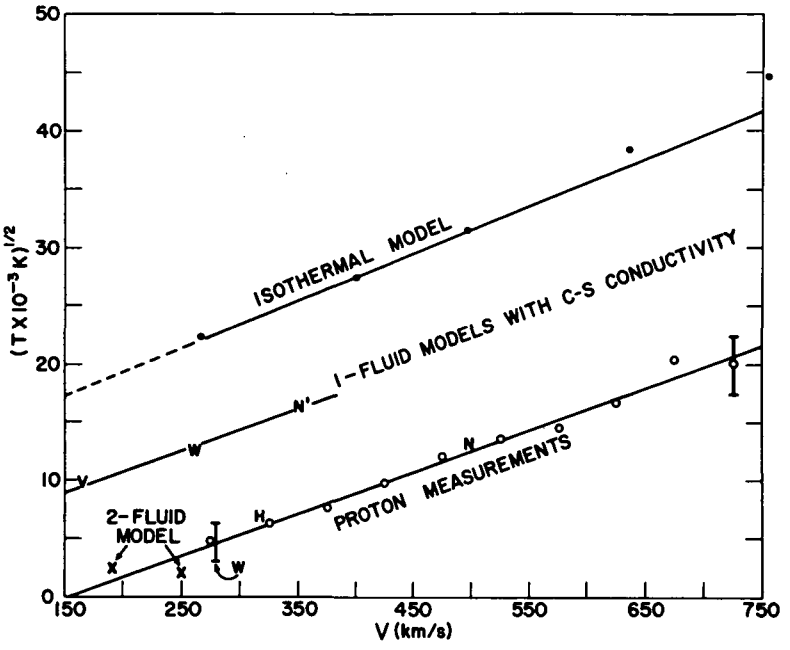


Figure 17.—Values of $T^{1/2}$ computed from 3-hour averages, plotted as a function of bulk speed V for intervals 250 km-s^{-1} $< V <$ 300 km-s^{-1}, etc., together with theoretical predictions and other observations as discussed in the text. Open circles—Explorer 34 observations, with variability indicated by error bar on the uppermost point; solid circles—Parker's solution to the Bernoulli equation for an isothermal corona, $T \sim V^2$. The line through the proton measurements is a least-squares fit to the data, and the other lines are arbitrarily drawn parallel to this line.

postulated $T^{1/2} - V_b$ relation up to 400 km-s^{-1}. Note that the upper two-fluid theory point at 250 km-s^{-1} lies below the line through the observations. Although agreement can be obtained in this way, as pointed out by Brandt and Wolff,* important processes are not taken into account, and nonthermal processes may not, in the end, be required. The accompanying chapter by Hundhausen** should be consulted for a discussion of the controversial theoretical ideas associated with heating and the applicability of theories of the solar wind.

*J. C. Brandt and C. L. Wolff, "On Solar Wind Heating", preprint, 1971.
**See Chapter 4, "Dynamics of the Outer Solar Atmosphere".

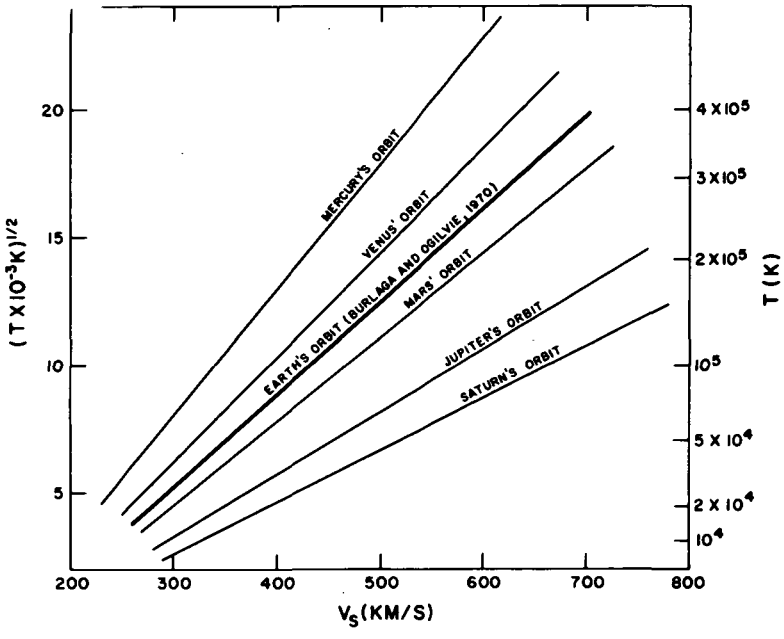


Figure 18.—*V-T* relations at heliocentric radii other than 1 AU.

An important use for this relation is in the identification of unusually hot regions in the plasma, to assist in the identification of instabilities and heating mechanisms (Burlaga et al., 1971). Figure 18 shows a prediction by Whang* of the form of the $T^{1/2}$-V_b relation in the vicinity of other planets in the solar system.

B. Disturbed Times

This section deals with observations taken at times following solar flares and interplanetary storms. Until recently, these have been dealt with on an event-by-event basis, in other words, morphologically. In recent times, however, the more detailed observations available have brought out regularities and common features.

A major solar flare releases 10^{31} to 10^{32} ergs of energy in a time of the order of 10^3 s. The result of this is to propel a shock wave into the interplanetary

*From a private communication.

medium, and about 30 of these have been observed as they passed spacecraft (Hundhausen, 1970). When such shocks encounter the Earth, the effect is usually to produce a sudden commencement magnetic storm (Burlaga and Ogilvie, 1969). Parker originally suggested that these shocks would be blast waves, that is to say, that the shocks would not be driven as far as 1 AU. If this were the case, both bulk speed and density would drop rapidly behind the shock. In the case of a driven disturbance, the shock is followed by interplanetary material which has passed through the shock front, and this material may be separated from the driver gas by one or more tangential discontinuities.

Hundhausen and Gentry (1969) have solved the time dependent hydrodynamic problem of the propagation of a shock between heliocentric shells of radius r_1 and r_2. This includes the pressure gradient and motion of the solar wind and assumes adiabatic flow, so that it is reasonably realistic except for the assumption of radial symmetry. The results show that to much better than order-of-magnitude accuracy, transit times of $\approx$50 hr, flare released energies of 10^{31} to 10^{32} ergs, and disturbance durations of the order of minutes are mutually consistent.

Observationally, the events show a wide variety, especially in the nature of the postshock flow. A classification based on 19 events, proposed by Hundhausen et al. (1970), divides them into *F* events for which the energy flux falls off behind the shock and *R* events for which it rises behind the shock. The total mass of material and its energy were determined for each event by integrating the excess rise above the preshock value; the average values were 3×10^{16} g and 5×10^{31} ergs, respectively. The assignment of flares to these shocks suffers from the usual uncertainty, but there is little evidence that they were associated with high speed streams, and the average energy is consistent with flare association at a time when very large flares were rare. The *F* class of events resembles the blast wave picture, except for the rise in flow speed observed after the shock, also noted earlier by Ogilvie and Burlaga (1969). A remarkable proportionality between the energy released and the total mass excess was obeyed by all events. If this is interpreted directly, it indicates an average energy release per proton of about 3 keV by the flare. The *R*, or driven, events tend to be the larger ones, so that a large energy release seems to occur over a longer time.

Examination of the presently available observations of interplanetary shock waves shows the average transit time is about 55 hr, the mean speed based upon that time is 750 km-s^{-1}, and the average of the speeds at the point of observation is 500 km-s^{-1}. This indicates that a typical interplanetary shock is not strong and that if the duration of the original disturbance is sufficient, it is driven. These values are

entirely consistent with typical flare energy releases. "Corotating" shocks at the leading edge of high speed streams have not been identified unambiguously and may not exist at 1 AU, although reverse shocks which propagate towards the Sun while being convected outwards with the plasma flow have been recently identified (Burlaga, 1970).*

This picture leads to two very interesting considerations. The first is the deviation of the shock front from spherical symmetry. Using arguments concerning the correlation between the magnitude of geomagnetic sudden commencements on the Earth and the heliographic coordinates of assigned flares, Hirschberg (1968) postulated that the radius of curvature of the corresponding average interplanetary shock is near 0.5 AU when it encounters the Earth. Such a radius of curvature is also in agreement with the results of Taylor (1969), who computed the normal directions to 36 likely shocks observed in satellite magnetometer data. Luckily, this radius is sufficiently large not to seriously invalidate the calculations of Hundhausen and Gentry. The Earth can probably encounter such a shock without entering the driven material, if the latter is "standing off" from a volume of rapidly moving ejecta. This feature may help to explain qualitatively some of the observed variability.

Second, an interesting observation is the association of abnormally large values of the He/H ratio immediately after some interplanetary shocks in the position where the driving material might be supposed to be. This is now a well established phenomenon, having been observed on at least five occasions by three different groups of investigators (see Hirschberg et al., 1970).

The presence of $^4He^{++}$ in the solar wind was postulated by Neugebauer and Snyder (1966) and confirmed by Ogilvie and Wilkerson (1969) using a method which performs a mass-per-unit-charge separation. Extensive studies have also been made by the VELA group (Robbins et al., 1970) and by the MIT group (Lazarus et al., 1970). The results of these studies agree in predicting an average proportion of about 4.5 percent by number, which may vary slightly over the solar activity cycle. The bulk speeds of the two species are equal to within experimental error, and their temperatures are in the ratio of 4:1, with a rather large spread, which indicates that these ionic species have a common velocity distribution rather than being in thermodynamic equilibrium at 1 AU, as they presumably were at the corona. This again indicates the presence of a noncollisional energy transfer mechanism, in this case between the protons and the helium ions. The observed range of variation of hourly average values of the He/H number ratio, derived from Explorer 34

*J. Binsack, private communication.

measurements (Ogilvie and Wilkerson, 1969), is shown in Figure 19. Determination of the He/H ratio characteristic of the Sun by spectroscopic methods gives 0.063 ± 0.015 (Lambert, 1967), by the observation of solar cosmic rays gives 0.055 to 0.07 (Durgaprasad et al., 1968), and by the mass-luminosity relation for 12 stars gives 0.077 ± 0.01 (Morton, 1968). There is therefore little doubt that the normal He content of the interplanetary medium is less than that of the Sun and that it rises above that of the Sun after some shock waves. Separation of ions by mass in the corona as a result of settling and diffusion (Geiss et al., 1970) would tend to produce in the corona a decreasing proportion of He as the heliocentric radius of the position of observation increases. Enhancement to values above those characteristic of the chromosphere is expected in the lower corona. A flare outburst presumably ejects material from the lowest regions, which suggests that the observed helium enhancements are associated with material from low levels in the corona. Hirschberg has suggested that flares may occur at places in the chromosphere where the

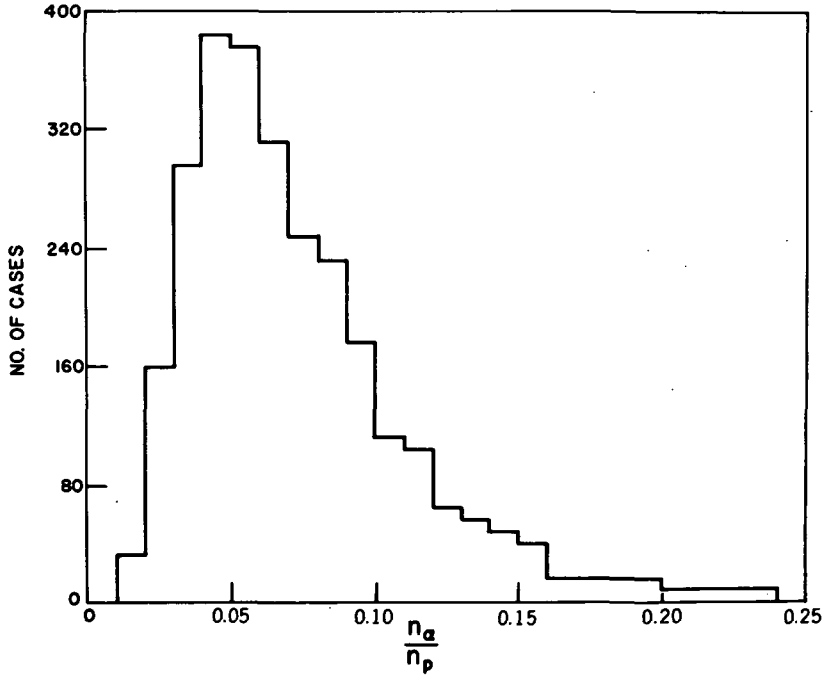


Figure 19.–Histogram of determinations of n_α/n_p for June 1967 to February 1968 by the Explorer 34 plasma experiment.

proportion of helium has fluctuated to a high value, and Robbins, that helium might be produced by thermonuclear reactions at the time of flares. The maximum observed values of the He/H ratio are well above the expected chromospheric values.

The helium enhancement is sampled several hours after observation of the shock. Material swept up from the ambient medium is found immediately behind the shock, separated from the helium-rich driver by one or more tangential discontinuities. If the path of the space probe through the disturbance does not intersect the driver material, but merely encounters the shock, then no helium enhancement is to be expected.

During 1967, fortunate circumstances allowed Lazarus et al. (1970) to observe the flow regions behind an interplanetary shock at two points separated by 0.1 AU, which showed that regarding observations as a function of time at 1 AU as being the equivalent of observations of the passage of a unique structure past the detector is a legitimate transformation. An intercalibration of the two detectors, one on Mariner 5 and the other on Explorer 34, was carried out at the time of the shocks seen by the latter detector at 0215 UT on June 25 and 1455 UT on June 26, 1967. Hourly averages of density, bulk speed, and most probable thermal speed are shown in Figure 20. During the quiet periods immediately before the shock on June 25 and around 1200 hours on June 26, agreement between both sets of observations was very good; disagreement between the density observations between 1800 on June 25 and 0600 on June 26 is regarded as being due to density fluctuations on a scale of 10^6 km. By August 11, 1967, the spacecraft were separated by 1.6×10^7 km, and both were situated close to the Sun-Earth line. In Figure 21, hourly averages of the fluid parameters derived from the Mariner 5 instrument, adjusted for the expected inverse-square density dependence, are plotted at times corresponding to radial convection at the bulk speed. These are compared with similar parameters observed near the Earth by the Explorer 34 instrument. The agreement of all three quantities is very good, except for the period 0600 to 1800 on August 11. The actual shock identification is ambiguous because of a telemetry gap in the case of Mariner 5 and the probable deflection of the plasma flow away from the sensitive aperture of the Explorer 34 instrument, which indicated an apparent density decrease at 0555 UT. Nonetheless, the very good agreement between the observations indicates that both spacecraft were observing the same material, which was moving without distortion over a length scale of 0.1 AU. It proved difficult in this case to be certain whether this disturbance should be ascribed to a flare or a corotating structure, but the evidence is consistent with the former explanation.

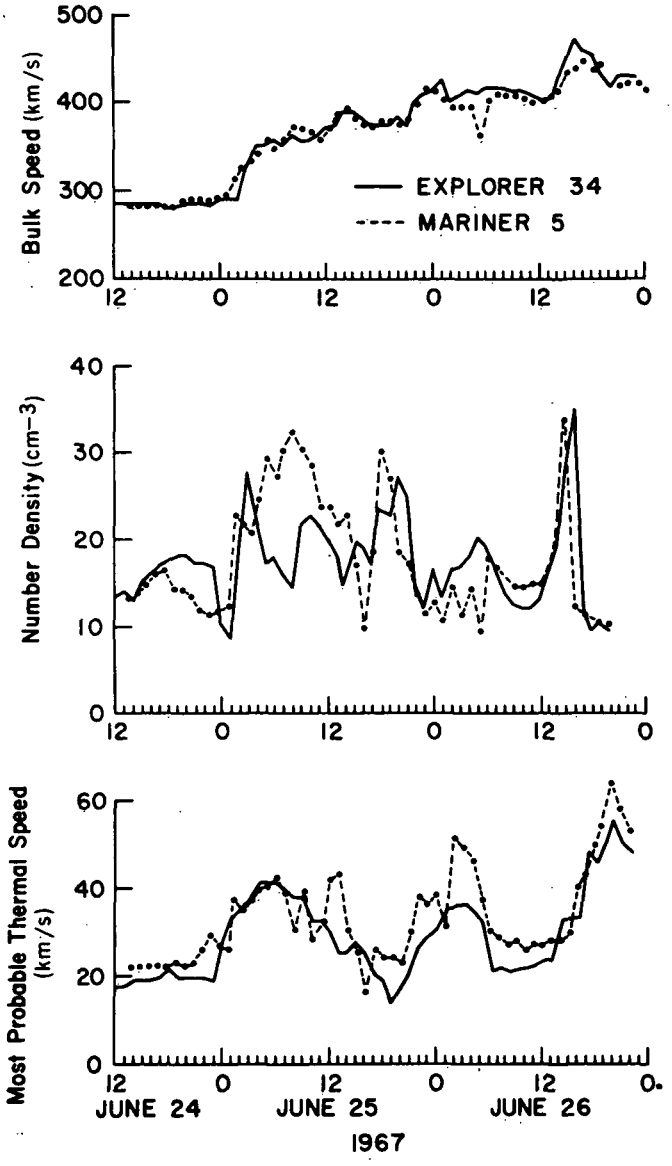


Figure 20.–Hourly averages of bulk speed, number density, and most probable thermal speed derived from the Explorer 34 and Mariner 5 plasma experiments for the June 25 and 26 events. Note that time duration of the density peak starting at 1500 UT on June 26 is longer for Explorer 34.

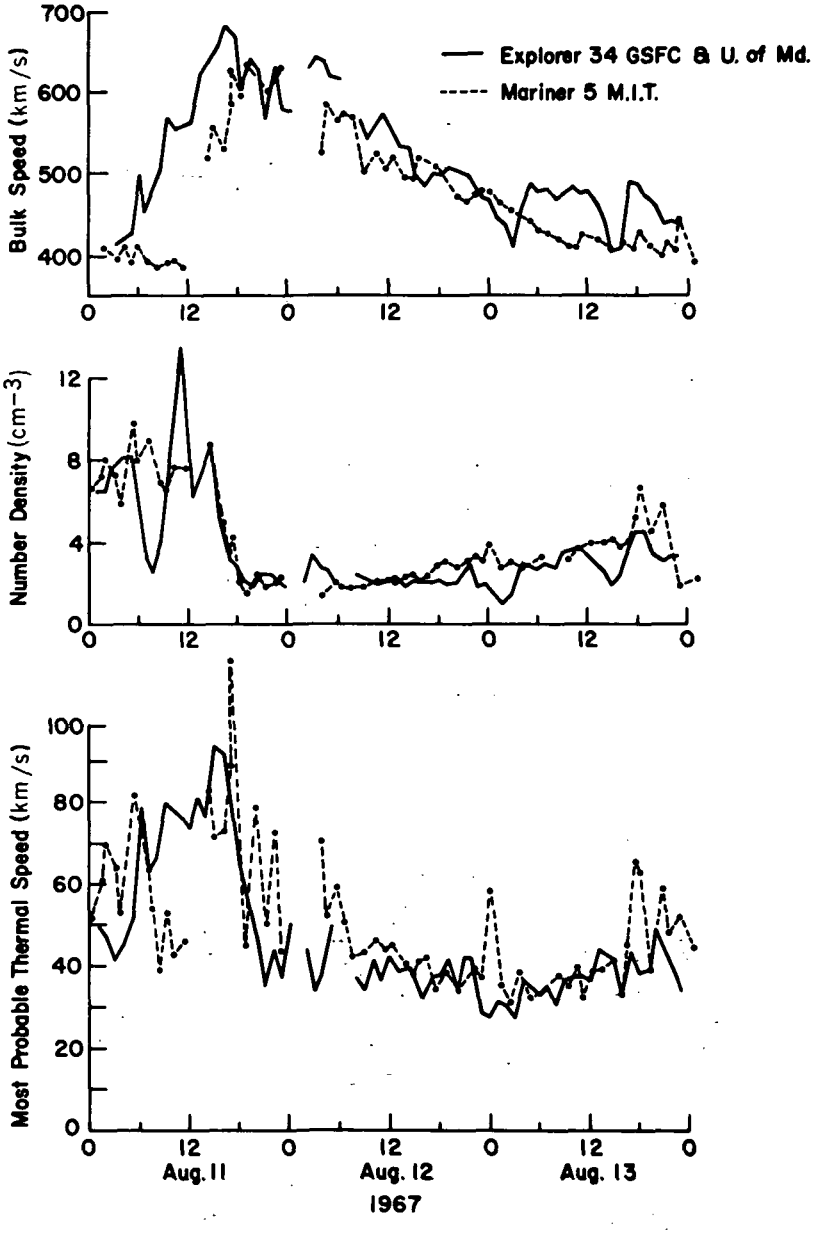


Figure 21.–Hourly averages of bulk speed, number density, and most probable speed derived from the Explorer 34 and Mariner 5 plasma experiments for the August 11 event. Explanation of this diagram is given in the text.

REFERENCES

Bame, S. J., Hundhausen, A. J., Asbridge, J. R., and Strong, I. B., "Solar Wind Ion Composition", *Phys. Rev. Lett.* 20:393, 1968.

Belcher, J. W., Davis, Leverett, Jr., and Smith, E. J., "Large Amplitude Alfven Waves in the Interplanetary Medium", *J. Geophys. Res.* 74:2302, 1969.

Blackman, R. B., and Tukey, J. W., *The Measurement of Power Spectra*, Dover, New York, 1958.

Burlaga, L. F., "Micro-Scale Structures in the Interplanetary Medium", *Solar Phys.* 4:67, 1968.

Burlaga, L. F., "Directional Discontinuities in the Interplanetary Magnetic Field", *Solar Phys.* 7:54, 1969.

Burlaga, L. F., Ogilvie, K. W., and Fairfield, D. H., "Microscale Fluctuations in the Interplanetary Magnetic Field", *Astrophys. J. (Letters)* 155:L171, 1969.

Burlaga, L. F., and Ogilvie, K. W., "The Causes of Sudden Commencements and Sudden Impulses", *J. Geophys. Res.* 74:2815, 1969.

Burlaga, L. F., "A Reverse Hydromagnetic Shock in the Solar Wind", *Cosmic Electrodynamics* 1:233, 1970.

Burlaga, L. F., and Ogilvie, K. W., "The Heating of the Solar Wind", *Astrophys. J.* 159:659, 1970.

Burlaga, L. F., Ogilvie, K. W., Fairfield, D. H., Montgomery, M. D., and Bame, S. J., "Energy Transfer at Colliding Streams in the Solar Wind", *Astrophys. J.* 164:137, 1971.

Colburn, D. S., and Sonett, C. P., "Discontinuities in the Solar Wind", *Space Sci. Rev.* 5:439, 1966.

D'Angelo, W., "Cesium Plasma Research", *Advances in Plasma Physics*, Interscience Publishers, New York, 1969, Vol. 2.

Durgaprasad, N., Fichtel, C. E., Guss, D. E., and Reames, D. V., "Nuclear Charge Spectra and Energy Spectra in the Sept. 2, 1966 Solar Particle Event", *Astrophys. J.* 154:307, 1968.

Eviatar, A., and Schulz, M., "Ion Temperature Anisotropies and the Structure of the Solar Wind", *Planet. Space Sci.* 18:321, 1970.

Fairfield, D. H., "Bow Shock Associated Waves in the Far Upstream Interplanetary Medium", *J. Geophys. Res.* 74:3541, 1969.

Geiss, J., et al., "On Acceleration and Motion of Ions in Corona and Solar Wind", *Solar Phys.* (in press), 1970.

Hartle, R. E., and Sturrock, P. A., "Two Fluid Model of the Solar Wind", *Astrophys. J.* 151:1155, 1968.

Hartle, R. E., and Barnes, A., "Non-Thermal Heating in the Two Fluid Solar Wind Model", *J. Geophys. Res.* 75:6915, 1970.

Hirschberg, J., "The Transport of Plasma from the Sun to the Earth", *Planet. Space Sci.* 16:309, 1968.

Hirschberg, J., Alksne, A., Colburn, D. S., Bame, S. J., and Hundhausen, A. J., "Observation of a Solar Flare Induced Interplanetary Shock", *J. Geophys. Res.* 75:1, 1970.

Hundhausen, A. J., "Direct Observations of Solar Wind Particles", *Space Sci. Rev.* 8:690, 1968.

Hundhausen, A. J., "Interpretation of Positive Ion Measurements", *J. Geophys. Res.* 74:3740, 1969.

Hundhausen, A. J., and Gentry, R. E., "Numerical Simulation of Flare Generated Disturbances in the Solar Wind", *J. Geophys. Res.* 74:2908, 1969.

Hundhausen, A. J., "Composition and Dynamics of the Solar Wind Plasma", *Rev. Geophys.* 8:729, 1970.

Hundhausen, A. J., Bame, S. J., and Montgomery, M. D., "Large-Scale Characteristics of Flare-Associated Solar Wind Disturbances", *J. Geophys. Res.* 75:4631, 1970.

Lambert, D. L., "Abundance of Helium in the Sun", *Nature* 215:43, 1967.

Lazarus, A. J., et al., "Interplanetary Shock Observations by Mariner 5 and Explorer 34", *Solar Phys.* 13:232, 1970.

Lazarus, A., et al., "Summary of Results from Mariner 5", *J. Geophys. Res.* (to be published), 1970.

Montgomery, M. D., Bame, S. J., and Hundhausen, A. J., "Solar Wind Electrons", *J. Geophys. Res.* 73:4999, 1968.

Morton, D., "The Abundance of Helium in A and B type Stars", *Astrophys. J.* 151:285, 1968.

Neugebauer, M., and Snyder, C., "Mariner 2 Observations of the Solar Wind", *J. Geophys. Res.* 71:4469, 1966.

Noble, L. M., and Scarf, F. L., "Conductive Heating of the Solar Wind", *Astrophys. J.* 138:1169, 1963.

Ogilvie, K. W., and Lind, D. L., "Electrons at the Bow Shock", *Trans. Amer. Geophys. Union* 49(4):733, 1968.

Ogilvie, K. W., McIlwraith, N., and Wilkerson, T. D., "A Mass-Energy Analyser for Space Plasmas", *Rev. Sci. Inst.* 39:441, 1968.

Ogilvie, K. W., and Burlaga, L. F., "Hydromagnetic Shocks in the Solar Wind", *Solar Phys.* 8:422, 1969.

Ogilvie, K. W., and Wilkerson, T. D., "Helium Abundance in the Solar Wind", *Solar Phys.* 8:435, 1969.

Parker, E. N., *Interplanetary Dynamical Processes*, Interscience Publishers, New York, 1963.

Robbins, D. E., Hundhausen, A. J., and Bame, S. J., "Helium in the Solar Wind", *J. Geophys. Res.* 75:1178, 1970.

Sari, J. W., and Ness, N. F., "Power Spectra of the Interplanetary Magnetic Field, *Solar Phys.* 8:155, 1969.

Scarf, L. F., et al., "Initial Results of Pioneer 8 VLF Electric Field Experiment", *J. Geophys. Res.* 73:6665, 1968.

Scarf, F., "Microscopic Structure of the Solar Wind", *Space Sci. Rev.* (to be published), 1970.

Scarf, L. F., Green, I. M., Siscoe, G. L., Intriligator, D. S., McKibben, D. D., and Wolfe, J. H., "Pioneer 8 Electric Field Measurements", *J. Geophys. Res.* 75:3167, 1970.

Schatten, K. H., et al., "A Model of Interplanetary and Coronal Magnetic Fields", *Solar Phys.* 6:442, 1969.

Siscoe, G. L., et al., "Power Spectra and Discontinuities of the Interplanetary Magnetic Field", *J. Geophys. Res.* 73:61, 1968.

Snyder, C. W., Neugebauer, Marcia, and Rao, U. R., "Solar Wind Velocity and its Correlation", *J. Geophys. Res.* 68:6361, 1963.

Taylor, H. E., "Sudden Commencement Associated Discontinuities in the Interplanetary Medium", *Solar Phys.* 6:320, 1969.

Vasyliunas, V. M., "Deep Space Plasma Measurements", *Methods of Experimental Physics* 9, 1969.

Whang, Y. C., Liu, C. K., and Chang, C. C., "A Viscous Model of the Solar Wind", *Astrophys. J.* 145:255, 1965.

Whang, Y. C., and Chang, C. C., "An Inviscid Model of the Solar Wind", *J. Geophys. Res.* 70:4175, 1965.

Wilcox, J. M., "The Interplanetary Magnetic Field", *Space Sci. Rev.* 8:258, 1968.

Wolfe, J. H., et al., "The Composition, Anisotropic and Non-Radial Flow Characteristics of the Solar Wind", *J. Geophys. Res.* 71:3329, 1966.

Page intentionally left blank

CHAPTER 6
LOWER ATMOSPHERES OF THE PLANETS

Donald M. Hunten
Kitt Peak National Observatory
Tucson, Arizona

I. INTRODUCTION

This discussion will largely ignore upper atmospheres, which may be defined as those regions where photochemistry and ionization are important. Lower atmospheres are therefore mixed, with the exception of condensable constituents, and somewhat stable chemically (examples of "somewhat stable" species in the Earth's atmosphere are O_2, CO, and CH_4). Once the composition is known, the vertical structure of an atmosphere is essentially defined by the temperature profile through the gas laws and the hydrostatic equation. We shall largely ignore the complications of horizontal transport and weather. In most cases, it can be assumed that the atmosphere is horizontally stratified and that this condition is maintained by dynamical processes. Motions in planetary atmospheres have been discussed in some detail by Goody (1969).

We shall begin by reviewing some elementary developments and defining basic terminology. Attention will then be focused in turn on the three planets Mars, Venus, and Jupiter, about which most is known; evidence on composition and structure will be critically discussed. Mercury and the remaining giant planets will receive only brief treatment because much less is known about them.

II. BASIC RELATIONS

A. Equilibrium Temperatures

The energy radiated by a planet of radius a is

$$4\pi a^2 \sigma T_e^4, \tag{1}$$

where $\sigma = 5.67 \times 10^{-5}$ erg-cm^{-2}-s^{-1}-K^{-4} is the Stefan-Boltzmann constant and T_e is the effective temperature; the actual temperature of the radiating surface may be

somewhat higher. For Mars and Mercury, the thermal time constant of the surface is so short that most of the radiation comes from the day side; instead of Equation 1, it is better to write

$$2\pi a^2 \sigma T_e^4 . \tag{2}$$

The energy received from the Sun is

$$\pi a^2 (1 - A) F / R^2 , \tag{3}$$

where F is the solar flux at the Earth (solar constant, 1.40×10^6 erg-cm^{-2}-s^{-1}) and R is the distance from the Sun in astronomical units. The albedo for solar radiation is A, which should properly be obtained as a flux-weighted average:

$$A = \int F_\lambda A_\lambda \, d\lambda \Big/ \int F_\lambda \, d\lambda . \tag{4}$$

Values of A are given by Allen (1963); they are, however, distinctly uncertain because of difficulties of measurement.

By combining Equation 1 with Equation 3, we obtain

$$T_e = [F(1 - A)/4\sigma]^{1/4} R^{-1/2}$$

$$= 280(1 - A)^{1/4} R^{-1/2} . \tag{5a}$$

Similarly, by combining Equation 2 with Equation 3, we obtain

$$T_e = [F(1 - A)/2\sigma]^{1/4} R^{-1/2}$$

$$= 333(1 - A)^{1/4} R^{-1/2} . \tag{5b}$$

Table 1 shows the values given by Equations 5a and 5b for the five innermost planets; Equation 5b, corresponding to a hemisphere, is used for Mercury and Mars and is indicated by an asterisk. Except for Venus and Jupiter, the "observed" values are crude impressionistic averages. The value for Earth refers to the surface, which is not really appropriate, because the effective temperature should be that of the radiating level. This level is near the tropopause and does, indeed, have approximately the expected temperature. This is an example of the well-known "greenhouse" effect. For Venus and Jupiter, the effective temperature is that of an

Table 1.—Calculated equilibrium temperatures.

Planet	A	R	T_e (calculated)	T_e (observed)
Mercury	0.06	0.387	527*	557
Venus	.85	0.723	205	210
Earth	.40	1	246	(290)
Mars	.15	1.523	529*	250
Jupiter	.58	5.20	99	134

atmospheric level or "cloud top". A comparison of the calculated and observed values for Jupiter indicates that it radiates more than twice the energy it receives. Because of its great interest, this result has been carefully checked, and it appears to be real (Aumann et al., 1969). Thus, Jupiter is probably generating internal heat roughly equal to that which it receives from the Sun.

B. Barometric Equations

Pressure p, density ρ, and temperature T are related by the ideal-gas law:

$$p = \rho kT/m = \rho RT/M$$

or

$$dp/p = d\rho/\rho + dT/T \,. \tag{6}$$

Deep in the Jupiter atmosphere, and to some extent in the Venus atmosphere, the gas may depart appreciably from Equation 6, but it is usually an excellent approximation. The hydrostatic equation is

$$dp/dz = -\rho g = -pmg/kT = -p/H \,, \tag{7}$$

where $H = kT/mg = RT/Mg$ is the scale height. If H is independent of height z, Equation 7 can be integrated to

$$p = p_0 \exp(-z/H)$$

and

$$\rho = \rho_0 \exp(-z/H) \,. \tag{8}$$

The constants of integration p_0 and ρ_0 are calculated at $z = 0$, which can be any arbitrary reference level. Though H is seldom independent of height over an

Table 2.–Planetary data: gravity g, pressures at surface P_s and cloud top P_c, scale height H, and dry adiabatic lapse rate Γ.

Planet	Gas	M	g (cm-s^{-2})	P_s (b)	P_c (b)	T (K)	H (km)	Γ (K-km^{-1})
Venus	CO_2	44	884	100	0.1	210	4.5	11
Earth	N_2,O_2	29.2	982	1.013		250	7	10
Mars	CO_2	44	376	.006		200	10	5
Jupiter	H_2	2	2600		2	130	20	1.9

appreciable range, the form of Equation 8 is still useful as a limit. It shows that pressure and density should be plotted as logarithms against height, and the curves will be nearly straight. It also brings out the significance of the scale height as the characteristic length of the phenomenon, the distance over which p and ρ vary by a factor of e. Typical values for the various planets are shown in Table 2.

The hydrostatic equation can also be integrated when H is a linear function of height: $H = H_0 + \beta z$, so that β is the scale-height gradient. We obtain

$$p = p_0[1 + (\beta z/H_0)]^{-1/\beta} = p_0(H/H_0)^{q+1}$$

and

$$\rho = \rho_0(H/H_0)^q . \tag{9}$$

In the last two expressions we have used the polytropic index q, defined by the relation ρ = constant $\times$ T^q so that $-1/\beta = q + 1$. Some values of β and q are given in Table 3. The pressure and density now vary at different rates; it is often convenient to define the density scale height as

$$H_\rho = -\rho/(d\rho/dz) = (1 + \beta)H = H[q/(q + 1)] . \tag{10}$$

C. Integrated Densities

Integrated densities are important because they relate to observable quantities: absorption spectroscopy to a near-vertical integration, and radio occultations to a

Table 3.–Adiabatic atmospheres: lapse rate Γ, specific-heat ratio γ, polytropic index q, and scale-height gradient β.

Planet	Γ (K-km^{-1})	γ	q	β
Venus				
100 b	7.3	1.19	5.26	-0.160
0.1 b	11	1.32	3.13	-0.242
Earth				
wet	6.5	(1.23)	4.35	-0.187
dry	10	1.4	2.5	-0.286
Mars	5	1.37	2.70	-0.270
Jupiter	1.9	1.4	2.5	-0.286

horizontal integration past the limb. It is convenient to use the number density, or concentration, $n(z) = \rho(z)/m$, and the integrated, or column, density

$$N(z_0) = \int_{z_0}^{\infty} n(z)\, dz\ .$$

For both Equations 8 and 9, the result is

$$N(z_0) = n(z_0)H(z_0) = n(z_0)H_0\ . \tag{11}$$

In view of Equation 11, H is sometimes called "the height of the homogeneous atmosphere", particularly in other languages.

In spectroscopic observation of a planetary atmosphere, the solar radiation usually enters at a local zenith angle θ_0, and the reflected radiation leaves at angle θ. The integrated density is then written ηN, where the air-mass factor is $\eta = \sec\theta_0 + \sec\theta$. It is frequently necessary to average this factor over those portions of the planet included in a particular observation.

The hydrostatic equation (Equation 7) may be written

$$p(z_0) = g\int_{z_0}^{\infty} \rho\, dz = g\rho(z_0)H(z_0)\ . \tag{12}$$

This relation is exact except for the small variation of gravity with height; pressure is proportional to integrated density. In terms of number density, Equation 12 becomes

$$p(z_0) = mgN(z_0) = MgN(z_0)/N_A\,, \tag{13}$$

where N_A is Avogadro's number.

In spectroscopic work it is customary to express the integrated density of a gas in cm-atm (m-atm, km-atm), the amount of gas in a column of the stated length at STP. A more correct but less common usage is "Amagat", the corresponding density unit, instead of atmosphere. Because the concentration at STP is Loschmidt's number per cm^3, we have the relation

$$1 \text{ cm-atm} = 1 \text{ cm-Agt} = 2.687 \times 10^{19} \text{ cm}^{-2}\,. \tag{14}$$

For H_2O, which condenses far below atmospheric pressure, the concept of "cm-atm" is not directly applicable, though it could be used through the relation in Equation 14. Instead, the common units are g-cm^{-2} and "microns of precipitable water", which means the depth of the liquid layer that would be formed by condensing all the water in the column. A precipitable micron is therefore 10^{-4} g-cm^{-2}, or 3.35×10^{18} molecules-cm^{-2} and is equivalent to 0.125 cm-atm.

So far we have neglected the curvature of the planet; we now turn to the opposite limit, an integration past the horizon. We require first a relation between path length x and height z; from Figure 1, we obtain

$$(a + z)^2 = a^2 + x^2$$

$$z \approx x^2/2a\,, \tag{15}$$

in which a term $z/2a$ has been dropped. Because the important range of z is of the order of a scale height, this approximation is an excellent one. For an exponential atmosphere, the integral along x is

$$N_h(z_0) = n(z_0) \int_{-\infty}^{\infty} \exp(-x^2/2aH)\, dx \doteq n(z_0)(2\pi aH)^{1/2}\,. \tag{16}$$

The air-mass factor for a full traversal is found by dividing Equation 16 by $n(z_0)\,H$; thus,

$$\eta_h = (2\pi a/H)^{1/2}\,. \tag{17}$$

Typical values of η_h are 50 to 100 for terrestrial planets and 470 for Jupiter.

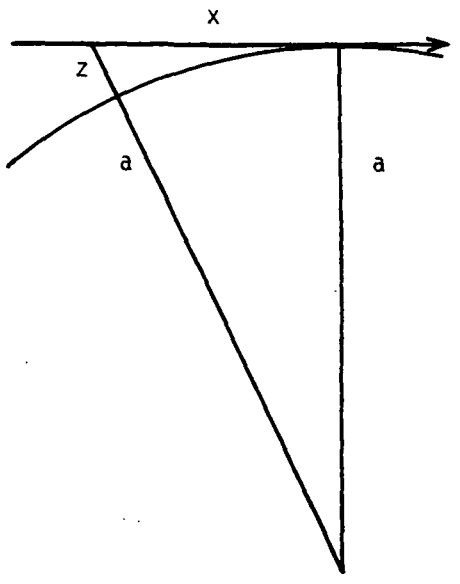


Figure 1.–Geometry for the calculation of horizontal air mass.

With a scale-height gradient the integral is much more difficult to evaluate, and analytic results are available only for certain values of the gradient β. If Equations 16 and 17 are not accurate enough, it is probably best to do the integral numerically.

Another quantity of interest for a horizontal ray is the refraction α. The optical path for a ray, relative to a vacuum path of the same length, is $L(z_0) = R(z_0)N_h(z_0)$, where R is the refractivity, the refractive index minus unity. In radians, α is just the derivative of L with respect to z_0; if R varies exponentially, we obtain

$$\alpha(z_0) = -L/H = -R(z_0)(2\pi a/H)^{1/2} = -\eta_h R(z_0) . \tag{18}$$

For example, at the Earth's surface, $R = 2.93 \times 10^{-4}$ and $\eta_h = 71$; we thus find $\alpha = 0.0208$ radian, or 72 minutes of arc. The observed depression of astronomical bodies at the horizon should be half as great, and the value is 35 minutes, which is in good agreement. The next derivative of Equation 18 gives the differential refraction, which is responsible for the observed attenuation of radio and optical signals in an occultation experiment.

D. Adiabatic Temperature Gradient

A gas in rapid vertical motion will take up the adiabatic temperature gradient, usually described by its negative, the adiabatic lapse rate Γ. Logarithmic differentiation of the adiabatic relation $p = \text{const} \times \rho^{\gamma}$ gives

$$dp/p = \gamma\, d\rho/\rho\,. \tag{19}$$

This relation may be used to eliminate $d\rho/\rho$ from Equation 6, and then from Equation 7 dp/p can be replaced by $-dz/H$; the result is

$$\left(-\frac{dT}{dz}\right)_{\text{ad}} = \Gamma = \frac{T}{H}\frac{\gamma - 1}{\gamma} = \frac{Mg}{R}\frac{C_p - C_v}{C_p} = \frac{Mg}{C_p} = \frac{g}{c_p}, \tag{20}$$

where C is the molar specific heat and c is the specific heat per gram. Values of Γ and γ (the ratio of specific heats) are shown in Table 3, as well as the corresponding polytropic index and scale-height gradient. The required relations are

$$\beta = \frac{dH}{dz} = \frac{R}{Mg}\frac{dT}{dz} = \frac{H}{T}\frac{dT}{dz} = -\frac{H}{T}\Gamma = -\frac{\gamma - 1}{\gamma} \tag{21}$$

and

$$q = 1/(\gamma - 1) = -(1 + \beta)/\beta\,. \tag{22}$$

In the Earth's troposphere, condensation of water reduces the lapse rate considerably; the mean situation has a "wet adiabatic" lapse rate of 6.5 K-km^{-1} and can be represented by an effective γ of 1.23.

Under what conditions should we expect to find an adiabatic lapse rate? One example that has been thoroughly studied is a stellar interior. Under many conditions, the outward flow of radiation by itself would produce a lapse rate much greater than adiabatic. Such a situation is unstable against vertical interchange of gas parcels, and the result is to set up a regime of free convection. The interchange of gas is an extremely efficient heat-transfer process, and it can be shown that the adiabatic lapse rate is exceeded by only a tiny amount (Schwarzschild, 1958). In a thin planetary atmosphere, such as that of Mars, a similar situation develops, though for a very different reason. Here, the surface is strongly heated by solar radiation; some of this heat is communicated to the atmosphere, which can only carry the

required flux by convection. Again, the lapse rate is expected to be very close to adiabatic.

Though free convection always implies the adiabatic lapse rate, the converse is not true. On Earth, for example, we are aware of constant horizontal motion in the form of various sorts of weather. These motions are mainly driven by the temperature difference between equator and poles. Clearly, they are potentially able to generate a considerable amount of vertical motion as well, and if these vertical motions are fast enough they will produce the adiabatic lapse rate (complicated by the presence of water vapor). With such forced convection, the actual lapse rate is slightly *less* than the adiabatic value, instead of slightly greater, but not necessarily by a detectable amount. For Earth, we cannot yet be sure whether free or forced convection dominates, or perhaps a combination. The beginnings of a treatment have been given by Gierasch and Goody (1969), but much more work is needed.

On Venus, a near-adiabatic lapse rate has been observed or inferred all the way from the cloud tops to the neighborhood of the surface. No known heat source is available to provide for free convection, and it seems more probable that there is a large-scale circulation driven by heating in the region of the cloud tops. If so, there is a small downward convective heat flux, balanced by a small upward radiative flux.

For Jupiter, with its internal heat source, it seems likely that free convection extends very deep and that the planet has some resemblance to a stellar interior, but observational evidence other than that discussed previously is almost entirely lacking. Further discussion in connection with individual planets will follow.

E. Radiative Heat Transfer

The opposite limit to an adiabatic atmosphere is one whose temperature profile is controlled by radiation. At typical temperatures of 100 to 300 K we are speaking of thermal radiation with a peak at 30 to 10 μm. To interact with the atmosphere, this radiation must be absorbed and reemitted; relevant absorbers are H_2O and CO_2 on Earth, CO_2 on Mars and Venus, and CH_4, NH_3, and H_2 (pressure-induced) on the giant planets. A realistic treatment must clearly be very complicated, and entire books have been written about it (Goody, 1964; Kondratyev, 1969). Here we shall introduce only the very simplest limiting case, which is nevertheless surprisingly appropriate to a real situation, as discussed by Goody (1964). The upward thermal flux F_t is constant; thus, no allowance is made for solar radiation absorbed within the atmosphere. The absorption coefficient k is independent of wavelength through the thermal part of the spectrum (gray absorption). This simplification is more

accurate than might be supposed, because only those wavelength bands that are actually absorbed need to be considered; the rest are simply transmitted without interaction. The independent variable is optical depth τ, measured downwards from the top of the atmosphere:

$$\tau = \int_z^\infty k\,dz\,.$$

If the radiation field is represented by two streams, one up and one down, the transfer is described by the Milne-Eddington relation:

$$T^4 = (F_t/2\sigma)(1 + 3\tau/2)\,. \tag{23}$$

At the lower boundary (the planetary surface), there must be a temperature jump; otherwise, there could be no upward flux. If the surface temperature is T_s and the atmospheric temperature just above the surface is $T(\tau^*)$, the relation is

$$T_s^4 - T^4(\tau^*) = F_t/2\sigma\,. \tag{24}$$

Thus, for the surface temperature we may combine Equations 23 and 24 to get

$$T_s^4 = (F_t/\sigma)(1 + 3\tau^*/4)\,. \tag{25}$$

The temperature jump is unstable against the setting up of convection; it follows that pure radiative equilibrium is always unstable when there is a lower boundary and an upward radiation flux. The magnitude of the jump goes as T_s^{-3} and can therefore be small at high temperatures.

As an example, for the Earth the mean value of F_t/σ is 3.64×10^9 K^4 (Goody, 1964), and τ^* is approximately 1. From Equation 25 we obtain T_s = 282 K, and from Equation 23 $T(\tau^*)$ = 260 K. Also, if we set $\tau = 0$, we find $T(0)$ = 207 K, which we may identify with the stratospheric temperature. As Goody discusses, a convective layer will be set up with a depth of 7.7 km; at greater heights the radiative curve will apply.

Another consequence of Equation 23 follows from the fact that most of the opacity will normally be concentrated near the surface, in a barometric type of distribution. If τ^* is considerably greater than 1, the lapse rate near the surface is

likely to become superadiabatic, quite apart from the jump at the surface. Again, this situation will be destroyed by the onset of convection.

The stratospheric temperature $T(0)$ can be obtained very simply, without recourse to Equation 23, by a method due to Gold and Humphreys. Let us suppose the effective temperature of the planet to be independently determined by radiative balance, as discussed above; the relation is $T_e^4 = F_t/\sigma$. Now we consider a volume element in a region with $\tau \ll 1$; it receives thermal radiation from a hemisphere and emits over an entire sphere; thus,

$$\sigma T^4(0) = \sigma T_e^4/2,$$

or

$$T(0) = 2^{-1/4} T_e = 0.841 T_e . \tag{26}$$

This result is consistent with Equation 23 but is much more general; it does not depend on any kind of grayness, but only on Kirchhoff's law that the same radiations are absorbed and emitted efficiently by any substance. However, Equation 26 still ignores any heating within the atmosphere by solar radiation; for this reason it applies only to a small part of the Earth's lower stratosphere. Ozone absorption below 3000Å provides a strong heat source that leads to a temperature maximum around 50 km. Similar effects may occur on other planets, though so far a temperature rise has been predicted only for Jupiter.

The greenhouse effect takes place when an atmosphere is relatively transparent in the visible, and more opaque in the infrared. Thus, solar radiation can reach the surface, whereas the escape of thermal radiation is blocked in the manner described by Equation 25, with F_t determined by the solar flux absorbed at the ground. If τ^* is large, there can be a correspondingly large temperature rise; however, such a large value may be difficult to reconcile with a high transparency in the visible. And it must always be remembered that the profile described by Equation 25 is unstable against convection, which will carry the flux with a smaller temperature rise.

III. MARS

Mars has a thin atmosphere, free of extensive cloudiness and transparent to the bulk of the Sun's radiation. Thus, most of the heat is deposited at the surface, and an understanding of atmospheric structure should be relatively easy in comparison with the other planets. A further simplifying factor is the lack of oceans.

Surface temperatures, which give the driving forces for atmospheric motions, can be calculated as if the atmosphere were absent. Radiometric observations of the day side have been analyzed by Gifford (1956). Since then, considerably more detailed observations have been obtained by Sinton and Strong (1960) (see also Leovy, 1966a), and a thorough analysis has recently been presented by Morrison et al. (1969). The data can be fitted by a model of radiation and conduction in a medium that turns out to be a poor conductor; a dust with a particle size around 200 μm is suggested. A brief report of the Mariner 6 and 7 results confirms the temperatures deduced for the night side (Neugebauer et al., 1969). Near the equator, the temperature reaches 300 K just after noon; by sunset it is 230 K, and just before dawn it is 180 K. Noon temperatures of 220 to 250 K are typical at 60° latitude. Dry ice is expected to condense out of the atmosphere (pressure 6.5 mb) at 148 K; Mariner 7 observed essentially this temperature for the south polar cap. Dry-ice frost can be expected in the early morning at high latitudes, in addition to snow in the winter polar region.

A. Atmospheric Composition

Nearly all our evidence on composition comes from Earth-based spectroscopy; radio occultations of spacecraft, discussed below, give information on total pressure and abundance of CO_2, the major constituent. Most of the information is obtained from vibration-rotation bands in the near infrared, where detector sensitivity is high; the stronger far-infrared bands cannot be studied from the Earth's surface in any case because of atmospheric absorption. For O_2, we use a forbidden electronic transition, and for O_3 its ultraviolet continuum.

The partial pressure of CO_2 is obtained from bands at 10 500Å (123-000) and 10 380Å (203-000), where the designations in parentheses refer to the three vibrational quantum numbers of the upper and lower states (Figures 2 and 3). The most complete study is that of Belton et al. (1968a); two other values (Giver et al., 1968; Carleton et al., 1969) obtained at about the same time are in close agreement. The abundance is found to be 78 ± 11 m-atm; by Equation 12, the corresponding partial pressure is 5.5 ± 0.8 mb. Much of the uncertainty is from the difficult laboratory measurement of the band strength, which requires a very long path of gas at a low pressure.

To obtain the total pressure spectroscopically, we can use a strong line, which is black at the center so that the equivalent width is entirely determined by the

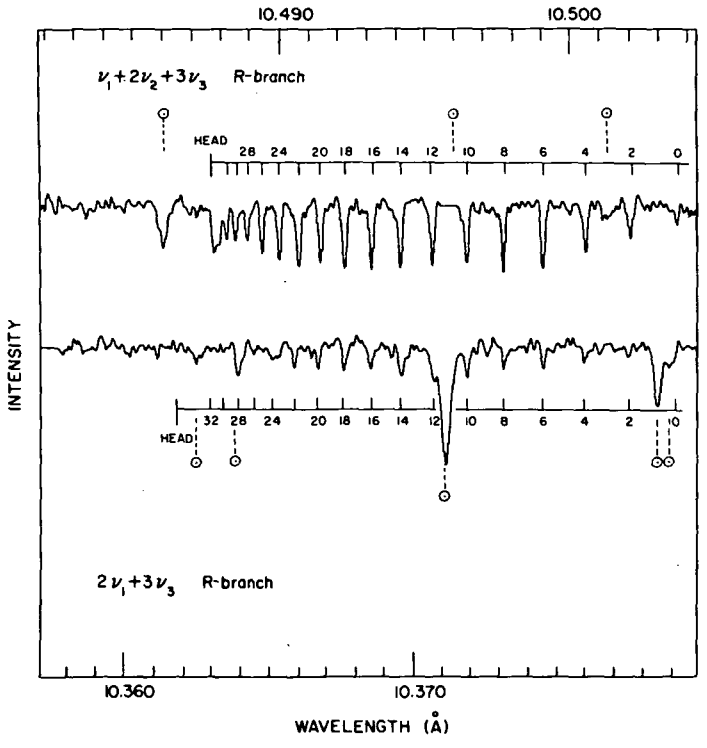


Figure 2.—Two CO_2 bands in the spectrum of Mars (Belton et al., 1968). Solar lines are flagged with circle-dot symbols.

Lorentz wings. In this situation, the equivalent width is proportional to the square root of the product of the amount of absorbing gas and the total pressure. The first analysis of Mars spectra based on this relation was given by Grandjean and Goody, and a detailed discussion of the early work with strong CO_2 bands is given by Chamberlain and Hunten (1965). These bands, in the 1.6- and 2.0-μm regions, are so complicated that they are not suitable for a really accurate result. Kaplan, Münch, and Spinrad (1964) obtained the first spectrum of a weak CO_2 band and were therefore able to derive both a CO_2 abundance and a total pressure. With better spectra, it has been possible to obtain more accurate results (Belton et al., 1968a; Giver et al., 1968; Carleton et al., 1969), but a considerable uncertainty has remained. The Connes Atlas (Connes et al., 1969) shows strong and weak bands of CO resolved into the individual lines and therefore more useful; according to Kaplan

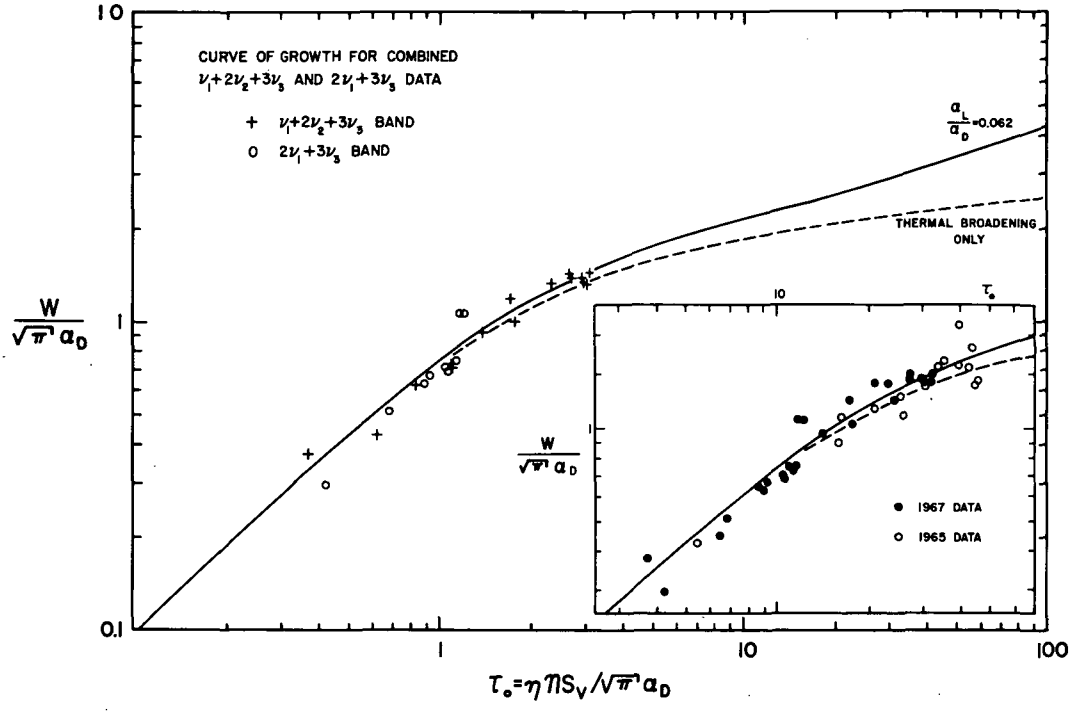


Figure 3.—Curve of growth for the lines of Figure 2. The inset includes additional data with a larger air mass.

et al. (1969), the surface pressure is found to be 5.3 mb, very close to the partial pressure of CO_2. Considering the possible errors in both results, we may take the fractional abundance of CO_2 as 80 to 100 percent.

A completely independent method is to use the occultation of the telemetry carrier of a space probe; this has been successful with Mariners 4 (Fjeldbo and Eshleman, 1968; Kliore et al., 1965), 6, and 7. A scale height can also be found, and with the available knowledge of the temperatures, the mean molecular weight must be close to that of CO_2. For pure CO_2, the mean of the six determinations is 6.4 mb; with an admixture of argon or nitrogen, the value is slightly greater. The agreement between the occultation and spectroscopic results is excellent.

The CO abundance from the weak 3-0 band is 5.6 cm-atm (Kaplan et al., 1969); the mixing ratio CO/CO_2 is therefore 8×10^{-4} (by number or by volume).

Water vapor is difficult to observe because of the strong absorption in the Earth's atmosphere. It is necessary to observe from a dry site at times when the planetary Doppler shift is a maximum; these times are close to quadrature, and occur 2 to 3 months before and after opposition. A typical shift of 0.3Å at 8200Å is about three times the typical spectral resolution. Spatial resolution of the planet is difficult to obtain because the image is much smaller than it is at opposition.

The band that is used most is the 211-000, especially the strong, isolated line at 8189Å. Unsuccessful searches were made in the 1930's by Adams and Dunham, but the first positive result was not obtained until 1963 (Kaplan et al., 1964). More recent work (Schorn et al., 1967; Owen and Mason, 1969; Schorn et al., 1969b) has shown that the amount is variable, sometimes falling below the detection threshold, which is around 10 to 15 precipitable microns, or 5×10^{19} cm^{-2}. The largest reported amount is 35 μm, and there is evidence that sometimes the observable water is confined to one hemisphere. If we assume that the water vapor is confined to a 2-km layer, while the effective depth of the CO_2 atmosphere is 10 km, the mixing ratio in that layer is 1.2×10^{-3} to 3×10^{-3}, or perhaps much less at those times when H_2O lines are not found.

Solar radiation in the 1950Å region penetrates to the surface of Mars and is capable of dissociating water vapor. A considerable part of the free hydrogen thus produced is expected to escape from the planet, and there is a corresponding substantial source of O_2. These processes have been treated in some detail (Hunten and McElroy, 1970), and the escape flux of H atoms has been estimated at around 10^8 cm^{-2}-s^{-1}. If the global mean abundance of H_2O is 3×10^{19} cm^{-2}, the atmospheric water vapor must be replenished every 3×10^{11} s, or 10^4 Earth years. Presumably, the vapor diffuses out of the ground from a buried layer of permafrost. In 5×10^9 years, the current loss rate implies a total consumption of about 2.5 m of solid ice.

Oxygen also requires for its detection a large Doppler shift and a small telluric air mass. Under favorable conditions, the strongest lines of the Fraunhofer *A* band at 7619Å are accessible to observation. Belton and Hunten (1968, 1969a) were able to remove most of the telluric absorption from the spectrum by forming the ratio with a solar spectrum. Of the two lines expected in the few angstroms observed, one was present and the other, though perhaps present, was displaced from the expected wavelength. They therefore suggested a "possible detection" of about 20 cm-atm, or an upper limit of the same amount. The corresponding mixing ratio with CO_2 is 2.6×10^{-3}. The escape rate of hydrogen suggested above would leave behind this much

O_2 in 10^{13} s, or 3×10^5 years; the presence of such a strong source lends credibility to the possible spectroscopic detection.

Ozone can, in principle, be detected through its strong ultraviolet continuum, with a maximum at 2550Å; it is responsible for the cutoff of the Earth's atmosphere at about 3000Å. A rocket spectrum by Broadfoot and Wallace (1970) has its short-wavelength limit just at the peak of the continuum; there is no clear evidence of absorption, and only an upper limit of 2×10^{-4} cm-atm can be set. Unfortunately, this difficulty must be common to any attempts to observe weak ozone absorption against the background of the planetary surface: Even if a small, broad dip is seen, there is no assurance that it is not a feature of the surface albedo instead of the atmosphere. Preliminary accounts of data from OAO 2 and Mariners 6 and 7 suggest a detection of approximately the above amount, but in view of the inherent ambiguity, it is best to regard this as still an upper limit.

Photochemical models exist that relate the O_2 and O_3, but there are still major uncertainties. Belton and Hunten (1968, 1969a) treated the system as if CO_2 and CO were photochemically inert, and obtained fair agreement between the two tentative abundances. McElroy and Hunten (1970) included photodissociation of CO_2 as a source of O atoms and added some further, somewhat speculative, reactions to oxidize the resulting CO. The range of results includes the "observed" ozone toward one end but goes several orders of magnitude smaller at the other.

Nitrogen and argon are the most likely gases that cannot be detected by absorption spectroscopy (except by very difficult measurements below 1000Å). As we have seen, the current estimates of the total pressure leave room for as much as 20 percent of another gas. On Earth, nitrogen is prominent in emission spectra, including the dayglow for which we have Mars observations from Mariners 6 and 7. These spectra show no signs of N_2, N_2^+, NO, or CN (Barth et al., 1969) and imply an N_2 mixing ratio probably less than 5 percent (Dalgarno and McElroy, 1970).

The dominance of CO_2 on Mars (and Venus) is not as strange as it appears. Terrestrial CO_2 does not remain in the atmosphere for long: It dissolves in water and then precipitates as carbonate. Rubey (1951) has estimated the amount so stored as 18 kg-cm^{-2}, or just under 18 atm. The much smaller value for Mars (15 g-cm^{-2}) suggests a much smaller degree of degassing for that planet; allowing for its smaller volume, the ratio is 1.5 percent. Corresponding to the Earth's 327 kg-cm^{-2} of water, Mars should have 270 g-cm^{-2}, which is close to the estimated consumption (220 g-cm^{-2}) from photodissociation and escape. The terrestrial ratio N_2/CO_2 is 7 percent by volume, just compatible with the available evidence for Mars.

B. Atmospheric Structure

The structure of an atmosphere is defined by the variation with height of the densities of the major constituents. For a dry, mixed atmosphere of known mean mass, the temperature profile is enough to specify the structure completely. Detailed models for the Martian lower atmosphere are discussed below; they predict a nearly isothermal stratosphere, at about 150 K, above a troposphere whose depth is some 15 km at low latitudes, decreasing toward the poles. There is strong convection on the day side, and a stable inversion at night.

CO_2 spectra, such as Figure 2, contain about 15 resolved lines, which give 15 points on the rotational energy distribution of the molecules. Further information is present in the band head, which consists of several unresolved lines from relatively high levels. Three parameters can be extracted with some confidence: a mean temperature, a maximum temperature, and a lapse rate. For the spectrum shown, which refers to the subsolar region, these results are 195 K, 310 K, and 4.6 K-km^{-1}, respectively (Belton et al., 1968a).

Even more information can be obtained from the occultation of a space probe. Because of the importance of these results and the necessity of understanding their limitations, the method will be discussed in some detail. A signal is transmitted at a wavelength of 14.3 cm from the Earth to the spacecraft. The received frequency is coherently multiplied by a factor of 11/10 and transmitted back to the Earth at the new wavelength, 13.0 cm, as the telemetry carrier. After a similar transformation at the station, the difference, or Doppler, frequency is recorded. By far the largest part of this is due to the radial velocity of the spacecraft with respect to the Earth (150 kHz for a typical velocity of 10 km-s^{-1}). However, previous tracking by the same system gives an accurate knowledge of the orbit, which permits the subtraction of this component. At this point it becomes convenient to think in terms of cumulative phase change, which corresponds to a displacement of the apparent position of the spacecraft from its true position. Small corrections must be made for the changing path length and electron density in the Earth's ionosphere. The remaining effects, due to the planetary ionosphere and atmosphere, are illustrated in Figure 4. These phase shifts represent the advance due to the ionosphere and the retardation due to the atmosphere, and an additional effect related to the bending of the ray by refraction.

The effect of the ionosphere gives valuable information about the upper atmosphere, but for our purpose it is a nuisance that must be eliminated. The electron-density profile can be obtained for the region that is traversed by the rays;

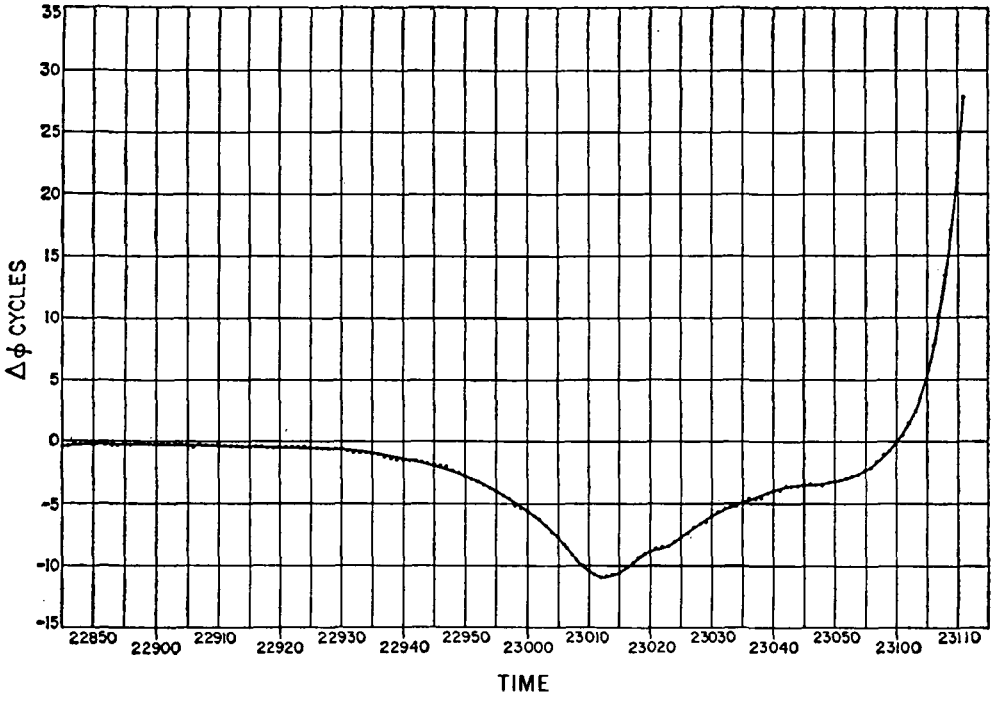


Figure 4.—Phase shifts from the entry of Mariner 4 into occulation by Mars (Kliore et al., 1965). The independent variable can be regarded as either time or relative height. The negative peak, due to the ionosphere, is at about 125 km.

but when the lower atmosphere is probed, the ionosphere is penetrated at two regions that have not been directly sampled. If the Martian ionosphere is as unpredictable as that of the Earth, a significant and inherent error is likely to occur. The effect is small for the large signal near the ground but could completely alter the picture for heights above 20 km. The ideal remedy would have been to use two frequencies; the large dispersion of a plasma would then permit an unambiguous separation of the two components, as well as a better allowance for the Earth's ionosphere. Lacking this refinement, we must at least remain aware of the resulting uncertainties.

The above description applies specifically to the day side of Mars. Complications due to the ionosphere are absent on the night side, where the electron density is too low to be detected. On the other hand, closed-loop operation is not practical,

because for the existing night-side measurements, the spacecraft has just emerged from behind the planet. Mariner 4 measurements suffered a serious perturbation when phase lock occurred at a grazing height of about 12 km. More recent measurements have been made in the open-loop mode, trusting the stability of the oscillator on the spacecraft.

Allowance for refraction, mentioned above, is straightforward, and the final data are a series of phase shifts $\phi(z_0)$ as a function of grazing height z_0. Using the refractivity R as in Equation 18, we have, from Equation 16,

$$\phi(z_0) = R(z_0)(2\pi aH)^{1/2}.$$

Analysis of the data therefore requires a simultaneous evaluation of R and H. The same situation arises in density measurements from satellite drag (King-Hele, 1959), and the remedy is to refer the result to a height $z_0 + H/2$. Indeed, it can be seen from Equation 15 that most of the phase shift is produced by the first half-scale height above z_0; the resolution of the method is approximately $H/2$. In practice, a process of integral inversion (Fjeldbo and Eshleman, 1968) is used to sharpen the height resolution somewhat, but the information thus gained is necessarily limited.

To convert the refractivity profile to a density profile, the composition of the gas must be known. The relationship is given by Fjeldbo and Eshleman (1968) for mixtures of CO_2, N_2, and A. The relative proportions of the latter two do not matter much, but CO_2 is considerably more refracting. The uncertainty is significant but not large for atmospheres of 80 to 100 percent CO_2, and it takes the form of a constant scale factor for both densities and temperatures.

To find a temperature profile, we use Equation 10 and the definition of scale height:

$$T = -\frac{mg\rho}{k(1+\beta)\,d\rho/dz} = -\frac{mg}{k(1+\beta)\,d(\log\rho)/dz}. \tag{28}$$

Because of the differentiation, the temperatures are much less certain than the densities; because of the logarithmic derivative, the errors grow rapidly with height. In particular, the temperatures above 15 km are very sensitive to the exact choice of zero point on the phase-shift scale. Lapse rates are obtained by still another differentiation and must be treated with extreme caution. Statistical treatments do not give a realistic estimate of the errors in such a case, which is dominated by systematic error.

We emphasize that much of the above discussion is specific to Mars, where the signal from the neutral atmosphere is exceedingly small. For a denser atmosphere, even the lapse rates carry great confidence.

Table 4 gives the surface pressures and temperatures found by the three Mars Mariners, on the assumption of essentially pure CO_2; for 80 percent CO_2, all the values would increase by about 10 percent. For Mariners 6 and 7, the revised temperature profiles of Rasool et al. (1970) are shown in Figure 5 (see also Kliore et al., 1969).

Table 4.—Data from Mariners 4, 6, and 7.

Parameter	Mariner 4		Mariner 6		Mariner 7	
	Day Side	Night Side	Day Side	Night Side	Day Side	Night Side
Pressure (mb)	4.5	8.0	6.0	7.6	4.9	7.5
Temperature (K)	160	210	250	164	224	205

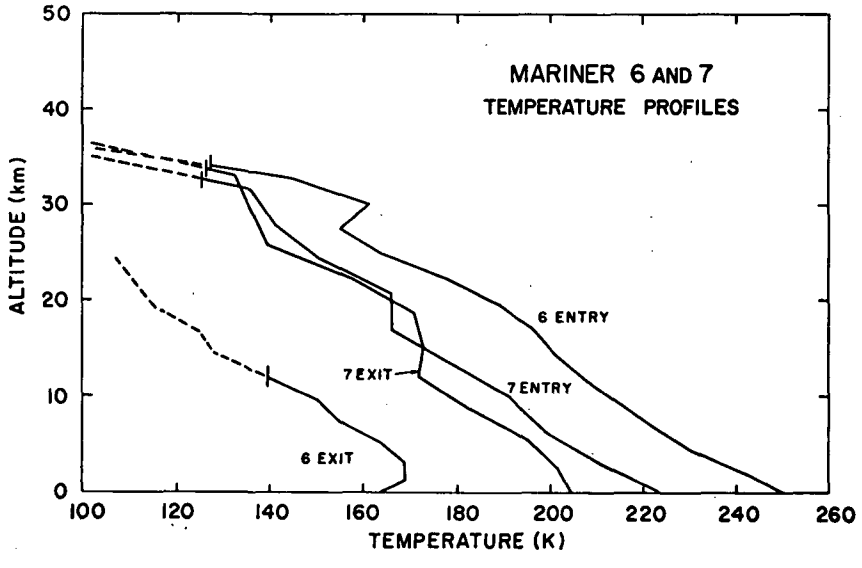


Figure 5.—Vertical temperature distributions in the atmosphere of Mars at the four occultation points of Mariners 6 and 7 (Rasool et al., 1970).

C. Convection and Circulation

Nearly all our information on motions in the Mars atmosphere comes from theoretical models; the three available studies cover different aspects and complement one another very well. As noted at the beginning of this section, dynamical modelling of the Martian atmosphere is much easier than for Earth because of the absence of many large nonlinear effects related to the presence of the oceans and to condensation of water vapor.

In the work of Gierasch and Goody (1968), stress was laid on the vertical structure at a few selected points. Transfer of radiation was modelled in detail for the CO_2, although during the day it is much less important than convection. Priestley's theory was used for free convection, and the boundary layer at the surface (corresponding in part to the temperature jump discussed above) was handled by an approximate method due to Kraichnan. These methods are semiempirical and somewhat controversial, but they do work in similar terrestrial situations. The diurnal variation of ground temperature was taken from the work of Sinton and Strong. Typical results for the equatorial region are shown in Figure 6, which deserves careful study. The day begins with a deep temperature inversion, built up by radiation during the night; a similar effect is familiar on Earth during calm, clear weather, but the depth of the inversion is only 100 to 200 m. At the foot of each curve is a mark showing the air temperature at the surface and a number showing the temperature of the surface; the difference is absorbed by the thin boundary layer. At noon, the jump amounts to 69 K, and convection reaches to 6.5 km; at 4 p.m., the values are 36 K and 16 km. This last height corresponds to the mean tropopause for low latitudes; above it is a nearly isothermal stratosphere at about 153 K. Convection ceases almost at the moment of sunset, and the radiative inversion starts to build up; at 8 p.m. it is already 500 m deep. At a slightly later time, the temperature jump disappears, as indicated by the absence of numbers for the ground temperature.

Some results are also given for 45° latitude in winter. Corresponding to a smaller range of surface temperature, the diurnal variation is less pronounced; the tropopause is found at 6.5 km, and the stratosphere is at 139 K.

Gierasch and Goody also give estimates of the vertical eddy diffusion coefficient K; in the early afternoon, it increases from about 10^8 $cm^2 \cdot s^{-1}$ near the ground to over 5×10^8 $cm^2 \cdot s^{-1}$ at 10 km. The corresponding mixing times are astonishingly short; using the approximation L^2/K for a distance L, we find about a minute for $L = 1$ km, or an hour for 10 km. The frequent presence of dust in the

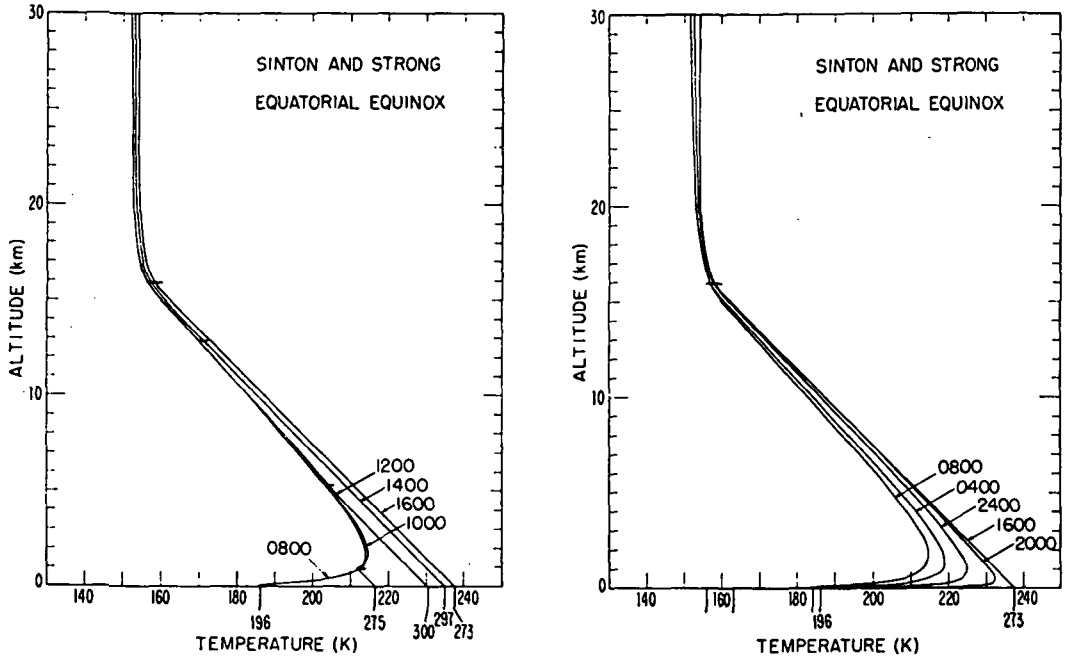


Figure 6.—Daytime (left) and nighttime (right) temperature profiles for Martian equatorial equinox (Gierasch and Goody, 1968). "Sinton and Strong" refers to the surface temperatures used as input.

atmosphere is hardly surprising. There are also important implications for reactive trace constituents such as ozone, which are brought frequently in contact with the surface, where they may be destroyed.

The study of Leighton and Murray (1966) concentrates on the formation and disappearance of the polar caps. The thermal model, based again on the Sinton and Strong data, showed that condensation of CO_2 was to be expected over the winter pole (this was confirmed by Gierasch and Goody) and that a cap should be built up whose behavior is remarkably like what is observed on Mars. The northern cap was predicted to be permanent, again as observed, because its summer is significantly cooler: This summer occurs when Mars is farthest from the Sun. Such a permanent cap could easily contain more mass than the entire atmosphere; Leighton and Murray argue that it closely regulates the partial pressure of CO_2 at the value appropriate to the cap temperature, controlled mainly by radiative equilibrium. This

study established CO_2 as the material of the caps beyond any reasonable doubt; the temperature measured by Mariner 7 provided full confirmation.

The previous opinion had been heavily in favor of water ice. on the basis of infrared spectra by Kuiper and Moroz that showed absorption between 1.5 and 1.8 μm. Leighton and Murray commented that this absorption could be due to H_2O dissolved in the solid CO_2 or forming a surface film as the CO_2 sublimes away. A careful laboratory study has since been made by Kieffer (1970), who confirms both suggestions. His evaluation of the planetary spectra leads him to prefer the second explanation, which is plausible because both observations were made during late spring or early summer for the cap in question.

Leighton and Murray estimate the mass transferred into and out of the caps as the equivalent of about 1 mb of atmosphere; there should be a semiannual pressure variation of about this magnitude. They also point out that the northern cap will eventually disappear as its summer moves to the neighborhood of perihelion, and a permanent southern cap will then start to form. Two effects are responsible: the rotation of the line of apsides of Mars' orbit (72 000-year period) and the precession of Mars' axis in the opposite direction (estimated at 180 000 years); the net period is about 50 000 years. The ice that is now permanently trapped in the north will evaporate and will slowly migrate to the south by what amounts to horizontal eddy diffusion. Such diffusion is a rather inefficient transport process, and it is estimated that the buildup of ice goes at a rate of about 1 cm per 200 years. For the same reason, a temporary cap of ice or snow is highly improbable; not enough water vapor can be transported in a single winter.

The third study was done by Leovy and Mintz (1969) (see also Leovy, 1966b). It concentrates on the meteorology of the planet and uses essentially the methods of numerical weather prediction. Again, the condensation of CO_2 near the winter pole was found. The principal results are given in numerous charts and graphs that are not easily summarized. A careful study of the original paper is recommended.

D. Surface Topography

All the studies discussed above assume the planetary surface to be a smooth spheroid; large-scale topography could have a profound influence and has received considerable attention in the last few years. Two Earth-based methods have been applied: radar, and fine-scale mapping of the surface pressure. Both give comparable vertical and horizontal resolution (1 km and 500 to 1000 km), but the optical method gives much better coverage in latitude because the radar measurements are

limited to regions that pass through the sub-Earth point. Additional data have been obtained by Mariners 6 and 7, from the infrared and ultraviolet spectrometers, with much better horizontal resolution but rather incomplete coverage. The infrared instrument measures the equivalent width of the 2-μm bands of CO_2 and is thus more specific than the ultraviolet instrument, which measures the albedo at a short wavelength.

The radar results are limited so far to three latitude circles: +22°, +11°, and +6°. They show peak-to-peak variations of 10 to 15 km and find most of the structure to be of large scale, similar to terrestrial continents (Pettengill et al., 1969; Rogers et al., 1970). The optical method involves the use of some 15 exit slits, one for each of the lines shown in Figure 2, to give high sensitivity and therefore a short observing time (Belton and Hunten, 1969b; Wells, 1969). Instead of 3 nights for the spectrum shown, only 3 minutes are needed for a single measurement. The available coverage is from +50° to -40° with a small gap in longitude due to poor observing weather. The results are similar to those obtained with radar and agree also in finding little correlation of height with visible markings.

IV. VENUS

Visual inspection of Venus shows that it is totally covered with clouds; but it is difficult to be sure that small gaps are absent, and the night side is invisible. Thermal mapping [in the 8- to 14-μm region (Murray et al., 1963)] demonstrates conclusively that the cloud is continuous over the whole planet; otherwise, hot spots or areas would be present. For a surface temperature of 750 K, the Milne-Eddington relation (Equation 25) requires an infrared opacity of over 200 to reduce the effective temperature to 210 K. Though part of this can be contributed by the gas of the thick atmosphere, much of it must be due to the clouds. The observed radiation is produced at or near the cloud tops, except for radio wavelengths.

The thermal maps of Murray et al. (1963) (Figure 7) show several other interesting effects. No day-night change is visible; the principal effect is limb darkening (which implies an increase of temperature with depth, or an equivalent effect of scattering) and a small amount of polar cooling. Goody (1965) has carried out a careful analysis to isolate these three effects; the most surprising result is that the night side is actually a few degrees warmer than the day side. It appears that the thermal structure is controlled by circulation rather than radiation.

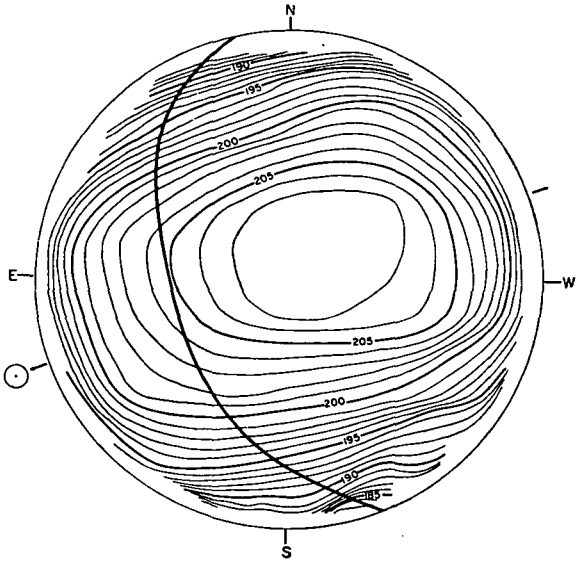


Figure 7.—Isotherms (K) for Venus in the 8- to 14-μm band (Murray et al., 1963). The terminator and the direction of the Sun are indicated.

The high surface temperature of 750 K is indicated by the radio brightness in the 10-cm region and is amply verified by measurements from spacecraft. A major problem is the origin of this high temperature. The greenhouse effect has received much study and strong advocacy, but it is difficult to imagine that most of the solar radiation can penetrate to the bottom of a medium whose infrared opacity is greater than 200. Again, we should probably have recourse to large-scale circulation. The second outstanding problem is the material of the clouds and how important they are in controlling the opacity.

A. Composition

In addition to Earth-based spectroscopy and radio occultation, we have direct chemical analysis of the Venusian atmosphere by the entry probes Veneras 4, 5, and 6. The best data are from Veneras 5 and 6, whose experiments were significantly more refined.

Carbon dioxide is the major gas, clearly established by Venera 4; the quoted concentration from Veneras 5 and 6 (Avduevsky et al., 1970a and 1970b) is 95 ± 2 percent. The method was to take a sample, absorb the CO_2 in potassium hydroxide, and measure the residual pressure, assumed to be due to N_2 or Ar. The result had been foreshadowed by spectroscopy (Belton et al., 1968b; Moroz, 1968), but ignorance of the optical properties of the cloud prevented the range from being narrowed to better than 20 to 100 percent. With the better spectral resolution now available, we could do better, but there is no reason to doubt the direct analysis.

Scattering in the clouds, discussed further below, complicates the interpretation of all spectroscopic measurements. If absorption lines of similar strength at nearby wavelengths can be compared, a relative abundance with respect to CO_2 can be obtained for another gas. If these precautions are not observed, the result can be seriously in error.

The observed CO_2 bands correspond roughly to the absorption in 10 to 20 km of gas at 0.1 b. Some regions of the spectrum are completely blanked out, and weak bands are to be found almost everywhere in the 1- to 3-μm region. Benedict has identified no fewer than 190 bands of seven isotopic forms; they include "hot" bands originating from the first excited vibrational level. The material for this work is the remarkable Connes Atlas (Connes et al., 1969), obtained by Fourier spectroscopy with a Michelson interferometer.

Carbon monoxide, though previously suspected in medium-resolution spectra by Sinton, was first clearly identified (Connes et al., 1968) in the Connes Atlas. Not only the principal isotopic form but also $C^{13}O^{16}$ and $C^{12}O^{18}$ are observed. The mixing ratio is $CO/CO_2 = 4.6 \times 10^{-5}$.

Hydrogen chloride and hydrogen fluoride are clearly present in the Connes Atlas (Connes et al., 1967); both isotopes of Cl can be seen. Comparison with a nearby isotopic CO_2 band, and also a hot band, gives a mixing ratio $HCl/CO_2 = 6 \times 10^{-7}$. This is not a small concentration if liquid water is present in the clouds; it implies that the drops would be strong hydrochloric acid, about 25 percent HCl by weight (Lewis, 1968 and 1969b). For HF, the mixing ratio is 5×10^{-9}.

For oxygen, the best limit is set by Earth-based spectroscopy at large Doppler shift. Belton and Hunten (1969a) studied Venus as well as Mars and set an upper limit $O_2/CO_2 \leqslant 8 \times 10^{-5}$. The experiment on Venera 4 gave a positive indication, but this claim was later withdrawn when the results from Veneras 5 and 6 were negative. In any case, it is difficult to imagine that the spectroscopic upper limit could be violated by almost two orders of magnitude.

Water vapor is highly controversial; the mixing ratios in the literature span almost four orders of magnitude. The Veneras made their measurements at a pressure level of about 0.6 b, and Venera 4 found about 0.1 to 0.7 percent H_2O. The Venera 5 and 6 results are given in preliminary fashion (Avduesky et al., 1970a and 1970b) as 4 to 11 mg/liter, which corresponds to 0.8 to 2.3 percent. The location is well below the visible cloud surface, which is at about 0.1 b.

Information on water vapor is also present in the microwave emission spectrum of the planet and from the absorption of radar signals. The transition from the low-temperature emission of the clouds to the high-temperature emission from the surface and lower atmosphere takes place in the wavelength region from 0.1 to 5 cm; the shape of the transition gives information on atmospheric opacity. The method is especially sensitive to H_2O, which has a strong line at 1.35 cm, very broad at high pressures. The best agreement is found (Pollack and Wood, 1968) with about 0.5 percent H_2O, and the upper limit is about 1 percent (Figure 8). This result disagrees with the upper part of the range reported for the direct measurements, but the brief report available may not be entirely accurate.

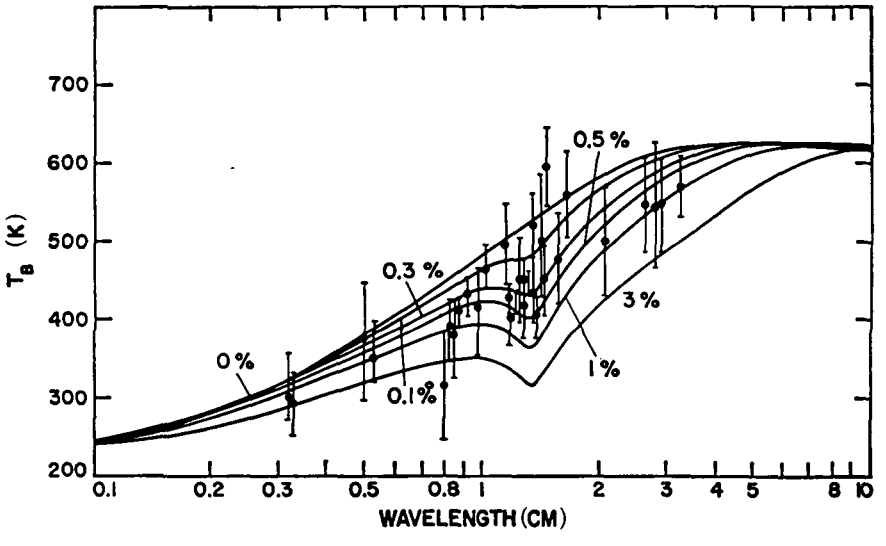


Figure 8.–Brightness temperature T_B as a function of wavelength for six values of the water-vapor mixing ratio in the lower atmosphere of Venus (Pollack and Wood, 1968). The points with error bars represent the collected observations.

Spectroscopic studies from Earth refer to a higher level, in the neighborhood of the very hazy "cloud tops". The results span a large range, but even the highest indicates much less moisture, a mixing ratio of about 10^{-4}. If both sets are correct, there must be a layer of ice or water cloud somewhere between; but as we discuss below, the visible clouds probably do not have this composition, and two different cloud layers seem to be required. Doppler-shifted spectra in the 8189Å region and balloon spectra at 1.13 μm show a weak but definite planetary line (Belton et al., 1968b; Spinrad and Shawl, 1966; Bottema et al., 1965) and give the mixing ratio of about 10^{-4}. A somewhat smaller upper limit was found in the Connes spectra (Connes et al., 1967) at longer wavelengths and by Owen (1967) at 8189Å. The conflict has been resolved by Schorn et al. (1969a), who have reported clear evidence for variability; they could not detect the 8189Å band in the spring of 1967, but it was clearly present the following winter. Aircraft spectra (Kuiper et al., 1969) are of much lower resolution but can work with stronger bands and also at longer wavelengths. The telluric absorption, though greatly reduced, has to be allowed for by separate spectra of the Moon; this procedure introduces additional uncertainty. Planetary absorption is definitely present, but the mixing ratio seems to be only a few parts in 10^6. This result does not take into account the scattering atmosphere or the possibility of weak CO_2 absorption in the neighborhood. Caution is therefore necessary until a more realistic analysis has been made. In any case, the evidence is clear that the H_2O is variable, and the highest mixing ratio seems to be about 10^{-4}.

B. Scattering in Clouds

In 1956, Chamberlain and Kuiper published a series of measurements of the CO_2 band at 8689Å over a range of phase angles ϕ (the angle between the Sun and the Earth, seen from Venus). On the average, light enters and leaves the atmosphere at zenith angles (θ_0, θ) equal to $\phi/2$. One would expect the absorption to vary as an airmass factor $\mu_0^{-1} + \mu^{-1} = 2 \sec(\phi/2)$, where $\mu = \cos\theta$. Instead, the opposite behavior is observed: The variation is approximated by $\mu_0 + \mu = 2\cos(\phi/2)$. The explanation put forward by Chamberlain and Kuiper, based on an earlier suggestion by van de Hulst, was that the absorption lines are formed primarily during a random traversal of the scattering medium, a mixture of gas and cloud particles. The variation as $\mu_0 + \mu$ is readily found in a simple model based on isotropic scattering, and this model has been explored in considerable detail (Belton et al., 1968b; Belton, 1968; Chamberlain, 1970; Chamberlain and Smith, 1970; McElroy, 1969a).

Many of the results can be transformed to be appropriate for realistic forward-scattering particles by the van de Hulst similarity relations, studied in detail by Hansen (1969). Additional similarities have been brought out by Belton et al. (1968b) and Chamberlain (1970). A prominent one is that even moderately weak lines absorb as a function of the product of pressure and amount of gas. Separation of these two variables is difficult and can only be made with very high spectral resolution at the centers of the lines. It is this difficulty that prevented a spectroscopic determination of the fraction of CO_2 in the Venusian atmosphere.

A special definition is required for the amount of gas in a scattering atmosphere because the path of the light is so complicated. The most useful quantity is the specific abundance, the amount of gas (in m-atm, say) in a scattering mean free path; it is a mixing ratio of gas with particles. The other important quantity in the theory is the single-scattering albedo of the particles, which controls the mean number of scatterings and therefore the path length of a photon (when gas absorption is not too strong). Thus, with very white particles, we expect many scatterings, a long path, and strong lines; with dark particles, there are fewer scatterings and weaker lines.

Unfortunately, it is very difficult to obtain the single-scattering albedo with enough accuracy to be useful. In the absence of enough information to define the problem completely, we must seek a way to use the partial information we do have. Fortunately, we can take advantage of the independent knowledge that the Venusian atmosphere is mostly CO_2, and we can use its lines as standards. Similarly, for the giant planets, the major gas is probably H_2. If we can compare lines at nearby wavelengths, the single-scattering albedo and the pressure should be nearly the same; if the lines are of comparable strength, the remaining complications cancel out, and a good mixing ratio can be obtained. This procedure is practical for Venus but much more difficult for the giant planets because H_2 lines are few, weak, and disturbed by pressure narrowing. Here one may be forced to give absolute amounts of a single gas, a notion that is unfortunately almost without meaning. On Venus, the lines of a single CO_2 band satisfy the restrictions almost ideally, and considerable faith can be put in the rotational temperatures derived from such bands.

C. Clouds

By far the most likely material for a cloud on Venus at observed temperatures is ice, simply by analogy with the Earth. But there are several lines of evidence that

make ice seem very improbable. After discussing this evidence, we shall turn to the difficult problem of finding a plausible alternative material.

We use the term *cloud tops* for the region observed spectroscopically, with the understanding that it is probably very hazy and many kilometers deep, with nothing that can be called a surface. From CO_2 line widths and the knowledge that CO_2 is the major gas, we find that the pressure is about 100 mb (Belton et al., 1968b). Excellent rotational temperatures have been obtained by a number of different groups on many different bands (Connes et al., 1968; Young et al., 1969); the result is 240 K, with error limits ±5 K. Such temperatures are essentially free of systematic errors and can be accepted with confidence as true gas temperatures. The vapor pressure of ice at 240 K is 0.3 mb; if the cloud particles are ice, the gas between them must contain water vapor at this partial pressure, or a mixing ratio of 3×10^{-3}. Even the largest observed amount, 10^{-4}, disagrees by a factor of 30, and the discrepancy is worse for many of the measurements. A mixing ratio of 10^{-4} could be obtained at 213 K, which is ruled out by the temperature observations. The above numbers are modified slightly by the presence of HCl, but as discussed by Lewis (1969b), the conclusion is unchanged.

The polarization of the light scattered by Venus has been observed by Coffeen and Gehrels (1969) and analyzed in detail by Coffeen (1969). With a somewhat intuitive allowance for multiple scattering, he found a refractive index in the range of 1.43 to 1.55 and a mean diameter of 2.5 μm. The calculations have recently been greatly refined by Hansen (Hansen and Arking, 1971), who finds the refractive index to be 1.45 ± 0.01 in the visible, with slight normal dispersion, and a mean diameter of 2.2 μm. The presence of features ascribed to the glory and rainbow strongly indicates the particles to be spherical. Ice, with its refractive index of 1.31, seems to be excluded quite apart from its nonspherical crystals.

Plotting the albedo of Venus as a function of wavelength yields information on the cloud, after allowance has been made for absorption by the gas. In the ultraviolet, Anderson et al. (1969) have given a detailed analysis of their rocket observations on a scattering model, which includes scattering and absorption by CO_2 and HCl. They find the single-scattering albedo of the particles to vary smoothly from 0.9 at 3500Å to 0.4 at 2000Å. Again, this is unlike the behavior of ice, unless its absorption is due to dissolved substances, which would include HCl. For the visible and infrared, the planetary albedo has been compiled by Kuiper (1969) and compared with a wide range of powdered solids. There are several absorption bands that do agree fairly well with ice, as discussed in some detail by Hansen and Cheyney (1968a and 1968b). The weakness of the bands can be explained by a small particle

size. Pollack and Sagan (1968) and Plummer (1970) have strongly advocated the identification as ice from this evidence, but this identification ignores the contrary spectroscopic and polarimetric data.

Many other candidates have been suggested, but as Lewis (1969b) points out, most of these are incompatible with the presence of HCl. His own geochemical study favors mercury and several of its compounds, in a variety of different layers. He suggests Hg_2Cl_2 for the topmost layer, with metallic liquid Hg the next layer below. Unfortunately, the refractive index of Hg_2Cl_2 is 2.3 (average for the two rays), and the material shows no infrared absorptions up to 2.5 μm (Kuiper, 1969). It does absorb in the ultraviolet, and might fit the observations there, but it seems too good a reflector in the blue. Earlier, Lewis (1968) suggested NH_4Cl, which has attractive features, but his geochemical study (1969b) gave far too small an abundance of NH_3; this does not seem a very strong argument. The infrared spectrum matches fairly well; the substance does not absorb in the ultraviolet but could be yellowed by the strong solar irradiation on Venus. The refractive index is 1.64. The use of $FeCl_2 \cdot 2H_2O$ gives a rather good fit all the way from 0.3 to 3 μm and is strongly advocated by Kuiper (1969). Further detailed studies of its geochemistry and optical properties are needed.

Rasool (1970) has noted several anomalous peaks in the attenuation data from Mariner 5. It is possible to work out a plausible correlation with the multiple cloud layers predicted in Lewis' model based on Hg and its compounds.

It should be noted that none of the above suggestions is really consistent with Hansen's interpretation of the polarimetric data. The C_3O_2 (carbon suboxide) polymer has been put forward repeatedly since the original qualitative suggestion by Harteck; Hansen points out that its refractive index and possible spherical shape are both appropriate. The real difficulty here is that too little is known about the production (presumably photochemical) and stability of the substance.

The water vapor measured by the Veneras should give an ice cloud with a bottom near 200 to 300 mb (Avduevsky et al., 1970a and 1970b). If present, this cloud must presumably be hidden from view by a layer of some other composition. Alternatively, as Lewis (1969b) suggests, the operation of the H_2O detector, based on electrical conductivity, may have been upset by the presence of HCl.

D. Atmospheric Structure

Temperature and pressure profiles are available from the occultation experiment on Mariner 5 and the direct probing of Veneras 4, 5, and 6. In no case do the

measurements reach the surface, and the necessary extrapolation has caused a good deal of controversy. An important element in reaching an agreed solution has been the planetary radius of 6053 km, obtained by Earth-based radar; the history of this measurement is discussed in a lively note by Smith (1970). Extrapolation of the spacecraft measurements to the surface gives a temperature of about 750 K, as inferred from the radio brightness, and confirms both results. The corresponding surface pressure is about 100 b. A diurnal variation of even 1 K in 117 days is impossible at such pressures (Hunten and Goody, 1969). An upper limit to the latitude variation of the surface temperature has been set by radio interferometry (Sinclair et al., 1970): The poles are less than 12 K cooler than the equator. Thus, the surface is essentially isothermal.

The Mariner 5 occultation data (Kliore et al., 1967) extend from 80 km, 10 mb, 240 K to 40 km, 5 b, 450 K (where the temperatures are for a composition of 90 CO_2:10 N_2). The lower limit of the range is set by critical refraction, which bends tangential rays down to the surface. Below about 70 km, the lapse rate is about 9 K-km^{-1}, close to adiabatic; from 70 to 80 km, it is much smaller, 0 to 2 K-km^{-1}. The appearance is that of a tropopause at about 70 km, which is also the area of the "cloud tops". Because the phase shifts are large, these data are much firmer than the corresponding ones for Mars.

Summaries of the Venera 4, 5, and 6 data, which include the overlapping Mariner 5 results, are given by Avduevsky et al. (1970a and 1970b) (see Figure 9) and Eshleman (1970). The Veneras carried radar altimeters to indicate the remaining distance to the surface. Unfortunately, the readings are in strong disagreement; but the average of Veneras 5 and 6 agrees well with the Earth-based radar radius. Surface conditions are thus indicated to be near 770 K and 100 b, and the lapse rate continues to be adiabatic all the way down. A small subadiabatic region near the surface is possible, but it cannot be deep if the thermal emission of Venus is to be reproduced.

In 1959, observations were obtained of a natural occultation by Venus of the bright star Regulus. In principle, this measurement contains as much information as a radio occultation, but in practice, there is a severe limitation in signal-to-noise ratio. Hunten and McElroy (1968) argue that the scale height can be obtained to little better than a factor of 2, and any attempt to find a gradient is illusory. However, a useful density is obtained at a known height from the center of the planet. The effect observed is differential refraction (discussed following Equation 18), and it is found that the received intensity is reduced by a factor of 1/2 when the refraction α equals the angle subtended by a scale height at the distance of the

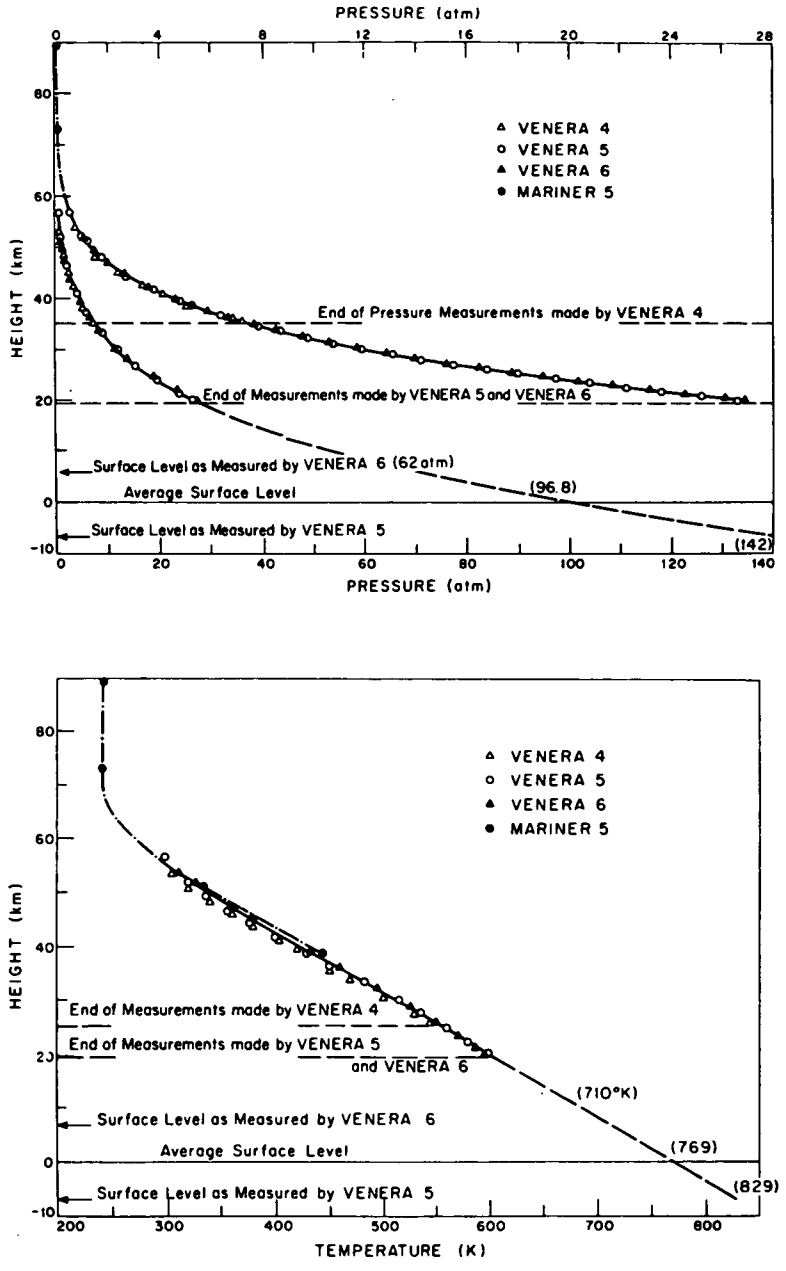


Figure 9.–Pressure and temperature profiles for Venus (Avduevsky et al., 1970a and 1970b).

Earth. For the conditions of the Regulus occultation, this angle is only 0.012 arc seconds. The density derived depends on $H^{3/2}$, and H is estimated (Hunten and McElroy, 1968; McElroy, 1969) as 4.4 km (for a temperature of 200 K). The result is a CO_2 number density of 3.2×10^{13} cm^{-3} at a radial distance of 6169 km, or a height of 116 km. The thermal structure of this region has been calculated by McElroy (1969b) for a condition of radiative equilibrium; he obtains a lapse rate of 2.0 $K\text{-}km^{-1}$ from 60 to 110 km, which fits the Mariner data well where they overlap.

E. Surface Temperature and Circulation

We now return to the fundamental question of why the surface temperature of Venus is so high. As we show in the introduction to this section, the opacity in the thermal infrared must be at least 200. In itself, this is not unreasonable for a cloudy atmosphere composed of gases with strong infrared absorption bands. Pressure broadening of these bands, and possible pressure-induced transitions, can produce heavy absorption over wide wavelength regions, but it is hard to imagine all the solar flux penetrating to the surface through such a medium. With less flux to the bottom, still more opacity is needed, and we have still not allowed for the heat transferred by convection. Some of these questions are considered in great detail by Sagan and Pollack (1969), but the case made for the greenhouse effect is still not convincing (Hunten and Goody, 1969). A considerable greenhouse effect can be obtained from the properties of forward-scattering cloud particles, but Samuelson (1967) is unable to reach a temperature greater than 480 K, which is still without allowance for convection.

One can imagine the temperature being supported by internal heat from the planet. On Earth, this heat is less than 0.1 percent of the solar flux absorbed by Venus. Even if we postulate a much larger flux, we run into difficulties with convection; and if the flux exceeds the solar heating, the extra energy would be observed in the thermal emission.

The most successful models envision a large-scale atmospheric circulation driven from the top, where solar radiation is most likely to be absorbed. Such a model has been worked out in detail by Goody and Robinson (1966), who find, at high levels, motion away from the subsolar point and towards the antisolar point. There is a slow, deep return circulation, and the temperature gradient is produced by adiabatic compression. The opacity of the atmosphere is required to be very large so that little heat is transferred by radiation: The mere 200 found above would be far

too small. It may be that the solar radiation is absorbed at a level low enough that diurnal variations (in 117 days) are small; the temperature contrast would then be between equator and poles. The model still applies, but now with two convective cells, one in each hemisphere; the motion is poleward at high levels and equatorward at low levels.

Another class of models has been inspired by the observational evidence for a retrograde atmospheric motion with a period of about 4 days. This evidence comes both from low-contrast cloud markings (Boyer and Guerin, 1969) and marginal measurements of Doppler shifts at the limb (Guinot and Feissel, 1968). It has been argued (Schubert and Whitehead, 1969; Schubert and Young, 1970; Malkus, 1970) that such a circulation is likely to occur in a CO_2 atmosphere rotating as slowly as that of Venus. Again, it is supposed, though only in general terms, that an opaque lower atmosphere would be set into sufficient motion to maintain an adiabatic lapse rate. If the cloud level does indeed rotate fairly uniformly in 4 days, the same should be true of all higher levels, despite the long period of planetary rotation. In a 58-day night, there is time for many photochemical equilibria to be drastically altered, an example being the concentration of atomic hydrogen. In a mere 2 days, one must expect much smaller diurnal effects in the upper atmosphere.

V. JUPITER

Again, on Jupiter, we find an atmosphere dominated by cloud, though the cover is not as complete as on Venus: Infrared hot spots are indeed observed in the dark North Equatorial Belt. There is no evidence either way on the presence of a solid surface, except from the radiation balance, which suggests the emission of substantial planetary heat approximately equal to the absorbed solar flux, and, therefore, perhaps a convective interior. Models of the planetary interior favor strongly the presence of convection all the way to the center. At any rate, any bottom to the atmosphere is far below the region we consider here. As a consequence, the zero point of a height scale is arbitrary, and there are almost as many choices as authors. Until there is an agreed definition, one must simply be aware of the problem.

As for Venus, spectroscopic data must be interpreted on a scattering model; but there are no convenient CO_2 bands spread through the spectrum to serve as standards, only a few exceedingly weak and narrow forbidden lines of H_2. Thus,

comparisons under the restrictions discussed for Venus are not readily made, and we remain uncertain of the exact composition. Finally, the low surface brightness of a planet at 5.2 AU adds to the difficulty of spectroscopic observations.

A. Composition

There is no real doubt that the main constituent is molecular hydrogen. By symmetry, this molecule has no normal (dipole) absorption in the infrared, but it is so abundant that quadrupole lines, and some pressure-induced dipole lines, are present. The latter have been observed by Stratoscope II, and Danielson (1966) has found that the broad, deep absorption from 2 to 2.5 μm can be explained by about 20 km-atm of H_2. Quadrupole lines are slightly more tractable (though they too give problems), and they have the advantage of being observable from the ground. The selection rule for rotation J is $\Delta J = +2, 0, -2$, giving O, Q, S branches instead of the P, Q, R found for dipole transitions.

Though the quadrupole lines are readily observed, their interpretation is complicated by collisional (or pressure) narrowing, which is also present in the Raman and radio lines of H_2. Ordinary pressure broadening occurs when collisions interrupt the phase of a radiating wave; but for certain special cases, the interruption occurs in only about 1 percent of the collisions. At suitable pressures, the radiation is thus coherently averaged over about 100 free paths of the molecule, and the appropriate Doppler velocity is averaged in the same way. As the pressure is raised from a low value, we have first a normal Doppler line, then a narrowed line whose shape is approximately Lorentzian, and finally ordinary broadening, though only 1 percent as strong as normal. More details are given by Fink and Belton (1969), who also derive curves of growth for both clear and scattering atmospheres. It may readily be imagined that a scattering atmosphere with narrowing offers serious problems of interpretation, and that is indeed the case. The best recourse is to use lines weak enough that the exact shape and width of the profile is not important. Unfortunately, such weak lines are the ones that are hardest to observe accurately.

Fink and Belton (1969) have obtained photoelectric observations of two lines of the 3-0 band around 8150Å, the $S(0)$ and $S(1)$, and an upper limit for the $S(2)$. The observed lines are probably saturated. The weak 4-0 band is represented by a single line, the $S(1)$ at 6367Å (Figure 10). Other (photographic) observations are given by Owen and Mason (1968). On a clear-atmosphere model, H_2 abundances are found to be 67 ± 17 and 87 km-atm by the two groups; Fink and Belton obtain a

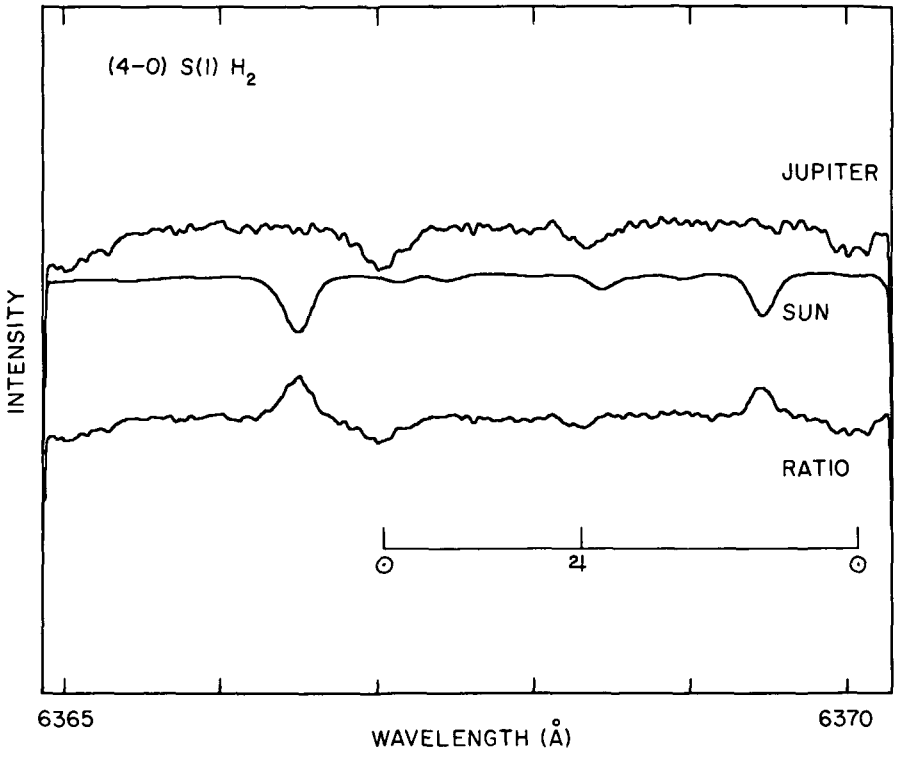


Figure 10.—The $S(1)$ line of the 4-0 quadrupole band of H_2 on Jupiter, most clearly seen in the ratio spectrum (Fink and Belton, 1969).

rotational temperature of 145 ± 20 K. For a scattering model, they find a specific abundance of 17 km-atm per mean free path.

Because we know very little about the properties of the clouds, we are forced to do what we can with the results for the clear atmosphere. The mean abundance of 75 km-atm corresponds to an H_2 partial pressure of 1.8 atm at the bottom of the column. Such a depth is certainly the farthest we can hope to see into the atmosphere in the near infrared; the depth is even less for the actual scattering atmosphere.

Helium is surely present, but it has no useful absorptions. The solar abundance corresponds to $He/H_2 = 0.11$ by number and is not ruled out. An upper limit is set by the pressure broadening of CH_4 lines, which suggests a total pressure less than 2.3

atm or a helium partial pressure of not more than 0.5 atm; because of the inherent uncertainties, the upper limit on He/H_2 should probably be taken as 1. Further discussion is given in McElroy's review (McElroy, 1969) of atmospheric composition of the Jovian planets. He also discusses in some detail a stellar occultation (Baum and Code, 1953) which appears to require $He > H_2$ for any reasonable temperature. As for Venus, the data can be fitted by a wide range of scale heights, and reasonable error limits are consistent even with pure H_2. The spectroscopic evidence should be given the greater weight.

The interior structure of Jupiter has been studied by DeMarcus (1958) and Peebles (1964), with the constraints set by the mean density and the mass distribution. Peebles concludes that the interior is probably convective and fully mixed all the way (or very nearly) to the center and that the He/H_2 ratio is about 1/8 by number. This result, which is close to the solar ratio, should be used until better information becomes available.

Methane is represented by many strong bands, even extending into the visible. Analysis is hampered by the fact that the structure of most of the bands is not understood. Based on a clear-atmosphere interpretation, the relative abundance to H_2 is found to be about 10^{-3}, but the uncertainty could be large.

Ammonia bands also appear at many wavelengths. This gas almost certainly condenses in the atmosphere of Jupiter and forms the visible clouds. The mixing ratio is thus a strong function of height, and it is not surprising to find different abundances at different wavelengths that penetrate to different depths. The near infrared gives 12 m-atm (Fink and Belton, 1969), whereas the ultraviolet near 2000Å gives only 2×10^{-3} cm-atm as an upper limit (Anderson et al., 1969). It seems unlikely that so great a difference could be caused by scattering in a gas; a likelier explanation is the existence of NH_3 "cirrus" at great heights.

For many other gases, upper limits have been set by Gillett et al. (1969). Fink and Belton (1969) find less than 40 μm precipitable H_2O, which would not be expected at 145 K but could well form another cloud layer at a greater depth.

B. Clouds

There is every reason to believe that the observed cloud deck is ammonia ice, undoubtedly very hazy toward the top. Absorptions are found to weaken slightly toward the limb, somewhat as expected for a scattering model and certainly not as expected for a clear atmosphere. The dark belts might be thought to be clearer, but

usually no temperature contrast can be observed. However, Gillett et al. (1969) have found the brightness temperature to be 225 K in the 5-μm region, where none of the known gases absorb strongly. Recently, Westphal (1969) observed indeed, in this wavelength band, localized hot spots within the North Equatorial Belt (5° to 20° latitude); there was also one such spot in another dark region. The most likely explanation by far is holes in the ammonia cloud which permit radiation to escape from a lower, hotter region where the opacity might be due to gas or to another cloud deck. The observed brightness temperature reaches 310 K; the actual temperature could be even higher if the spots are smaller than the field of view.

Lewis (1969a) has performed a study of other expected cloud layers, based on two assumed compositions. Below the ammonia ice layer, he finds the principal cloud to be an ammonia-water solution, perhaps solid in the upper parts. Above this, a layer of water ice may or may not be present. Reaction of H_2S and NH_3 produces NH_4SH, ammonium hydrosulfide, which could form another layer above the ice. One of these models is illustrated in Figure 11.

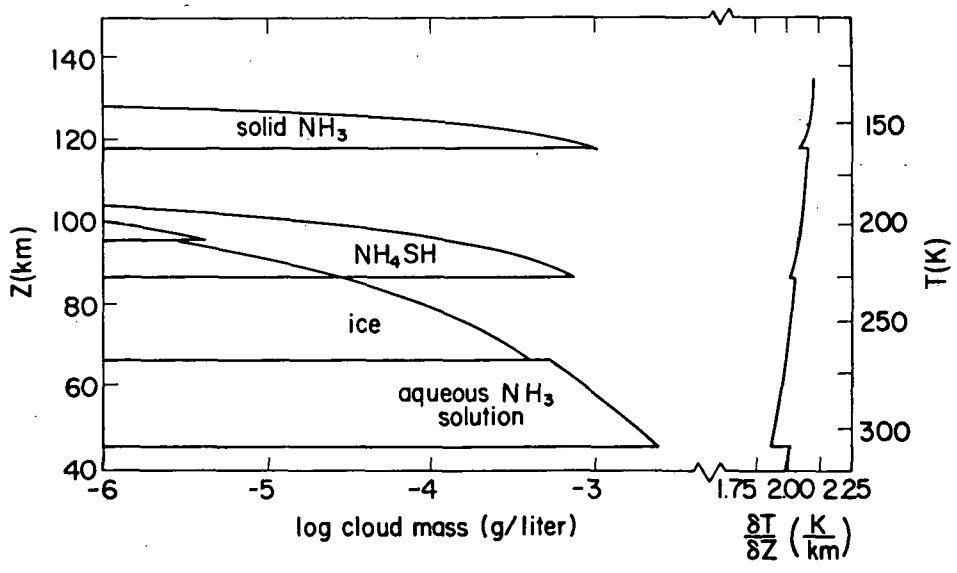


Figure 11.–Cloud masses and calculated lapse rates for clouds on Jupiter in a model with solar composition (Lewis, 1969a).

C. Thermal Structure

We shall restrict ourselves to the general region visible optically, down to a pressure of a few bars and a temperature of a few hundred kelvins. The fundamental work in the cloud region is that of Trafton (1967), also discussed by Trafton and Münch (1969). At the low temperatures of Jupiter, long wavelengths are important, and much of the opacity above 10 μm is contributed by pressure-induced dipole absorption (rotational and translational) in H_2. A convectively unstable region was found for pressures greater than 0.7 b, and the adiabatic lapse rate was assumed for greater pressures. Emerging flux was represented by the effective temperature T_e, which was varied from 110 to 130 K. The stratospheric temperature was found to be about 100 K.

A complementary study for the region above the clouds has been made by Hogan et al. (1969). They assumed a variety of cloud-top pressures and temperatures, in the range of 2 to 6 and 210 to 230 K, treating the cloud surface as a blackbody radiator. After inserting the opacities of H_2, CH_4, and NH_3, they calculated the temperature profiles for radiative equilibrium, with the usual allowance for convection. Solar heating was included; it is particularly important in the 3.3-μm band of CH_4 and gives a temperature inversion similar to that produced on Earth by ozone. The most successful model (No. 3) has a minimum temperature of 115 K at the 170-mb level and becomes constant, 140 K, at 20 mb.* Infrared and radio emissions were predicted and compared with observation; the fit is rather good. Figure 12 shows the comparison with the infrared spectrum of Gillett et al. (1969). We have already noted the "window" which permits 5-μm radiation to escape from hotter regions deep in the atmosphere, quite apart from Westphal's hot spots. A small peak at 7.7 μm is fitted by the models with a temperature inversion and confirms this feature nicely; the opacity of methane at this wavelength is great enough to put the effective radiating level above the temperature minimum.

Both Trafton and Hogan et al. confirm the necessity for a planetary heat flow roughly equal to the solar heating; the effective temperature for the model just discussed is 127 K (compare Table 1). They do not support the possibility that the radiation beyond 14 μm might be weak enough to restore a balance with solar heating alone. In any case, Aumann et al. (1969) (see also McElroy, 1969a, and Hogan et al., 1969) have measured the radiation over the whole range of 1.5 to 350 μm and find directly T_e = 134 K.

*See Figures 9 and 10, Chapter 12, "Evolution of Planetary Atmospheres", by Rasool.

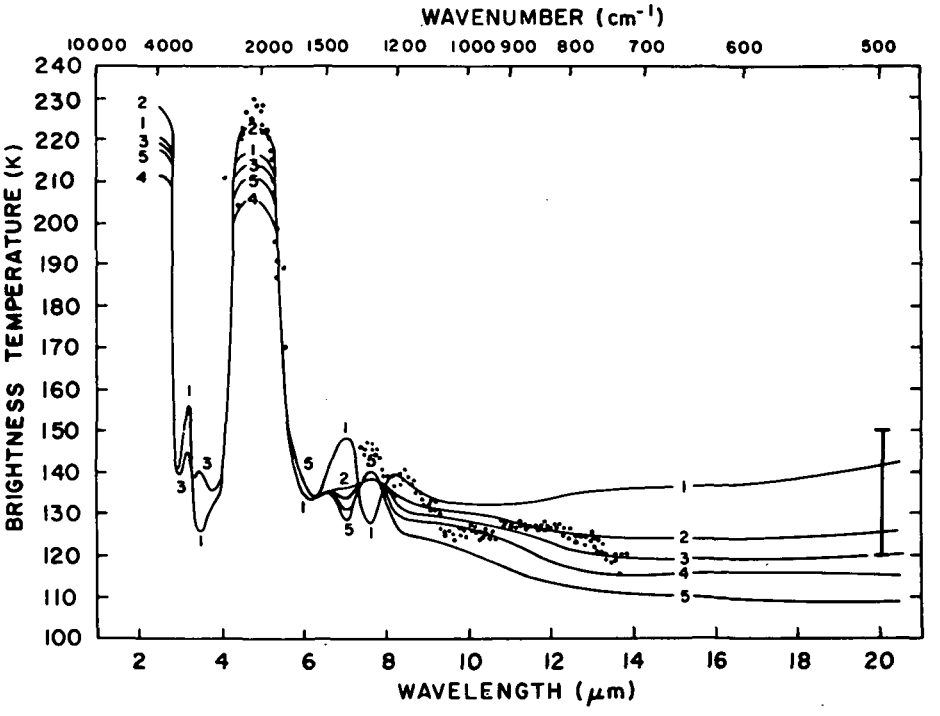


Figure 12.—Computed and observed infrared brightness temperatures of Jupiter. The dots correspond to the measurements of Gillett et al. (1969); calculations are by Hogan et al. (1969).

We should repeat here the words of caution from Section II on radiative-balance calculations. Gierasch and Goody (1969) have attempted to estimate the radiative and dynamical time constants in the region of the cloud tops and find the latter to be much shorter. If large-scale motions do indeed control the thermal profile, the calculations discussed here are not really applicable. We may hope that they approximate the truth (as they do on Earth), but we cannot prove it.

VI. OTHER PLANETS

A. Saturn, Uranus, Neptune

It is likely that the general structures of the atmospheres of Saturn, Uranus, and Neptune resemble that of Jupiter, with allowance for different gravity and increasing distance from the Sun. The lower temperatures cause ammonia to freeze

out at deeper levels; its absorption is probably absent on Saturn, and certainly on Uranus and Neptune. On the other hand, the methane absorptions become stronger, presumably because clearer atmospheres permit deeper penetration of light, and perhaps also, on Uranus and Neptune, because these planets are richer in heavier elements, as their mean densities suggest. The pressure-induced H_2 absorptions should be stronger in a clearer atmosphere, and the far wings of the infrared bands may encroach upon the visible; H_2 by itself could possibly explain the green visual appearance of Uranus and Neptune. A review is given by McElroy (1969a).

B. Mercury

Tests for CO_2 absorption on Mercury have been made by Belton et al. (1967) and Bergstralh et al. (1967); the upper limit is 0.04 mb. Thermal calculations (Belton et al., 1967) suggest that gases even as heavy as CO_2 and Ar will escape from an atmosphere of normal structure, and at most, a thin atmospheric remnant is to be expected. At a surface pressure of about 2×10^{-6} mb, the exosphere is in contact with the surface and should assume its temperature, which is low enough to prevent escape, perhaps even of O and H_2O. Such an atmosphere, containing at least CO_2, CO from photodissociation, and Ar, is thus entirely possible. It is still substantial enough to support an ionosphere and a rich dayglow spectrum but offers no obstacle to meteorities and solar-wind protons.

REFERENCES

Allen, C. W., *Astrophysical Quantities,* Athlone Press, London, 1963.
Anderson, R. C., Pipes, J. G., Broadfoot, A. L., and Wallace, L., *J. Atmos. Sci.* 26:874, 1969.
Aumann, H. H., Gillespie, C. M., and Low, F. J., *Astrophys. J. (Letters)* 157:L69, 1969.
Avduevsky, V. S., Marov, M. Ya., and Rozhdestvensky, M. K., *J. Atmos. Sci.* 27:561, 1970a.
Avduevsky, V. S., Marov, M. Ya., and Rozhdestvensky, M. K., *Radio Sci.* 5:333, 1970b.
Barth, C. A., Fastie, W. G., Hord, C. W., Pearce, J. B., Kelly, K. K., Stewart, A. I., Thomas, G. E., Anderson, G. P., and Raper, O. F., *Science* 165:1004, 1969.
Baum, W. A., and Code, A. D., *Astron. J.* 58:108, 1953.
Belton, M. J. S., Hunten, D. M., and McElroy, M. B., *Astrophys. J.* 150:1111, 1967.
Belton, M. J. S., *J. Atmos. Sci.* 25:596, 1968.
Belton, M. J. S., Broadfoot, A. L., and Hunten, D. M., *J. Geophys. Res.* 73:4795, 1968a.
Belton, M. J. S., and Hunten, D. M., *Astrophys. J.* 153:963, 1968.
Belton, M. J. S., Hunten, D. M., and Goody, R. M., *The Atmospheres of Venus and Mars,* J. C. Brandt and M. B. McElroy, eds., Gordon and Breach, New York, 1968b, p. 69.

Belton, M. J. S., and Hunten, D. M., *Astrophys. J.* 156:797, 1969a.
Belton, M. J. S., and Hunten, D. M., *Science* 166:225, 1969b.
Bergstralh, J. T., Gray, L. D., and Smith, H. J., *Astrophys. J.* 149:L137, 1967.
Bottema, M., Plummer, W., and Strong, J., *Ann. Astrophys.* 28:225, 1965.
Boyer, C., and Guerin, P., *Icarus* 11:338, 1969.
Broadfoot, L., and Wallace, L., *Astrophys. J.* 161:303, 1970.
Carleton, N. P., Sharma, A., Goody, R. M., Liller, W. L., and Roesler, F. L., *Astrophys. J.* 155:323, 1969.
Chamberlain, J. W., and Hunten, D. M., *Rev. Geophys.* 3:299, 1965.
Chamberlain, J. W., *Astrophys. J.* 159:137, 1970.
Chamberlain, J. W., and Smith, G., *Astrophys. J.* 160:755, 1970.
Coffeen, D. L., *Astron. J.* 74:446, 1969.
Coffeen, D. L., and Gehrels, T., *Astron. J.* 74:433, 1969.
Connes, P., Connes, J., Benedict, W. S., and Kaplan, L. D., *Astrophys. J.* 147:1230, 1967.
Connes, P., Connes J., Kaplan, L. D., and Benedict, W. S., *Astrophys. J.* 152:731, 1968.
Connes, P., Connes, J., and Maillard, J. P., *Atlas des spectres infrarouges de Venus, Mars, Jupiter, et Saturne*, Editions du Centre National de la Recherche Scientifique, Paris, 1969.
Dalgarno, A., and McElroy, M. B., *Science* 170:167, 1970.
Danielson, R. E., *Astrophys. J.* 143:949, 1966.
DeMarcus, W. C., *Astron. J.* 63:2, 1958.
Eshleman, V. R., *Radio Sci.* 5:325, 1970.
Fink, U., and Belton, M. J. S., *J. Atmos. Sci.* 26:952, 1969.
Fjeldbo, G., and Eshleman, V. R., *Planet. Space Sci.* 16:1035, 1968.
Gierasch, P., and Goody, R., *Planet. Space Sci.* 16:615, 1968.
Gierasch, P. J., and Goody, R. M., *J. Atmos. Sci.* 26:979, 1969.
Gifford, F., *Astrophys. J.* 123:154, 1956.
Gillett, F. C., Low, F. J., and Stein, W. A., *Astrophys. J.* 157:925, 1969.
Giver, L. P., Inn, E. C. Y., Miller, J. H., and Boese, R. W., *Astrophys. J.* 153:285, 1968.
Goody, R. M., *Atmospheric Radiation. I. Theoretical Basis*, Clarendon Press, Oxford, 1964.
Goody, R. M., *J. Geophys. Res.* 70:5471, 1965.
Goody, R. M., and Robinson, A. R., *Astrophys. J.* 146:339, 1966.
Goody, R. M., *Annual Reviews of Astronomy and Astrophysics*, Annual Reviews, Inc., Palo Alto, California, 1969, Vol. 7.
Guinot, B., and Feissel, M., *J. des Observateurs* 51:13, 1968.
Hansen, J. E., and Cheyney, H., *J. Atmos. Sci.* 25:629, 1968a.
Hansen, J. E., and Cheyney, H., *J. Geophys. Res.* 73:6136, 1968b.
Hansen, J. E., *Astrophys. J.* 158:337, 1969.
Hansen J. E., and Arking, A., *Science* 171:669, 1971.
Hogan, J. S., Rasool, S. I., and Encrenaz, T., *J. Atmos. Sci.* 26:898, 1969.
Hunten, D. M., and McElroy, M. B., *J. Geophys. Res.* 73:4446, 1968.
Hunten, D. M., and Goody, R. M., *Science* 165:1317, 1969.
Hunten, D. M., and McElroy, M. B., *J. Geophys. Res.* 75, 1970.
Kaplan, L. D., Münch, G., and Spinrad, H., *Astrophys. J.* 139:1, 1964.

Kaplan, L. D., Connes, J., and Connes, P., *Astrophys. J.* 157:L187, 1969.
Kieffer, H., *J. Geophys. Res.* 75:501 and 75:510, 1970.
King-Hele, D. G., *Nature* 183:1224, 1959.
Kliore, A., Cain, D. L., Levy, G. S., Eshleman, V. R., Fjeldbo, G., and Drake, F. D., *Science* 149:1243, 1965.
Kliore, A., Levy, G. S., Cain, D. L., Fjeldbo, G., and Rasool, S. I., *Science* 158:1683, 1967.
Kliore, A., Fjeldbo, G., Seidel, B. L., and Rasool, S. I., *Science* 166:1393, 1969.
Kondratyev, K. Ya., *Radiation in the Atmosphere,* Academic Press, New York, 1969.
Kuiper, G. P., *Comm. Lunar Planet. Lab (Univ. of Arizona),* No. 101, 1969.
Kuiper, G. P., Forbes, F. F., Steinmetz, D. L., and Mitchell, R. I., *Comm. Lunar Planet. Lab (Univ. of Arizona),* No. 100, 1969.
Leighton, R. B., and Murray, B. C., *Science* 153:136, 1966.
Leovy, C., *Icarus* 5:1, 1966a.
Leovy, C., *Science* 154:1178, 1966b.
Leovy, C., and Mintz, Y., *J. Atmos. Sci.* 26:1167, 1969.
Lewis, J. S., *Astrophys. J. (Letters)* 152:L79, 1968.
Lewis, J. S., *Icarus* 10:365, 1969a.
Lewis, J. S., *Icarus* 11:367, 1969b.
McElroy, M. B., *J. Atmos. Sci.* 26:798, 1969a.
McElroy, M. B., *J. Geophys. Res.* 74:29, 1969b.
McElroy, M. B., and Hunten, D. M., *J. Geophys. Res.* 75:1188, 1970.
Malkus, W. V. R., *J. Atmos. Sci.* 27:529, 1970.
Moroz, V. I., *Soviet Astron. A. J.* 11:653, 1968.
Morrison, D., Sagan, C., and Pollack, J. B., *Icarus* 11:36, 1969.
Murray, B. C., Wildey, R. L., and Westphal, J. A., *J. Geophys. Res.* 68:4813, 1963.
Neugebauer, G., Münch, G., Chase, S. C., Jr., Hatzenbeler, H., Miner, E., and Schofield, D., *Science* 166:98, 1969.
Owen, T., *Astrophys. J. (Letters)* 150:L121, 1967.
Owen, T., and Mason, H. P., *Astrophys. J.* 154:317, 1968.
Owen, T., and Mason, H. P., *Science* 165:893, 1969.
Peebles, P. J. E., *Astrophys. J.* 140:328, 1964.
Pettengill, G. H., Counselman, C. C., Rainville, L. P., and Shapiro, I. I., *Astron. J.* 74:461, 1969.
Plummer, W. T., *Icarus* 12:233, 1970.
Pollack, J. B., and Sagan, C., *J. Geophys. Res.* 73:5943, 1968.
Pollack, J. P., and Wood, A. T., Jr., *Science* 161:1125, 1968.
Rasool, S. I., *Radio Sci.* 5:367, 1970.
Rasool, S. I., Hogan, J. S., Stewart, R. W., and Russell, L. H., *J. Atmos. Sci.* 27:841, 1970.
Rogers, A. E. E., Ash, M. E., Counselman, C. C., Shapiro, I. I., and Pettengill, G. H., *Radio Sci.* 5:465, 1970.
Rubey, W. W., *Bull. Geol. Soc. Amer.* 62:1111, 1951 (reprinted in *The Origin and Evolution of Atmospheres and Oceans,* P. J. Brancazio and A. G. W. Cameron, eds., John Wiley & Sons, New York, 1964).
Sagan, C., and Pollack, J. B., *Icarus* 10:274, 10:290, 10:301, and 10:314, 1969.

Samuelson, R. E., *Astrophys. J.* 147:782, 1967.
Samuelson, R. E., *J. Atmos. Sci.* 25:634, 1968.
Schorn, R. A., Spinrad, H., Moore, R. C., Smith, H. J., and Giver, L. P., *Astrophys. J.* 147:743, 1967.
Schorn, R. A., Barker, E. S., Gray, L. D., and Moore, R. C., *Icarus* 10:98, 1969a.
Schorn, R. A., Farmer, C. B., and Little, S. J., *Icarus* 11:283, 1969b.
Schubert, G., and Whitehead, J., *Science* 163:71, 1969.
Schubert, G., and Young, R. E., *J. Atmos. Sci.* 27:523, 1970.
Schwarzschild, M., *Structure and Evolution of the Stars,* Princeton University Press, Princeton, New Jersey, 1958.
Sinclair, A. C. E., Basart, J. P., Buhl, D., Gale, W. A., and Liwschitz, M., *Radio Sci.* 5:347, 1970.
Sinton, W. M., and Strong, J., *Astrophys. J.* 131:459, 1960.
Smith, W. B., *Science* 169:1001, 1970.
Spinrad, H., and Shawl, S. J., *Astrophys. J.* 146:328, 1966.
Trafton, L. M., *Astrophys. J.* 147:765, 1967.
Trafton, L. M., and Münch, G., *J. Atmos. Sci.* 26:813, 1969.
Wells, R. A., *Science* 166:862, 1969.
Westphal, J. A., *Astrophys. J.* 157:L63, 1969.
Young, L. D. G., Schorn, R. A., Barker, E. S., and MacFarlane, M., *Icarus* 11:390, 1969.

BIBLIOGRAPHY

The Atmospheres of Mars and Venus, J. C. Brandt and M. B. McElroy, eds., Gordon and Breach, New York, 1968.
Icarus 10(3):353-421, 1969.
J. Atmos. Sci. 25(4):533-671, 1968.
J. Atmos. Sci. 26(5):795-1001, 1969.
J. Atmos. Sci. 27(4):523-560, 1970.
Radio Sci. 5(2):121-533, 1970.

Page intentionally left blank

CHAPTER 7
THE COMPOSITION OF PLANETARY ATMOSPHERES

Tobias Owen
Department of Earth and Space Sciences
State University of New York at Stony Brook

I. INTRODUCTION

This paper, a review of our knowledge of planetary atmospheres, will concentrate on the outer planets since other papers in this volume discuss Mars and Venus,* but some information about these two planets will be included as well. Our interest in the atmospheres of the planets goes beyond a mere tabulation of their compositions; the relation of atmospheric composition to the question of the origin and evolution of the planets and their atmospheres and the problem of the origin of life on Earth should be kept in mind. The nature of the comets also has a bearing on these larger questions, so a brief discussion of some new observations of these distant members of the solar system will be included.

To begin this review, we should recall the classical division between the inner (or terrestrial) planets, and the outer (or Jovian) planets, with the outer group divided into Jupiter and Saturn on the one hand and Uranus and Neptune on the other. These divisions are made very grossly on the basis of mean density and mass, the inner planets having higher mean densities and smaller masses than the outer planets (Table 1). Attention is called to the new value for the mean density of Neptune, which is now very close to the accepted value for that of Uranus.

The outer planets must be vastly different in chemical composition from the rocky inner planets, since they are composed primarily of the lightest elements and thus are closer to the cosmic abundance scale. On the other hand, the higher mean densities and lower masses of Uranus and Neptune compared with those of Jupiter and Saturn suggest that the former two planets must be somewhat deficient in the lighter gases.

*See Chapter 6, "Lower Atmospheres of the Planets", by Hunten, and Chapter 12, "Evolution of Planetary Atmospheres", by Rasool.

Table 1.—Planetary characteristics.

Planet	Mass*	Radius*	Density (g/cm^3)
Mercury	0.06	0.38	5.6
Venus	0.81	0.96	5.1
Earth	1.00	1.00	5.5
Mars	0.11	0.53	4.0
Jupiter	318	11.19	1.3
Saturn	95	9.47	0.7
Uranus	15	3.73	1.6
Neptune	17	3.87	1.7
Pluto	<0.18	(0.47)	<9.8

*Values for mass and radius are given in multiples of the mass and the radius, respectively, of the Earth.

Pluto has been a problem because of its derived mean density $\bar{\rho}$ of 40 g/cm^3, which has been mentioned occasionally as evidence that Pluto is a terrestrial planet. However, this density is much too high for a planet of Pluto's size, so it actually means that the measurement of either the mass or radius of Pluto is in error. Recent studies indicate that Pluto's mass must be less than one-fifth of the Earth's and may be as low as one one-hundredth, though this is still very uncertain.

The satellites in the outer solar system also exhibit large variations in mean density. Jupiter's four major satellites mimic the run of densities in the solar system itself, ranging from $\bar{\rho}$ = 4.0 g/cm^3 for J I to $\bar{\rho}$ = 2.1 g/cm^3 for J IV. Two of Saturn's moons, Titan and Tethys, have mean densities of 2.4 and 1.2 g/cm^3, respectively. By comparison, the Earth's moon has $\bar{\rho}$ = 3.3 g/cm^3, and the density for uncompressed rock on Earth is 2.5 g/cm^3.

Let us now examine the relative ability of the planets and satellites to retain atmospheres. This depends on the escape velocity—a function of planetary mass and radius—and on the root-mean-square velocity of the molecules making up the

atmosphere—a function of atmospheric temperature. The temperature depends on solar distance and albedo, if we ignore for the moment the subtleties of atmospheric structure and composition. The list we obtain from such a calculation runs as follows (in decreasing order of ability to retain an atmosphere): Jupiter, Saturn, Neptune, Uranus, Earth, Venus, Pluto, Triton, Mars, Titan, J III, J I, J IV, J II, Mercury, and the Moon.

Since we are confining ourselves to a discussion of atmospheres, we obviously want to know how far down this list we can go. At present, the answer is Titan, with Pluto and Triton excepted. There are some intriguing indications of atmospheres on J I and J II, but other explanations of the evidence are possible as well.

Pluto and Triton owe their relatively advanced positions on this list primarily to the low temperatures calculated for an equilibrium with incoming solar radiation. It may be that these temperatures are so low that any detectable atmospheric constituent, whether retained from the primordial nebula or produced by crustal outgassing, would be frozen out. However, the observational tests for the presence of atmospheres on these bodies are still rather primitive because both objects are so faint they are difficult to observe.

The most likely candidate for a detectable atmospheric constituent retained from the primordial nebula is methane. This gas exhibits an absorption at 6190Å that is prominent in the spectra of the outer planets. Dr. Robert Hardie and the author have examined the spectrum of Pluto in this region but found no trace of the methane absorption. From its absence, we could set an upper limit of 20 m-atm on the methane abundance. This is not a very sensitive limit if Pluto's surface temperature is as low as 40 K, the value corresponding to solar heating. In that case, we would only expect a few centimeter-atmospheres of methane to be able to exist in equilibrium with the solid.

Triton may be some 10 to 20 K warmer, and we could then expect 1 to 10 m-atm of methane. This amount would permit detection of the stronger absorption band at 8900Å. A preliminary observation of this region of Triton's spectrum was kindly obtained for the writer by Dr. J. B. Oke with a multichannel spectrometer attached to the 200-inch telescope. The resulting spectrum shows a small absorption at 8900Å, which may be evidence of a methane atmosphere. Further observations are required to confirm this feature. An interesting sidelight to these observations is the fact that the color of Triton (from 3800Å to 9500Å) appears to be very similar to that of the asteroid Pallas, not at all what one would expect for an ice-covered satellite.

II. TECHNIQUES

Before proceeding to a discussion of those planets known to have atmospheres, it may be useful to introduce a few definitions and to discuss briefly some of the methods used in obtaining the results we will consider. Most of our information about the composition of planetary atmospheres has been derived from the interpretation of planetary spectra. In this work, the spectra are recorded by direct scanning or on photographic plates in long-focus spectrographs positioned at the coudé focus of large telescopes. The image of a planet is cast on the slit of the spectrograph and remains stationary, except for rotation, as the telescope follows the object. Thus, one can obtain spatial resolution on the planetary disk on the resulting spectrogram if compensation for the image rotation is provided.

The spectrum of a planet observed from the Earth's surface consists of three superimposed spectra: The solar, planetary, and terrestrial atmospheres all leave their imprint. Attenuation by the atmosphere of the planet is given by the expression

$$I_\nu = I_{\nu 0} \exp(-\alpha_\nu \eta u),$$

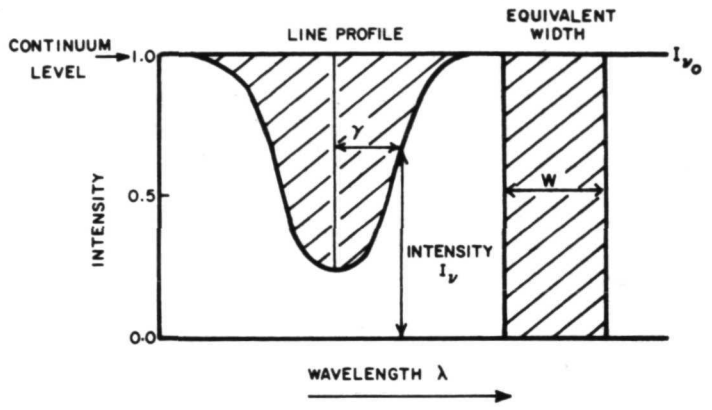
where α_ν is the absorption coefficient, u is the optical thickness, and η is the air-mass factor that is required to compensate for the differing geometry of observation over the planet's surface. Obviously, at the limb of a planet, we are looking through more atmosphere than at the center, and this must be taken into account in analyzing the observations.

It is customary to express the amount of gas u in units of meter-atmospheres, where 1 m-atm is the amount of gas at normal temperature and pressure (NTP) in a path 1 m long (about 2.687×10^{21} molecule-cm^{-2}). If we reduce the Earth's atmosphere to these units, we find a vertical column is equivalent to 7.99 km-atm.

What one observes in the spectrum is an absorption line or set of lines formed by a particular atmospheric constituent. A typical line profile is shown in Figure 1. The half-width γ of the line will depend on the local pressure and temperature according to the relationship

$$\gamma = \gamma_0 \frac{P}{P_0} \left(\frac{T_0}{T}\right)^{1/2}, \tag{1}$$

where the subscripts refer to NTP conditions. The total absorption of a line like this is given in terms of the equivalent width W, defined as the width of a rectangular



1. $W = \eta u S$
2. $W = 2(S\gamma\eta u)^{1/2}$
3. $\gamma = \gamma_0 \frac{P}{P_0}\left(\frac{T_0}{T}\right)^{1/2}$
4. $S = \int \alpha_\nu d\nu$

Figure 1.—Schematic representation of an absorption line (after A. Unsold).

cross section extending from the continuum to the zero level and having the same area as the absorption line. Mathematically, we define the equivalent width as

$$W = \int \frac{I_{\nu 0} - I_\nu}{I_{\nu 0}} d\nu$$

$$= \int 1 - e^{-\alpha_\nu \eta u} d\nu \,. \tag{2}$$

For a weak line in which no saturation occurs, this expression reduces to

$$W = \int \alpha_\nu \eta u \, d\nu = \eta u S \,, \tag{3}$$

where S is the integrated line strength, a function of temperature. The quantities S and γ are intrinsic to a given transition and can be measured in the laboratory.

The task of the observer may thus be viewed as follows: first, to identify absorption lines in the planetary spectrum that are caused by a constituent in the planet's atmosphere, and then to measure their wavelengths to identify the absorber. The total absorption is then measured to obtain W, and η is determined from the geometry of the observation. If S has been measured in the laboratory for this particular transition, it will be possible to obtain the abundance from Equation 3, provided that a reasonable estimate of local temperature can be made. If one has sufficient resolution to measure γ, it should also be possible to obtain some information on pressure. Pressure may also be determined from the integrated absorption of a strong line. Under favorable circumstances, temperature estimates may be obtained directly from the spectra by evaluating the relative intensities of a series of rotational lines occurring in a single vibrational transition of a given molecule. Hence, a large amount of information is available in the spectrum if we can make a proper analysis.

III. THE ATMOSPHERE OF JUPITER

In practice, this analysis often is not easy to make. It has been known for many years that the atmosphere of Jupiter contains methane, but the exact amount of this gas and the atmospheric temperature have been very difficult to determine. The reason may be seen from inspection of Figure 2, which is a reproduction of a

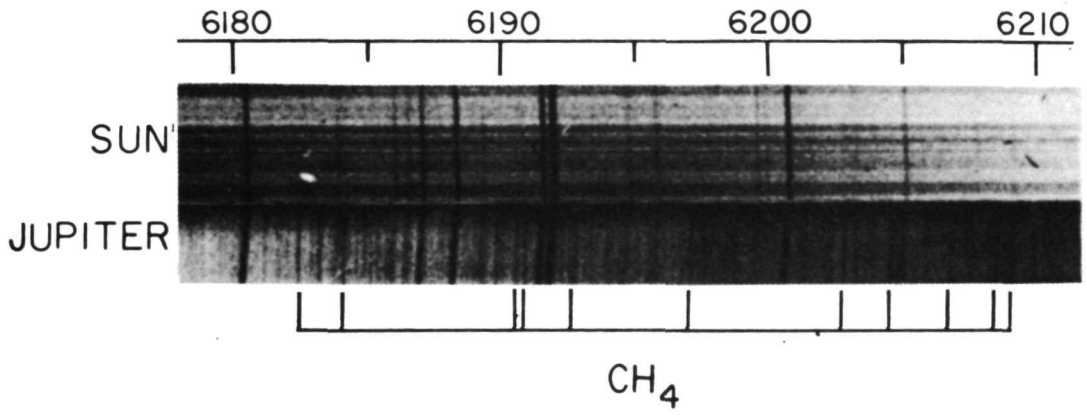


Figure 2.–The 6190Å methane band in the spectrum of Jupiter, with solar comparison (courtesy McDonald Observatory).

moderately prominent methane band at 6190Å. There are an enormous number of weak lines shown, and we have no idea to what quantum states in the molecule these individual transitions correspond.

There is the further difficulty that the atmosphere of Jupiter contains aerosols of various kinds; thus, the path of an incident photon can be quite complex. In particular, we expect that weak absorption lines will correspond to a longer effective path length than will strong lines. The scattering power of the aerosols will almost certainly exhibit some wavelength dependence. Thus, in making our abundance estimates, we must try to use lines of comparable intensity formed in the same region of the spectrum, and we must content ourselves with relative rather than absolute abundances.

The following values for Jovian atmospheric constituents are obtained from consideration of those methane and hydrogen absorption bands which can be analyzed (cf. Figures 3 and 4) and from the best estimate that can be made of the ammonia abundance:

$$H_2 = 75 \pm 15 \text{ km-atm},$$

$$CH_4 = 50 \pm 15 \text{ m-atm},$$

and

$$NH_3 = 12 \pm 5 \text{ m-atm}.$$

These values are derived from absorption bands occurring in the infrared, but it has not yet proved possible to satisfy all of the criteria we established for such investigations, since the methane lines that have been analyzed are considerably stronger than the hydrogen or ammonia lines, and thus the methane abundance may be underestimated. On the other hand, the hydrogen lines, since they arise from quadrupole transitions, have very different profiles from the methane or ammonia lines, being considerably narrower and thus perhaps formed slightly higher in this scattering atmosphere than an equivalent line of Lorentzian profile. Hence, the methane/hydrogen comparison may not be too bad. A study of weak methane lines would help to resolve this uncertainty.

In any event, the composition of this atmosphere is obviously very different from that of the Earth. The large excess of hydrogen is reminiscent of the Sun and

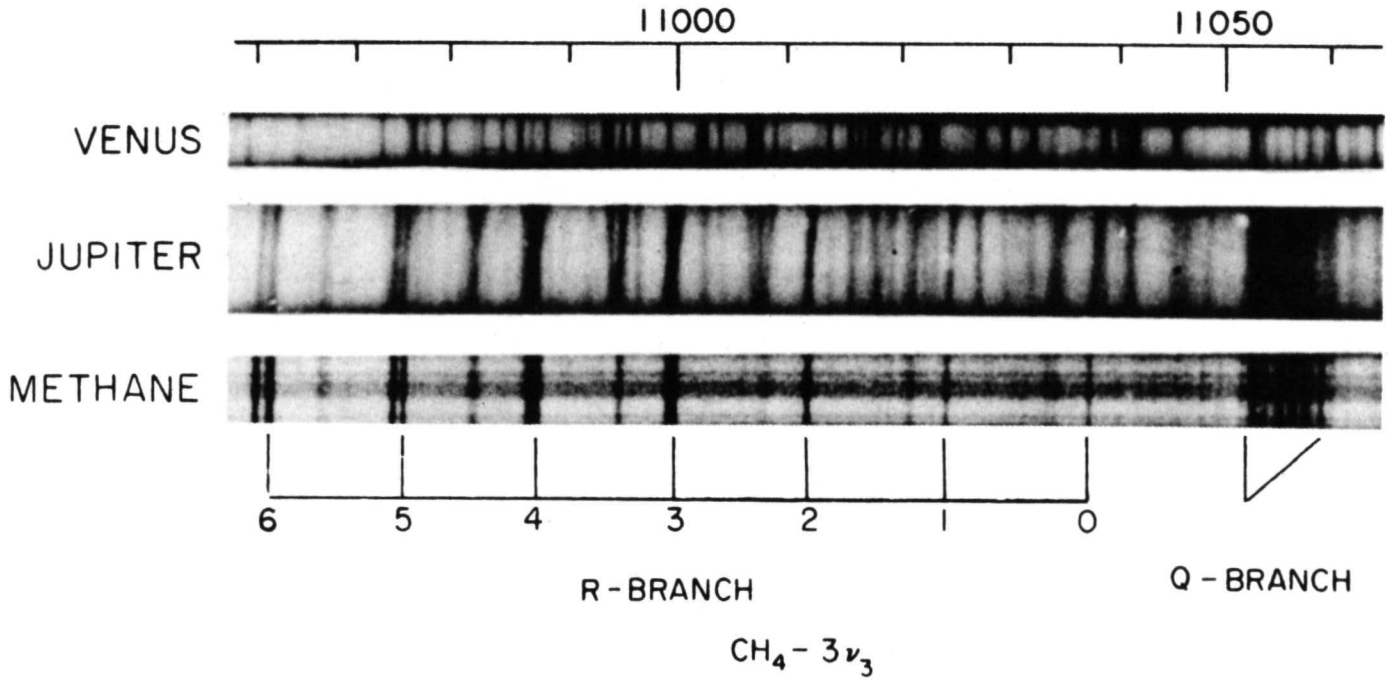


Figure 3.–The $3\nu_3$, 11 057Å methane band in the spectrum of Jupiter. The spectrum of Venus given for comparison contains only telluric and solar lines (courtesy McDonald Observatory).

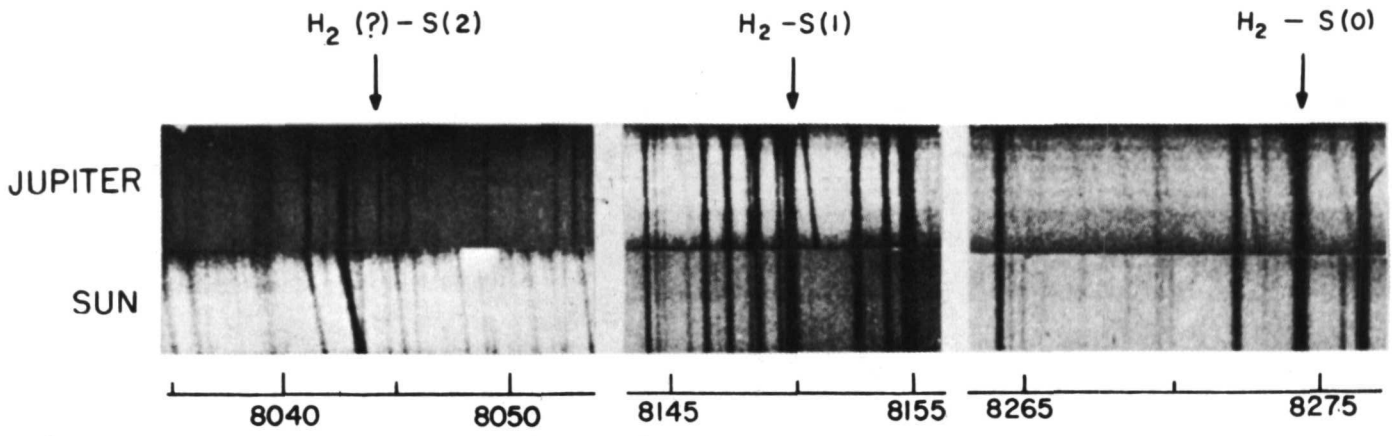


Figure 4.–Hydrogen lines from the 3-0 quadrupole band in the spectrum of Jupiter with solar comparison. Solar and planetary lines appear tilted in the spectrum of Jupiter, owing to the Doppler effect produced by the planet's rotation (courtesy McDonald Observatory).

stars, so we might compare the ratios of the equivalent element abundances to solar values:

Body	H/C	H/N
Jupiter	3000	1.5×10^4
Sun	2880	1.2×10^4

In view of the large uncertainties in both sets of numbers, the agreement seems very good. We expect that if these two ratios agree, the other elements will also be present on Jupiter with solar relative abundances, since likely loss mechanisms would lead to the preferential escape of light gases.

We conclude, therefore, that within present uncertainties, Jupiter has a composition identical to that of the Sun, immediately suggesting that the planet has retained the original mix of materials that condensed to form it from the primordial solar nebula. If this is true, we have an opportunity to explore the composition of a small region of our galaxy as it existed 4.5 billion years ago.

Additional tests of this conclusion are immediately evident. Helium is the second most abundant element in the Sun—how much is present on Jupiter? Unfortunately, this is a very difficult question to answer. The strongest ground-state helium line is at 584Å, a very difficult region in which to work, so the problem has been approached indirectly. The most effective method used to date involves the study of the profiles of weak absorption lines. As we have seen above, the half-width of such lines can be used to obtain an estimate of the local pressure if one can estimate the temperature with some precision [the temperature varies only as the square root (Equation 1)]. We have independent estimates of the pressure from the total abundance of hydrogen, and we can then ask how much helium must be added to produce the observed half-width. The result is an upper limit on the helium abundance of 20 km-atm. This implies a ratio of H/He $\geqslant 9$ for Jupiter, as compared with the solar value of 11^{+7}_{-5}. Again, we have consistency with the idea of solar abundances in the Jovian atmosphere; but, again, the uncertainties are large.

Another test of this hypothesis, and one which offers the opportunity for considerable cosmological insight as well, is afforded by the study of isotopic ratios, especially C^{12}/C^{13} and H/D. On the Earth $C^{12}/C^{13} \approx 90$, but the ratio is as low as

4 in the atmospheres of some carbon stars and has intermediate values in the interstellar medium. Deuterium has not been detected in stars or the interstellar medium; H/D ≈ 6500 on the Earth. Both C^{13} and H^2 are formed by astrophysical processes of great cosmological interest, and one theory for the origin of the solar system suggests that in the present solar system there should be decreasing amounts of these isotopes with increasing distance from the Sun.

We may look for $C^{13}H_4$ on Jupiter, but this requires a laboratory program to determine the wavelengths and intensities of the lines, since they cannot be computed *a priori*. Such a program is presently underway and should soon yield the necessary data. In the case of deuterium, we may look for HD, CH_3D, or NH_2D. The first of these appears to offer the best hope of success in view of the very large amount of hydrogen present in Jupiter's atmosphere. Once again, laboratory spectroscopy must provide some basic data, but on the basis of preliminary studies already published, it appears that H/D > 3000 in the Jovian atmosphere.*

All of the above conclusions are based on a simple atmospheric model that ignores the presence of scattering particles. We have taken this step deliberately, accepting the restriction that we can discuss only relative abundances, and we must exercise considerable caution even in deriving these. The need for such precautions is illustrated by the investigation of clouds and haze.

If the planet were to be photographed through an interference filter centered on a strong methane band, we would expect to see a rather uniform disk, brighter at the center than at the limbs, which is in keeping with the increase in absorption toward the limbs resulting from the longer optical path. Instead, we find a great deal of structure, as illustrated in Figure 5. This structure indicates the presence of an inhomogeneous haze of scattering particles, the lighter regions corresponding to less absorption and, hence, to the presence of a denser haze than the darker regions. The limb darkening appears more pronounced in the darker regions, which suggests that they are clearer, an interpretation that is consistent with polarization observations. This haze is presumably ammonia cirrus which lies at higher levels in the atmosphere than the main cloud deck.

The Great Red Spot appears particularly bright in such pictures, and it is tempting to associate its color with this high haze. The color of this object, as well as

*An absorption in the spectrum of Jupiter at 4.73 μm was first observed by G. Münch and G. Neugebauer in the summer of 1970 at relatively low resolution. These authors suggested several possible identifications, among them CH_3D and HCN. One year later, R. Beer and collaborators obtained a high-resolution spectrum of this band, positively identifying the absorber as CH_3D. This result is in press at the present writing; an abundance and a value of H/D must await further analysis.

Figure 5.—The appearance of Jupiter in a wavelength region of strong methane absorption: 8800Å to 9000Å (courtesy McDonald Observatory).

those of other regions on the disk, remains an intriguing problem. Chemical equilibrium studies in which a solar mixture of elements is allowed to combine in a large variety of ways do not indicate the production of any colored compounds in the spectroscopically accessible region of the atmosphere, with the possible exception of $(NH_4)_2S$. This compound may form in the equilibrium between NH_3 and H_2S and may be yellow in the frozen state, and it has been suggested as a possible constituent of the main cloud deck, along with NH_4SH. Other sulfur-ammonia compounds could result from photochemical reactions in the upper atmosphere, as well as from electrical discharges within the clouds.

The latter phenomena should also lead to the production of organic polymers, many of which are colored. This possibility is especially attractive because of the many laboratory experiments in which mixtures of gases similar to the composition

of the Jovian atmosphere have been irradiated by UV light or subjected to electrical discharges, with a resultant formation of complex organic molecules. These experiments have been oriented specifically toward the problem of the origin of life on Earth, and one thus has an added incentive to discover how far such reactions have proceeded on Jupiter.

In addition to H_2S, the chemical equilibrium studies indicate that large amounts of H_2O would also form from a solar abundance mixture. It appears that the only reason we have not detected either substance spectroscopically is because of their low vapor pressures at the ambient temperatures in the region above the lower cloud deck. At still lower levels in the atmosphere, clouds of NH_4OH and, possibly, ice should form.

An opportunity to probe the lower atmosphere is afforded by observations at 5 μm. Through one of nature's happy coincidences, this region of the spectrum is a window through both the telluric and Jovian atmospheres; neither methane nor ammonia have strong absorptions here. One would thus expect to see rather deeply into the Jovian atmosphere at this wavelength, but it has been surprising to discover just how far. Observations with high spatial resolution have indicated that there are regions on the disk, notably the North Equatorial Belt (NEB), where radiation corresponding to a temperature in excess of 300 K can be detected. The exact temperature cannot be determined, because the size of the emitting region is smaller than the aperture of the photometer used to make the observations. The emission from the rest of the planet at these wavelengths is very low, although there is a tendency for dark belts to be warmer than the general background. This is not true of the Great Red Spot, however.

Thus, we have an intriguing problem. Why is it that the atmosphere is so much more transparent over the NEB than elsewhere? This must imply a break in the clouds, and one may then ask what sets the lower bound in this region. If it is another cloud layer, the composition of the clouds must be different–perhaps these are the water clouds mentioned above. If it is a gas phase constituent, some interesting possibilities present themselves; among them are water vapor and CH_3D. Clearly, this is a region of the spectrum that should receive further exploration.*

IV. SATURN, URANUS, AND NEPTUNE

Jupiter is to the student of the outer planets what the Sun is to the stellar astronomer. We can examine it in much more detail than we can the other objects;

*See footnote on page 253.

thus, we hope to infer some of their properties from our understanding of Jupiter. However, we already know on the basis of bulk composition that this is not an absolutely reliable procedure, and our investigation of these planets must proceed as carefully as possible.

The same hydrogen lines observed on Jupiter have been studied on Saturn, with the result that an abundance of 190 ± 15 km-atm has been obtained. We expect that methane must also be present in greater abundance than on Jupiter, and available evidence appears to support this assumption, although we do not yet have a detailed analysis of the $3\nu_3$ band on which to base such an estimate (see Figure 3). Ammonia has been reported in the spectrum of Saturn, but the evidence is weak and unreliable. If the infrared temperatures derived for this planet are correct, the vapor pressure of ammonia probably would be too low to permit its detection. Saturn also should have an atmospheric window at 5 μm, and an increase in radiation has been observed there, but areal coverage comparable to that of Jupiter has not yet been obtained.

Saturn is known to have a layer of clouds, but the layer is much more uniform in color than that of Jupiter and apparently less active as well. Photographs of the planet taken through the same interference filter used for Jupiter reveal that the Equatorial Zone is the brightest region on the planet, the rest of the disk fading out uniformly toward the poles as a result of the strong methane absorption. The Equatorial Zone is peculiar in another respect, being yellowish in color and hence dark in the UV. Again, it may be that we are dealing with sulfur-ammonia compounds, as suspected in the case of Jupiter.

We cannot end our discussion of Saturn without mentioning its remarkable system of rings and its largest satellite, Titan, the only satellite known to have an atmosphere. The spectrum of the rings reveals that the particles that compose them are partially coated with ice. At blue and violet wavelengths, the rings exhibit a decrease in reflectivity indicative of an exposed rocky surface that has reflecting properties similar to the surfaces of Mars and Jupiter's satellite Io. The ice deposits are revealed through solid-state absorption bands in the 1- to 2.5-μm region. Solid ammonia would not be stable over the lifetime of the solar system at this distance from the Sun, so it is not surprising that it is not found in the ring spectrum.

Titan was found to have an atmosphere of methane by Kuiper over 20 years ago. However, we know almost nothing about it. In particular, it would be interesting to know the total atmospheric pressure and then to see what other gases

(such as argon) might be present. Furthermore, the atmospheric photochemistry, in the absence of the large amounts of free hydrogen that exist in the atmospheres of Jupiter and Saturn, might lead to the equilibrium production of other hydrocarbons. Observations would not be easy to make because Titan is relatively faint, but they are still within the present state of the art.

Uranus and Neptune are even more difficult to observe, but it has been known for many years that their spectra are dominated by absorption bands of methane. In fact, there is so much absorption from this gas that Neptune is distinctly greenish in color. It has proved more difficult to determine the amount of gas responsible for this absorption, since once again we do not have access to the $3\nu_3$ band, the only methane band that has been adequately analyzed in the laboratory. However, a band with similarly open structure has been found at 6800Å (Figure 6), and this system holds the promise of being analyzable once a suitable laboratory program has been carried out. For the moment, one must rely on direct comparisons with laboratory spectra, using weak lines such as those making up the 7500Å series of absorptions.

The 4-0 quadrupole lines of hydrogen have been observed in the spectrum of Uranus but not Neptune. Both planets exhibit pressure-induced dipole bands of hydrogen, however, which can also be analyzed to obtain an abundance of this gas. Thus far, this has only been done for Uranus. The results of these various analyses are summarized in Table 2.

Table 2.–Observed composition of the Jovian planets.

Planet	H_2 (km-atm)	NH_3 (m-atm)	CH_4 (m-atm)	H/C*
Jupiter	75 ± 15	12 ± 3	50 ± 15	3000
Saturn	190 ± 50	–	350 (115 ± 50)**	1100 (3300)**
Uranus	400 ± 100	–	3500 ± 1500	230
Neptune	–	–	(6000 ± 2500)	–
Pluto	–	–	<20	–

*The solar value of H/C is 2880.

**The figures given in parentheses are based on the assumption that the same correction will apply to the methane abundance on Saturn as was found to be necessary on Jupiter.

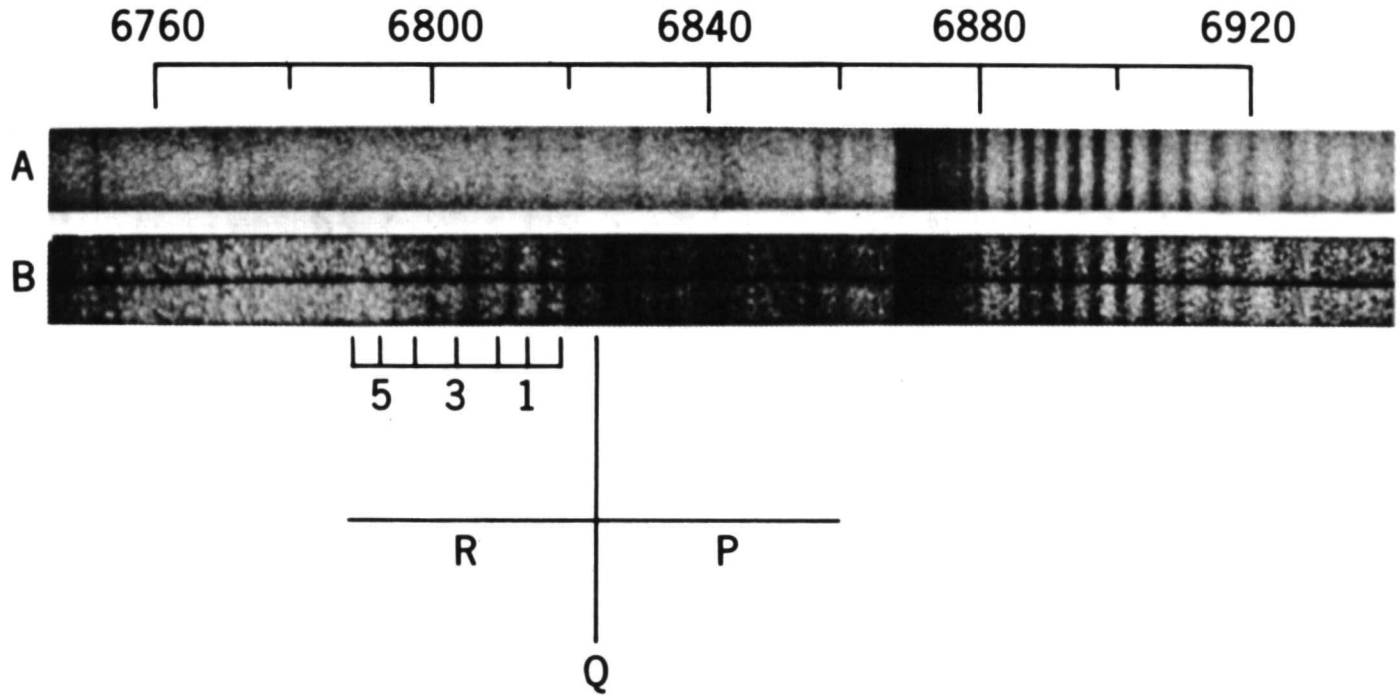


Figure 6.–The 6800Å methane band in the spectrum of Uranus (B) with a comparison spectrogram of Jupiter (A). The molecular absorption with a band head at 6867Å is telluric oxygen (courtesy Kitt Peak National Observatory).

No gases other than methane and hydrogen have been detected in the spectra of these planets; the atmospheres are undoubtedly too cold for ammonia to exist in detectable amounts. No evidence for clouds has been found on either planet, so it appears that we are dealing with clear, deep, Rayleigh-scattering atmospheres above very low cloud surfaces. Studies at radio wavelengths for which penetration into the atmospheres is greatest have indicated temperatures in excess of 200 K, but these observations are still in a preliminary state.

V. MARS AND VENUS

Since Mars and Venus are discussed in detail in other papers in this volume,* the remarks here will be limited to a few details relevant to this survey. Some new observations of Mars have been carried out with the Orbiting Astronomical Observatory (OAO) in the spectral region below 3000Å. As expected, the brightness of the planet increases near the short-wave limit of the observations at 2000Å. This increasing brightness is caused by the λ^{-4} dependence of Rayleigh scattering, and, indeed, the observations fit a Rayleigh atmosphere composed of 78 m-atm of CO_2, the composition derived from infrared observations. There is, however, an interesting discrepancy: a dip at 2600Å, which can be neatly accounted for by the addition of 0.025 part per million (ppm) of ozone. Unfortunately, this interpretation is not unique, since the atmosphere is optically thin at these wavelengths; thus, a solid-state absorption produced by some material on the planet's surface could also cause this feature. An especially interesting possibility is C_3O_2 or one of its polymers. The presence of this substance has been invoked to explain the spectral reflectivity of Mars at longer wavelengths and is known to have an absorption at 2600Å. Further study is needed to resolve this ambiguity.**

The absence of other absorption features in this region permits us to set the following upper limits (in ppm) on other possible atmospheric constituents:

$H_2S < 2$	$N_2O_4 < 5$
$SO_2 < 2$	$NO_2 < 10$
$NH_3 < 2$	$NO < 10$

*See footnote on page 243.

**A new analysis of the UV spectrum making use of revised solar photometry indicates that the 2600A absorption is spurious. The abundance given for ozone thus becomes an upper limit. Observations of a dip at 2600A occurring only in spectra of the south polar cap have been reported by the Mariner 6 and 7 UV spectrometer team (Barth et al., 1969).

The sensitivity of these limits is indicative of the high UV absorption cross section of these molecules. It will be difficult to exceed this sensitivity even with an *in situ* analysis.

It should be apparent that in leaving the outer solar system, our orientation has changed essentially from an astrophysical to a geophysical one. We are no longer dealing with atmospheres of hydrogen, methane, and ammonia but expect instead to find carbon dioxide, nitrogen, and water. However, the residual geocentrism that has haunted astronomy with special fervor since Ptolemy must be guarded against when comparing these planets with the Earth.

Ten years ago, it was widely assumed that the atmospheres of Mars and Venus were composed primarily of nitrogen, like that of the Earth. The observational evidence that was available did not contradict this view. With the wisdom of hindsight, however, we can see that it should have been remembered that the present terrestrial atmosphere is by no means representative of the relative amounts of volatiles outgassed by the Earth's crust over geologic time. In order of decreasing abundance, the principal volatiles were water vapor, carbon dioxide, nitrogen, and argon. The absence of large amounts of carbon dioxide in the Earth's atmosphere is the result of its precipitation in the form of carbonate rocks, brought about by an equilibrium reaction between carbonates and silicates that requires the presence of liquid water. The abundance of terrestrial life has also played a role in this equilibrium, as evidenced by the limestones formed of shells of marine creatures.

Early searches for water in the atmospheres of Venus and Mars led to negative results, so it might have been expected that if these planets were indeed similar to the Earth but somehow deficient in water, their atmospheres should be composed predominantly of carbon dioxide. In fact, it turns out that both planets have atmospheres that are over 90 percent CO_2, which again offers an attractive parallel between their atmospheric evolutions and the Earth's. For a more detailed investigation of this possibility, knowledge of the atmospheric water vapor abundance is very important.

In the case of Venus, this problem is particularly vexing. Studies of the planet's spectrum have led to several conflicting results. The author feels that the most reliable observations are those performed in a high-flying aircraft by G. P. Kuiper and his associates at the University of Arizona. These studies led to the detection of roughly 2 μm of precipitable water. In other words, if all the water vapor in the one-way path in the atmosphere that produces the spectrum were condensed out, it would correspond to a layer only 2 μm thick. For comparison, water vapor in the Earth's atmosphere at sea level is typically several precipitable centimeters.

This should be the end of the story; unfortunately, it is not. The Soviet space probes that entered the atmosphere of Venus reported much more water than this at lower atmospheric levels. Whereas it is possible to imagine ways of trapping the water at lower levels, the temperature at the level of the atmosphere probed by the spectroscopy is sufficiently high (above 240 K) as to make this unlikely. It may be that the Soviet measurements are in error; the results from additional probes are awaited with great interest.

For Mars, the situation is much clearer. The 8200Å water vapor band has been observed with high resolution at several oppositions, and Martian water vapor lines are definitely present. Again, the amount of precipitable water is small by terrestrial standards, being of the order of 25 μm. Variations in the abundance appear to occur with the seasons on Mars, water vapor being most abundant in the hemisphere in which late spring is occurring; but additional observations are required to verify this correlation.

If the analogy with terrestrial outgassing is to hold rigorously on Mars, approximately 7 m of water per square centimeter are needed to be in proportion to the known amount of carbon dioxide in the Martian atmosphere. Undoubtedly, some of this water has been dissociated by UV radiation, with subsequent escape of the hydrogen and crustal trapping or escape of the oxygen. The hydrogen corona on Mars observed by Mariners 6 and 7 corresponds to a present escape rate for hydrogen of 10^8 atom-cm^{-2}-s^{-1}. Assuming that this rate has been constant over the last 4.5 billion years, we may anticipate a loss of 7.1×10^{24} molecule-cm^{-2} of water vapor, or about 2 m of water. Certainly, the escape rate is limited at the present time by the low tropopause temperature (below 200 K) and the UV absorbing ability of CO_2, which will protect water vapor near the ground. This latter factor may well have been considerably less important at an earlier era when outgassing had not proceeded to its present extent; therefore, the loss of water may have been more rapid in the past.

There is also the possibility that a large proportion of the missing water is still present but in the form of permafrost or adsorbed soil moisture that we cannot yet detect. However, it is also possible, that the Martian atmosphere has evolved quite differently from the Earth's, that contributions from cometary and meteoritic impacts have been more significant than crustal outgassing. The Viking Lander mission planned by NASA for 1975 should provide answers to some of these questions.

To conclude this discussion of Mars, we may ask whether any of these new results shed any light on the popular conception of Mars as an abode of life. The

certainty that the Martian atmosphere contains water vapor is definitely a hopeful sign. However, this should not be interpreted to mean that pools of water exist on the surface. The mean atmospheric surface pressure is close to the triple point of water at 6.1 mb. Hence, only in low areas, where we may anticipate pressures in the range from 10 to 15 mb, would water be stable in the liquid form. Even there, evaporation would be exceedingly rapid, and sublimation from frost rather than melting and evaporation would probably occur. On the other hand, water may be available in the soil, particularly if permafrost is present (but we can anticipate only very thin coatings of soil particles). If the presence of ozone is confirmed, Martian organisms would have partial protection from biolethal radiation at 2600Å. They would pay for this protection by being in direct contact with the ozone since, unlike the terrestrial situation, ozone on Mars will be concentrated near the ground. However, the present upper limit is only one-twentieth of that leading to a Los Angeles smog alert, which even some forms of terrestrial life are known to survive.

On balance, then, we can say that conditions on Mars are perhaps slightly more favorable for the existence of life than we had been willing to admit in the wake of the initial reports from Mariners 6 and 7. Definitive evidence is still lacking, and we must again defer to the Viking mission to provide that evidence.

VI. THE COMETS

A discussion of comets will end this summary of the composition of planetary atmospheres. Comets are important members of the solar system and obviously must provide some clues to the question of its origin. Therefore, any review of the general problem should include a consideration of their composition. Such a discussion is particularly timely owing to the recent appearance of two bright comets: Tago-Sato-Kosaka (TSK) and Bennett.

Two of the most extraordinary characteristics of comets are associated with their orbits rather than their compositions. Comet orbits are remarkable for their spherical distribution in space (as opposed to the essentially coplanar distribution of planetary orbits) and their immense aphelion distances, occasionally reaching almost halfway to the nearest star. In both instances, we are speaking of new or long-period comets, since short-period comets have aphelia typically within the orbit of Jupiter and orbits that tend to be confined more toward the plane of the planetary system. These orbital characteristics are strikingly similar to those exhibited by the globular clusters with respect to the galaxy and might lead one to infer that the comets, like the globulars, are among the oldest members of their parent system, perhaps formed

before the planets themselves. However, whether or not this kinematic similarity really bespeaks a similar mode of origin or is merely coincidental remains to be determined.

Cometary spectra are very different from those of the planets. Typically, we observe a continuum consisting of reflected sunlight on which is superimposed an emission spectrum composed of various series of molecular and atomic features. We interpret the continuum as resulting from sunlight scattered by dust and other solid particles, whereas the emission lines arise from fluorescing gases. Long-period comets, presumably arriving in the vicinity of the Sun for one of the first times in their existence, tend to be much richer in dust than the short-period comets. This has led to the distinction between "new" and "old" comets, although all comets have presumably the same age. A study of the molecular emission lines indicates that they arise from radicals, such as C_2, C_3, CN, and OH, and not from stable molecular species, such as CH_4 and H_2O. This has led to the so-called icy conglomerate model of the comet nucleus, elaborated by F. Whipple, in which the nucleus consists of stable ices and hydrates which sublime as the comet reaches the Sun. The parent molecules rapidly become dissociated and ionized and subsequently produce the observed emissions.

We thus have a picture of a vast store of comet nuclei in the outer reaches of the solar system beyond Pluto, occasionally sending members in toward the inner solar system as a result of perturbations of nearby stars. Some of these comets are trapped by the gravitational influence of Jupiter, becoming short-period objects that gradually lose mass during repeated perihelic journeys. In studying these objects, then, we again have the opportunity to investigate conditions that existed during the early history of the solar system, but now the matter we investigate is in a new state of concentration, intermediate between that of the terrestrial planets and Jupiter and Saturn.

Two comments are in order at this point. First, as H. Brown suggested long ago, the cometary composition is rather similar to that apparently exhibited by Uranus and Neptune. Second, the radicals that we observe in the spectra of comets are virtually identical to those observed in interstellar space. This is especially interesting in view of the recent discovery of stable molecules such as H_2O, NH_3, and HCN in interstellar space. Thus, we are confronting objects that contain important clues to the origin of the solar system, but the clues are presently very difficult to interpret because we have no good theory for the origin of the comets themselves. It is tempting to identify this process with the same sequence of events that led to the observed fractionation of light gases at the distance of Uranus and Neptune, but it is by no means established that this was the case.

With these considerations in mind, we may turn to the observations of the two recent comets: TSK was primarily gaseous and thus ranks as an "old" comet, whereas Bennett was very dusty and thus qualifies as "new". The weak continuum of TSK made it possible to study the emission line spectrum in considerable detail. In particular, it was possible to investigate the intensity of the isotopic band of carbon, $C^{12}C^{13}$. As in the case of the outer planets, we wish to know the value of C^{12}/C^{13} to obtain information about the place and mode of origin of these bodies. For example, there are scientists who still believe that comets originate as a result of encounters between the Sun and interstellar dust clouds. In this case, we would expect to find varying values of C^{12}/C^{13}, in accord with the variations in this ratio observed in the interstellar medium. Only one comet has been investigated in this way before, and the resulting value of C^{12}/C^{13} was 75 ± 15, which is in reasonably close agreement with the telluric value of 90. Our studies of TSK are still not complete but suggest a value greater than or equal to 60. An exact determination must await a study of the NH_2 spectrum, since a band of NH_2 is blended with the C^{12}/C^{13} feature used for this analysis.

Other regions of the spectrum investigated for Bennett and TSK have not led to any remarkable new results, with the exception of the UV below 3000Å. Once again, the OAO was employed to investigate this previously inaccessible region, and the results were very rewarding. Both comets were found to exhibit immense hydrogen coronas, fluorescing brightly in the light of Lyman α. This phenomenon had been predicted a number of years ago by L. Biermann as a result of the dissociation of water vapor molecules subliming from the icy nucleus.

It is becoming clear that to make real progress in cometary studies, we must send a probe to a comet's vicinity. Ideally, one would like to achieve rendezvous with a bright comet while it is still far from the Sun and then travel in with it, observing the changes that occur with decreasing solar distance. In practice, this is extremely difficult to do, since the orbit must be known with precision well in advance of perihelion. Only Halley's comet satisfies this constraint, and its retrograde orbit places severe energy requirements on candidate spacecraft.

A more modest mission would involve sending a spacecraft on a flyby mode past a well-known short-period comet. It would be possible to fly through the coma and close to the nucleus and make *in situ* measurements of magnetic fields and composition as well as to observe the nucleus with the high spatial resolution that cannot be achieved from the ground. Possible candidates for such missions include D'Arrest, Encke, and Kopf.

VII. CONCLUSIONS

The inhabitants of the outer solar system—planets, satellites, and comets—appear to exhibit very directly the results of processes that occurred in the early history of the formation of the solar system. We are becoming accustomed to dealing with the rocky part of this fossil record as it is exhibited by the exposed surface of the Moon, but to investigate the history of the much more massive icy and gaseous fraction of the original material, we must study the giant planets.

It is striking nonetheless that in this remote region of our solar system, we find evidence of compounds and chemical processes that play a crucial role in theories for the origin of life on the Earth. These initial processes, leading only to the production of complex organic molecules, have long since ceased to occur on the Earth but may be taking place at the present time on Jupiter. What stage of complexity has been achieved? On the other hand, the study of comets, especially in the context of the new host of molecules discovered in the interstellar medium, suggests that some rather elaborate organic compounds may have been present in the original nebula itself.

Further investigation of these bodies is also bound to improve our presently misty thinking about the formation of the solar system and the general process of star formation. A key role in these considerations will be played by better values of the abundances of atmospheric gases, including isotopes.

In the inner solar system, we find ourselves once again tempted by apparent similarities among Venus, Earth, and Mars. It seems possible at present to conclude that the atmospheres of all three planets evolved similarly from crustal outgassing. In the case of Venus, this process was very complete, and the volatiles have remained in the atmosphere. In fact, the total amount of CO_2 on Venus is very similar to that evolved (and redeposited) on the Earth. One can account for the present absence of large amounts of water vapor in one of several ways, all leading to the transport of the water to the upper atmosphere where it can be dissociated. In this view, Mars simply has not outgassed as much as the Earth, perhaps because the planet never differentiated. As we have seen, there are again satisfactory means for disposing of the unseen water.

However, we should keep in mind several viable alternatives to this attractive picture. Perhaps Venus was deficient in water from the beginning. The high mean density of Mercury indicates that this planet lost an appreciable fraction of the more volatile silicates at some point during its formation. It may be that temperatures in the solar nebula were high enough at the distance of Venus from the Sun to prevent

large amounts of water-bearing minerals to be accreted. A search for other elements that might be fractionated in such a process would be helpful. The apparent absence of mercury compounds in the clouds of Venus might be an example of this effect.*

The atmosphere of Mars is sufficiently tenuous that cometary impact may be proposed as a significant contributor of volatiles. The meteoritic bombardment of the surface may have been more significant in liberating gases than a terrestrial type of crustal outgassing. On the other hand, the Martian atmosphere may once have been more dense than it is now, with significant losses having occurred with the loss of a primordial, induced planetary magnetic field, which occurred after radioactive heating of the planet's interior. The distribution of small craters can be interpreted as evidence of an early process of erosion that is now absent, which would tend to support such a model (A. Binder).

These examples do not exhaust the list of alternative explanations but serve to illustrate some likely possibilities. As in the case of the outer planets, we still lack the information required to come to grips with the basic problems raised in the introduction to this review. Nevertheless, the directions our new data are taking us seem promising, and one can anticipate very significant progress from the results of space missions and ground-based observing programs already scheduled for this decade.

REFERENCE

Barth, C. A., Fastie, W. G., Hord, C. W., Pearce, J. B., Kelley, K. K., Stewart, A. I., Thomas, G. E., Anderson, G. P., and Raper, O. F., "Mariner 6: Ultraviolet Spectrum of Mars Upper Atmosphere", *Science* 165:1004, 1969.

BIBLIOGRAPHY

General Reviews

McElroy, M. B., "Atmospheric Composition of the Jovian Planets", *J. Atmos. Sci.* 26:798-812, 1969.

Moroz, V. I., "Physics of Planets", NASA Technical Translation F-515, April 1968, Chap. 5.

Owen, T., "Jupiter and the Outer Planets", *Earth and Extraterrestrial Sciences* 1:89-97, 1970.

Owen, T., "The Atmosphere of Jupiter", *Science* 167:1675-1681, 1970.

*However, see the discussion of possible Mercury clouds at levels in the atmosphere not accessible to spectroscopic probing on page 227 in Chapter 6, "Lower Atmospheres of the Planets", by Hunten.

Rasool, S. I., "Jupiter: 'Rosetta Stone' of the Solar System", *Astronaut. Aeronaut.*, pp. 24-27, October 1968.

Sources of Detailed Information

"The Atmospheres of the Jovian Planets", papers from the Third Arizona Conference on Planetary Atmospheres, *J. Atmos. Sci.* 26(5), part 1, 1969.

"A Symposium on Jupiter and the Outer Planets", *Icarus* 10:355-411, 1969.

Additional Articles on Specific Topics

Duncombe, R. L., Klepczynski, W. J., and Seidelmann, P. K., "Mass of Pluto", *Science* 162:800, 1968.

Freeman, K. C., and Lynga, G., "Data for Neptune From Occultation Observations", *Astrophys. J.* 160:767, 1970.

Gehrels, T., Herman, B. M., and Owen, T., "Wavelength Dependence of Polarization. XIV. Atmosphere of Jupiter", *Astron. J.* 74:190, 1969.

Giver, L. P., and Spinrad, H., "Molecular Hydrogen Features in the Spectra of Saturn and Uranus", *Icarus* 5:586, 1966.

Lebofsky, L. A., Johnson, T. V., and McCord, T. B., "Saturn's Rings: Spectral Reflectivity and Compositional Implications", *Icarus* 13:226, 1970.

McGovern, Wayne E., "Upper Limit of Hydrogen and Helium Concentrations on Titan", *Proc. IAU Symp. 40 Planetary Atmospheres*, D. Reidel, Holland, 1971.

Margolis, J. S., and Fox, K., "Studies of Methane Absorption in the Jovian Atmosphere: I. Rotational Temperature From the $3\nu_3$ Band", *Astrophys. J.* 157:935, 1969.

Margolis, J. S., and Fox, K., "Studies of Methane Absorption in the Jovian Atmosphere: II. Abundance From the $3\nu_3$ Band", *Astrophys. J.* 158:1183, 1969.

Mason, H. P., "The Abundance of Ammonia in the Atmosphere of Jupiter", *Astrophys. Space Sci.* 7:424, 1970.

Moroz, V. I., "The Spectra of Jupiter and Saturn in the 1.0-2.5μ Region", *Soviet Astron. A. J.* 10:457, 1966.

Münch, G., and Younkin, R. L., "Molecular Absorptions and Color Distributions Over Jupiter's Disk" (abs.), *Astron. J.* 69:553, 1964.

Owen, T., "An Identification of the 6800Å Methane Band in the Spectrum of Uranus and a Determination of Atmospheric Temperature", *Astrophys. J.* 146:611, 1966.

Owen, T., "Comparisons of Laboratory and Planetary Spectra IV. The Identification of the 7500Å Bands in the Spectra of Uranus and Neptune", *Icarus* 6:108, 1967.

Owen, T., and Mason, H. P., "The Abundance of Hydrogen in the Atmosphere of Jupiter", *Astrophys. J.* 154:317, 1968.

Poll, J. D., "Estimate of the H_2 Abundance in the Atmosphere of Uranus From the Pressure Induced Spectrum", *Proc. IAU Symp. 40 Planetary Atmospheres*, D. Reidel, Holland, 1971.

Teifel, V. G., "Spectrophotometry of the Methane Absorption Bands at 0.7-1.0μ on the Disk of Jupiter", *Soviet Astron. A. J.* 10:121, 1966.

Page intentionally left blank

CHAPTER 8
INTERIOR STRUCTURE OF GIANT PLANETS

R. Smoluchowski
Princeton University
Princeton, New Jersey

I. INTRODUCTION

The behavior of matter at high pressures and temperatures is fairly well understood, and consequently the theory of the interior structure of most of the stars is rather satisfactory. Similarly, one expects the interiors of extremely dense, and extremely hot stars, such as white dwarfs and neutron stars, to have a structure characterized by a unique equation of state, which admittedly is not well known in many cases. The situation is quite different in the case of the interiors of planets, which have relatively low pressures and temperatures. Here, not only are the equations of state poorly known and complicated by chemical considerations, but, also, the past history and the nonequilibrium conditions caused by slow reaction rates play an important role. In recent years, our knowledge of the interiors of the outer, or giant, planets has made considerable progress. It is the purpose of this review to provide an overview of these recent developments. The bulk of the paper is devoted to Jupiter because it is the best-known outer planet. There is essentially nothing to say about Pluto. We begin with a brief recapitulation of the foundations of the theory of planetary interiors.

II. THEORETICAL METHODS

Before a suitable model for the interior of a particular planet is proposed and tested, several quantities should be known: radius R, mass M (average density $\bar{\rho}$), chemical composition, gravitational field outside of the planet, oblateness ϵ, and an appropriate equation (or equations) of state. As Table 1 indicates, the outer planets have densities that are much lower than 4- to 5.5-g-cm^{-3} densities of the inner planets. The exception is very poorly known Pluto, whose density may be as high as

Table 1.—Basic data concerning the outer planets. The value of T_{av} corresponds to solar irradiation alone; T_b is the observed brightness temperature. Parameters J and K are explained in text.

Parameter	Jupiter	Saturn	Uranus	Neptune	Pluto
Mass*	317.9	95.1	14.5	17.3	0.2 to 0.8
Radius*	11.19	9.47	3.69	3.92	0.5 to 0.6
Density (g-cm^{-3})	1.334	0.688	1.68**	1.59	7.7
Oblateness	1/15.4	1/9.5	1/18	≈1/51	
J	0.02206 ±0.00022	0.02501 ±0.00003		0.0074 ±0.0007	
K	0.0025 ±0.0014	0.0039 ±0.0003			
Visual bond albedo	0.73	0.76	0.93	0.84	
Balometric albedo	0.45	0.61	0.42	est. 0.42	
T_{av} (K)	105	77	52	41	
T_b (K) (1.5 to 100 μm)	134	97	(20 μm) 55 (1 mm) 220	– (1 mm) 180	
Energy flux (× solar)	2.7	2.4			

*Values for mass and radius are given in multiples of the mass and radius, respectively, of the Earth.
**Recent unpublished results obtained by R. E. Danielson, M. Tomasko, and B. Savage suggest a density for Uranus of 1.2 g-cm^{-3}.

7 g-cm^{-3}. (It should be remembered, however, that not too long ago, its density was believed to be 40 g-cm^{-3}!) It is clear that the other outer planets consist primarily of the lighter elements. Further conclusions can be drawn by estimating the theoretical limits for the central pressure p_c and central density ρ_c:

$$1 \leqslant p_c(8\pi R^4)(3GM^2)^{-1} \leqslant (\rho_c/\rho)^{4/3} , \tag{1}$$

where G is the gravitational constant. The average pressure $\bar{p}$ is equal to $0.4p_c$. If one assumes that the laws of ideal gases apply, then

$$1 \leqslant \bar{T}(5kR)(\mu m_0 GM)^{-1} \leqslant (\rho_c/\bar{\rho})^{1/3} , \tag{2}$$

where $\bar{T}$ is the average temperature, k is the Boltzmann constant, μ is the molecular weight, and m_0 is the atomic mass unit in grams. Furthermore, the following relation must be obeyed:

$$\rho_c > (5I)^{-3/2}(2MR^2)^{3/2}\bar{\rho}\,,$$

where I is the moment of inertia. These limits may indicate which phases and compounds are stable in a particular planet.

Once a choice of elements and phases is made, the appropriate equations of state imply a relation between density ρ, pressure p, and temperature T. Usually, it is first assumed that $T = 0$, $\rho = \phi(p)$, and $p = f(\rho)$ and that the equations of hydrostatic equilibrium hold:

$$\frac{d^2\rho}{dr^2} + \frac{2}{r}\frac{d\rho}{dr} + \left(\ddot{f}/\dot{f} - \frac{1}{\rho}\right)\left(\frac{d\rho}{dr}\right)^2 + 4\pi G\rho^2/\dot{f} = 2\omega^2\rho/\dot{f}$$

and

$$\frac{d^2p}{dr^2} + \frac{2}{r}\frac{dp}{dr} - (\ddot{\phi}/\dot{\phi})\left(\frac{dp}{dr}\right)^2 + 4\pi G\phi^2 = 2\omega^2\phi\,,$$

where G is the gravitational constant, $r < R$ is the radius, and ω is the angular velocity. Later on, corrections for the temperature gradient can be introduced. These corrections require a knowledge of the surface, or atmospheric, temperature and some information about the mechanism of internal heat transport and about the heat sources.

The low densities of the outer planets suggest hydrogen as a major constituent. Figure 1 shows the Gibbs free energy for solid hydrogen, as a function of pressure, for both the molecular and metallic forms.* The existence of metallic hydrogen has been predicted theoretically by Wigner and Huntington (1935). It follows from Figure 1 that, depending upon the details of the calculation and extrapolation procedures, the critical transformation pressure occurs somewhere between 1×10^{12} and 3×10^{12} dyne-cm^{-2}. Usually, the value 2×10^{12} dyne-cm^{-2} is assumed. It should be pointed out, however, that Alder (1961) suggested a transition pressure that is about 10 times higher. His arguments were based on a comparison of interatomic spacings in substances known to have similar molecular-to-metallic transitions. The radius at which this transition pressure exists is very important for a

*C. W. Beckett and R. C. Thompson (National Bureau of Standards), private communication.

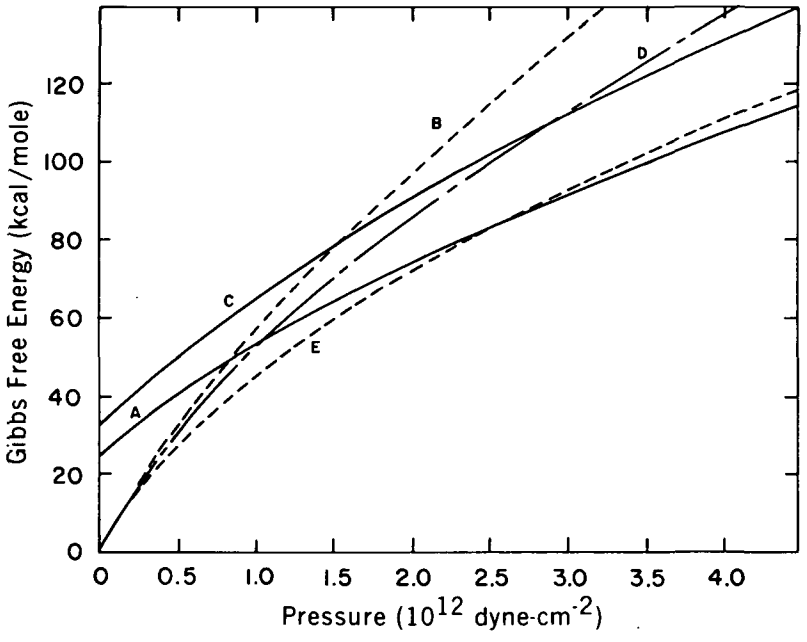


Figure 1.—Gibbs free energy of solid hydrogen (full lines—metallic; dashed lines—molecular) in various approximations. Curve *A*: Wigner and Huntington, 1935; Kronig, de Boer, and Korringa, 1946; March, 1956; Carr et al., 1961. Curve *B*: Buckingham et al., 1957. Curve *C*: Carr, 1962; Carr et al., 1961. Curve *D*: Fisher. Curve *E*: Gordon and Cashion, 1966.

planetary model not only because of the difference in density of the two forms of solid hydrogen, but also because the expected high thermal conductivity in the metallic solid, as opposed to the rather low thermal conductivity in the molecular solid, affects the rate of heat transport and the convection patterns.

Integration of the hydrostatic equations from the surface of the planet radially inward does not provide unique results unless additional factors, such as the total mass M, the oblateness ϵ, and the external gravitational field $V(r)$, are introduced. The external gravitational field can be described by

$$V(r) = -\frac{GM}{r}\left[1 - J_2(R/r)^2 P_2 \sin\theta - J_3(R/r)^3 P_3 \sin\theta - J_4(R/r)^4 P_4 \sin\theta + \ldots\right],$$

where the J_i are multipole gravitational coefficients (which vanish for a sphere), P_i are the appropriate Legendre functions, and θ is latitude. If, as for the Earth, $J_3 \neq 0$, the equatorial plane is not a plane of symmetry, and one can draw conclusions about the rigidity of the planet. The conventional notation for the multipole coefficients is J and K, which are defined by

$$J_2 = (2/3)J = (C - A)/MR^2 ,$$

where A and C are the moments of inertia around the polar and an equatorial axis, respectively, and $J_4 = -(4/15)K$. Coefficients J and K are the ones usually quoted. It is interesting to note that a reliable value of J_3 for Earth was obtained only after analyzing the precession of the perigee and the periodic variation of the eccentricity of artificial satellites with highly inclined orbits (King-Hele et al., 1965). Jupiter's well known Galilean satellites have large orbits with vanishing inclinations. Only the orbit of Amalthea (which is nearest to the planet, has a measurable eccentricity, and has an inclination of about half a degree) could provide information about J_3 and about the rigidity of the planet. Unfortunately, the observations of Amalthea are extremely difficult (Sudbury, 1969), and, as yet, no conclusions can be drawn. The significance of the parameters J_2 and J_4 (or J and K) in the theory of planetary interiors lies in the fact that they are directly related to the radial distribution of density in the outer layers of the planet. Thus, the correct planetary model obtained from the integration of the hydrostatic-equilibrium equations has to yield the observed multipole parameters. Many excellent reviews of the methods of calculating planetary models are available (Wildt, 1961); thus, no further details will be given here.

III. JUPITER

A. Basic Models

It is clear from the brief summary of the theoretical method of evaluating the structure of planetary interiors that the problem of the giant planets would be relatively easy if they contained only hydrogen. However, this is not the case, and the various ways in which the presence of about 10 percent atomic helium and of other elements is taken into account lead to a variety of models. The basic problem

is whether hydrogen and helium occur at a fixed ratio throughout the planet or whether helium is more highly concentrated at lower radii because of gravitational differentiation. This is a very difficult question to answer unambiguously because it involves assumptions about the early history of the planet, as well as about the diffusive and convective transport mechanisms. Peebles (1964) concluded that the diffusive motion of helium (presumably substitutional) in solid metallic hydrogen requires a thermal activitation energy which is much higher than the available kT, and that no differentiation could have taken place in a time of the order of 10^9 years. On the other hand, consideration of the motion of helium as an interstitial atom indicates that the required activation energy could be a factor of 10 smaller than kT, so that the atom could drift under the gravitational acceleration. A free motion with a mean path equal to the distance between interstitial sites would permit gravitational differentiation in a time less than 10^6 years. Although this time is undoubtedly an underestimate, it is possible that some differentiation has taken place. Its degree is difficult to assess, particularly because of the poor knowledge of the velocities in the convection which opposes differentiation.

The first complete model of Jupiter was obtained by De Marcus (1959), who assumed $T = 0$ and interpolated between the equations of state of hydrogen and helium. Nearly all subsequent investigations of Jupiter start with this now classical model, which has been described in many places (Öpik, 1962). De Marcus' results are shown in Figure 2. The phase transition of solid hydrogen occurs at a radius slightly larger than $r/R_j = 0.8$, and it shows up as a discontinuity in pressure, density, and helium content. The central core (which starts at $r/R_j = 0.1$) contains primarily helium, with the addition of heavier elements and perhaps also silicates to obtain a correct total mass of the planet. It is interesting to note that the hydrogen abundance turns out to be about 80 percent by weight, and this percentage is quite insensitive to variations of the equations of state within the limits of the theoretical uncertainty. Peebles (1964) further developed the theory of the Jovian interior by considering in detail the effect of the temperature gradient and the depth of the atmosphere. He assumed that outside of a high-density core, helium and hydrogen occur in a fixed ratio, which implies an exceedingly efficient mixing due to convection. In this respect, his model differs considerably from that of De Marcus. By fitting the model to the known values of J and K and considering the most likely abundance of elements, Peebles concluded that Jupiter has a deep, adiabatic atmosphere and a total hydrogen content close to that obtained by De Marcus. Figure 3 shows the radial dependence of density for isothermal and adiabatic

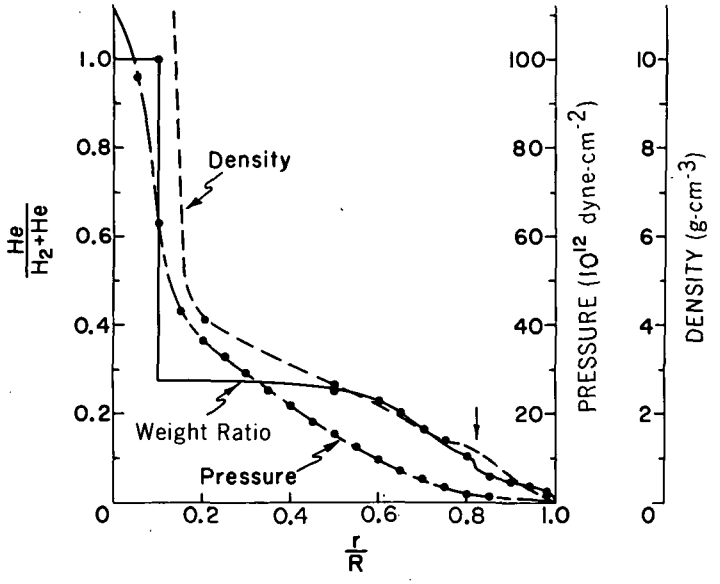


Figure 2.–De Marcus (1959) model of Jupiter.

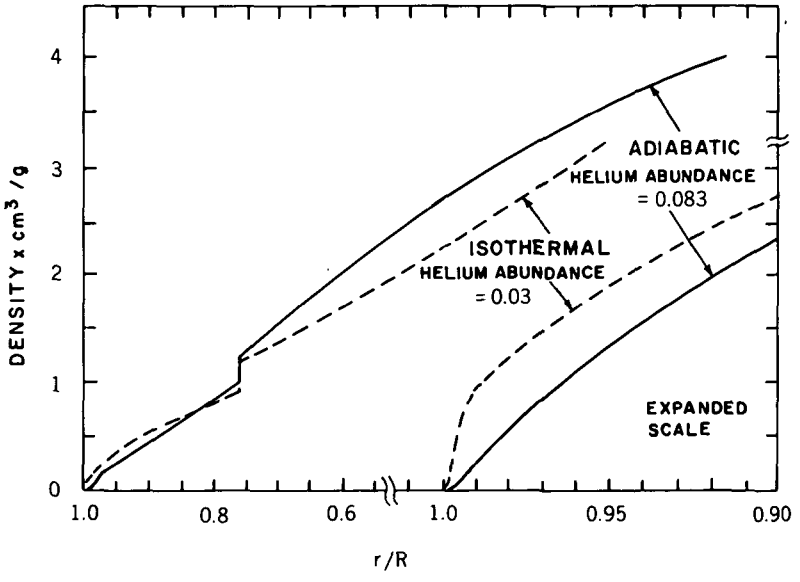


Figure 3.–Radial dependence of density in Jupiter, according to Peebles (1964).

temperature gradients and the helium content chosen to yield the observed value of J. De Marcus (1965) pointed out recently that the uncertainty of our knowledge of the parameter K for Jupiter is greater than the limits imposed by the hydrostatic model, so that only J is of any significance for the models presently considered.

The next step in the analysis of the Jovian interior was the application by Smoluchowski (1967) of physico-chemical arguments in analyzing the hydrogen-helium equilibrium at high pressures and temperatures. Using the known H_2-H_2, H_2-He, and He-He interactions, he concluded that helium will be easily soluble in solid H_2 but essentially insoluble in metallic hydrogen until pressures of about 12×10^{12} dyne-cm^{-2} are reached. At these pressures, helium can become ionized by the surrounding metallic hydrogen, can dissolve in it, and can form a hydrogen-helium alloy. This critical pressure occurs on Jupiter at about half of the planetary radius. Also, he estimated the melting temperatures of the various solids using semiempirical formulae and analogies with other alloy systems (Smoluchowski, 1970a). The result is that the molecular hydrogen layer near the transition pressure to the metallic form and most of the latter is solid, but the hydrogen-helium alloy in the deeper interior of the planet ($r/R_j \lesssim 0.4$) may be liquid. It follows, also, that the transition from solid to liquid molecular hydrogen comes nearest to being the surface of the planet because the existing temperatures and pressures place the fluid molecular hydrogen in a supercritical state in which the transition from the liquid to the gaseous forms is continuous. This does not preclude, of course, the existence of a sharp change in the radial density gradient in the atmosphere at the boundary between the lower, more liquid part, which may contain considerable amounts of dissolved helium, and the upper, more gaseous part, which would be expected to contain much less helium. These physico-chemical considerations of the H-He system play an important role in such phenomena as convection, heat transport, magnetic field, and perhaps the motion of the Red Spot, to be discussed below. In Figure 4, the radial distributions of helium in the various models are compared semiquantitatively. It should be pointed out that the Smoluchowski model has not been subjected to a gravitational analysis, and so the amount of helium contained in the molecular hydrogen layer is not known. Because of convection, there is presumably no radial concentration gradient in the liquid parts of the planet, that is, in the H-He alloy region and in most of the molecular hydrogen layer. The situation is less clear for the "pure" metallic-hydrogen layer in which, according to Figure 4, there is no helium at all. On the other hand, if there is helium in this layer (as in models 1, 2, and 3 in Figure 4), the layer is a two-phase system: pure metallic hydrogen with inclusions of liquid helium, i.e., not a solid solution. All the

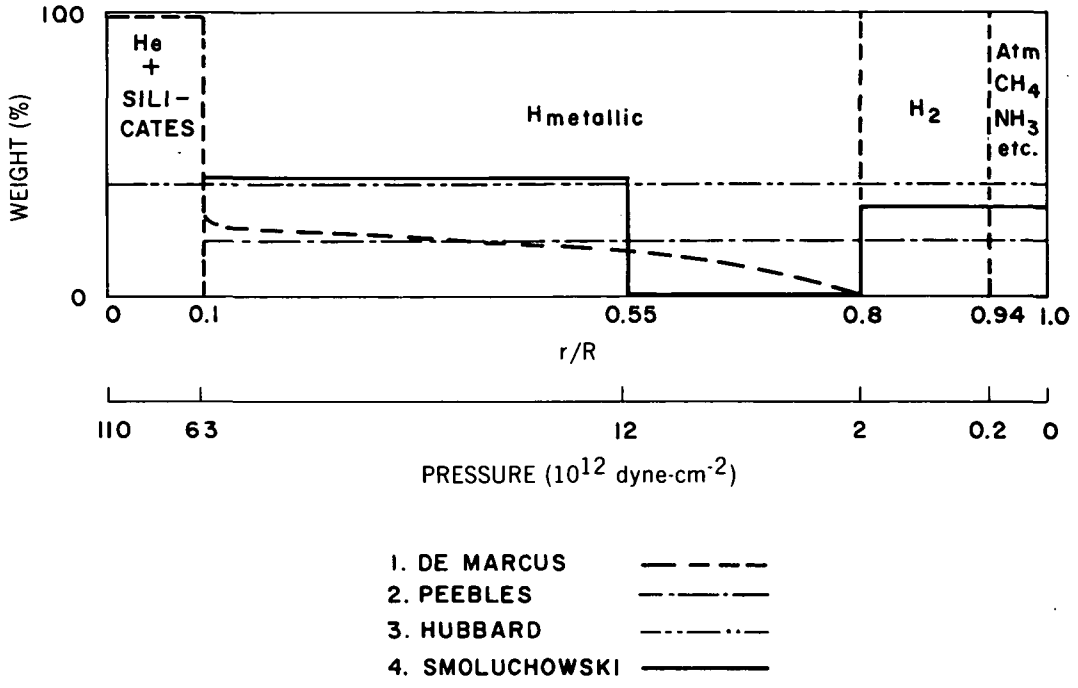


Figure 4.—Radial distribution of helium in Jupiter according to various models: gravitational differentiation in model 1, complete convective mixing in models 2 and 3, absence of mixing in the pure metallic-hydrogen layer in model 4. Model 4 contains 60 percent hydrogen, the other models around 80 percent.

conclusions drawn above concerning the solubility of He in H, melting temperatures, and so forth, are valid in the presence or absence of a concentration gradient.

Most recently, the thermal structure of Jupiter was analyzed in great detail by Hubbard (1968a, 1969, and 1970), who assumed initially the validity at high temperatures of density and pressure profiles used by Peebles. He focused his attention on the theory of metallic hydrogen: its electrical and thermal conductivities, opacities, melting temperatures, and Debye characteristic temperatures at various densities (including the effect of helium). Furthermore, he assumed that the planet is a chemically homogeneous body at all radii, which simplified the model to such an extent that the structure of the planet became overdetermined. As mentioned above, this assumption may not be valid in the solid parts of the interior. Hubbard pointed out that lowering the hydrogen abundance from about 80 percent

to less than 60 percent by weight leads to higher temperatures throughout the planet and to the absence of a central, high-density core. Table 2 shows the radial dependence of temperature for three models calculated by Hubbard. These models differ in many ways, principally in hydrogen content, which in model *a* is near 78 percent whereas in models *b* and *c* is below 60 percent. Also, the equations of state differ in the three models, the most sophisticated equation being used in the preferred model, model *c* (see Figure 5), which includes the effect of electron screening on the vibrational modes of the metallic hydrogen. The huge differences in the calculated central temperature should be noted. In Figure 5, the temperature plots are compared with the corresponding estimated melting temperatures. At the bottom of the atmosphere, all three models give temperatures of the order of 1000 to 2000 K, which is close to those postulated by Peebles. In these calculations, convection is described by means of the mixing-length theory, in which viscosity plays no role, even for temperatures well below melting. Further discussion of these results is given in the next section.

Table 2.–Radial dependence of temperature (in K) in Jupiter (Hubbard, 1968a, 1969, and 1970).

r/R_j	Model		
	a	*b*	*c*
0	4000	14 000	7500
0.1	3920	13 950	7400
0.2	3800	13 500	7200
0.3	3680	13 000	6900
0.4	3400	12 000	6350
0.5	3200	11 000	5780
0.6	2900	10 000	4950
0.7	2500	8 000	4200
0.8	2000	6 400	3600
0.9	1500	4 900	2750

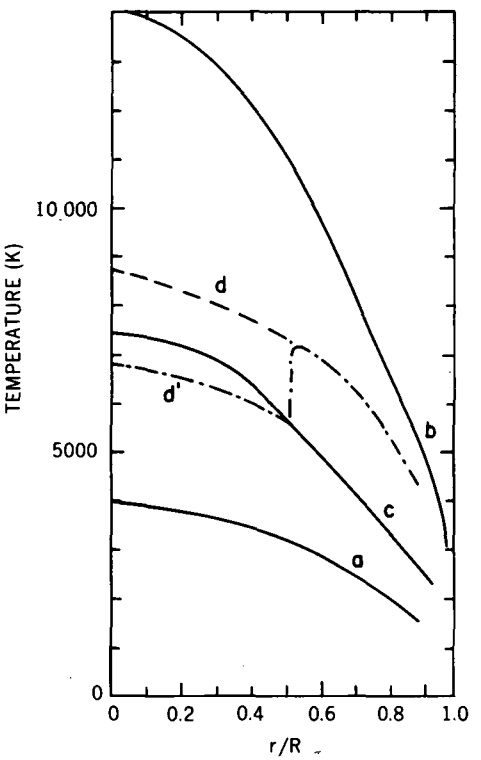


Figure 5.—Radial dependence of temperature in Jupiter according to Hubbard's three models (*a, b, c*) and of the melting temperature for pure hydrogen (*d*) and for a 15-percent (by number) solid solution of helium in metallic hydrogen (*d'*) (Smoluchowski, 1970a; Hubbard, 1968a, 1969, and 1970).

B. Heat Emission

It has been pointed out by F. Low (1966 and 1969) that Jupiter emits about 2.7 times more radiation than it receives from the Sun. Five possible sources of this extra 10^{33} ergs per year have been considered (Smoluchowski, 1967): natural radioactivity, nuclear reactions, change in the gravitational constant G, primordial heat, and gravitational shrinkage. The first source turns out to be several orders of magnitude too low, considering the ratio of radioactive to nonradioactive elements in the solar composition, which is expected to be similar to the Jovian composition. The second source requires internal temperatures one or two orders of magnitude

higher than any theory can justify. The third source, suggested by Dicke,* is based on the conclusion that the gravitational constant G is decreasing at the rate of $\dot{G}G^{-1} = 10^{-11}$ per year. Such a decrease would speed up the rotation of the central metallic, denser part of Jupiter with respect to the less dense outer molecular layer. The resulting frictional heat could be the source of the extra energy. Figure 6 shows the expected change in the angular velocity ω as a function of the planetary radius if there were no friction. Figure 7 shows the radial dependence of the force in the equatorial plane required to keep ω constant in the whole planet. Both plots show discontinuities at the boundary of the inner core ($r/R_j = 0.1$) and at the boundary of the metallic mantle ($r/R_j = 0.8$) where sudden jumps of density occur. Assuming that the rate of change of G is constant and that the Jovian interior has not changed drastically since its origin some 4×10^9 years ago, we can calculate the total energy

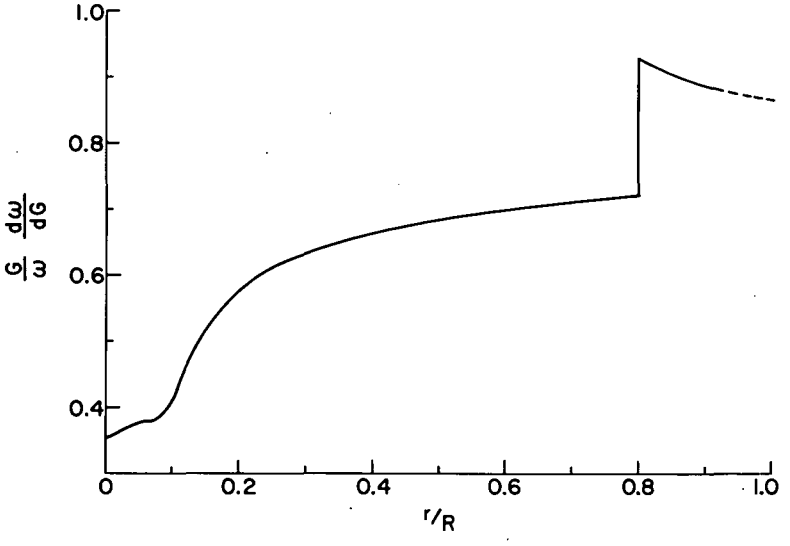


Figure 6.—Change of angular velocity caused by a decrease of the gravitational constant at various radii in Jupiter, with the assumption that there is no friction between layers.

*Private communication.

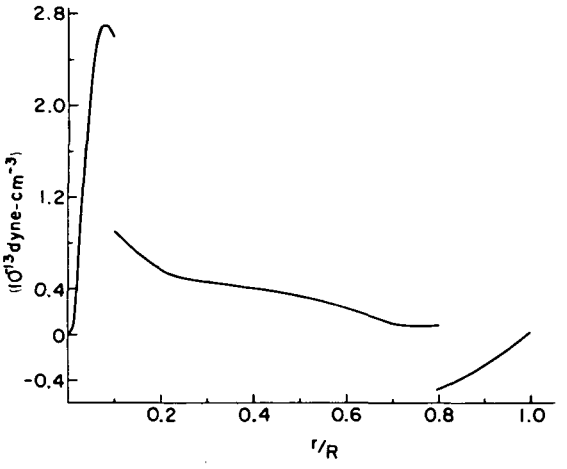


Figure 7.—Force in the equatorial plane of Jupiter necessary to keep the angular velocity constant for $\dot{G}G^{-1} = 10^{-11}$ per year (from R. H. Dicke, private communication).

evolved and its rate of production. It turns out that this effect also is too small by a few orders of magnitude.

The fourth possibility, a slow leakage of the primordial heat, has been discussed in detail by Hubbard (1968a, 1969, and 1970) and later by Bishop and De Marcus (1970). The calculated cooling times depend critically on the assumed surface flux and on the efficiency of heat convection and conduction in the Jovian interior. In support of this primordial source is the fact that the ratio of heat emission to heat received from the Sun is, within 10 percent, the same for Jupiter and for Saturn, which emits 2.4 times as much energy as it receives (Hubbard, 1968b). On the other hand, the vastly different structure and temperature profiles of the interiors of the two planets make it quite unlikely that the heat transport in both planets has always been equally efficient. Figure 8 illustrates the dependence of the conductivity of metallic hydrogen on its density, according to various theoretical approximations (Bishop and De Marcus, 1970). The range of variation is one to two orders of magnitude. Considerable uncertainty arises also from poor knowledge of the influence of dissolved helium on the conductivity of pure hydrogen. With various assumptions concerning these quantities, Hubbard obtained cooling times of the order of 10^9 to 10^{12} years. In any case, it is certain that if Jupiter were all

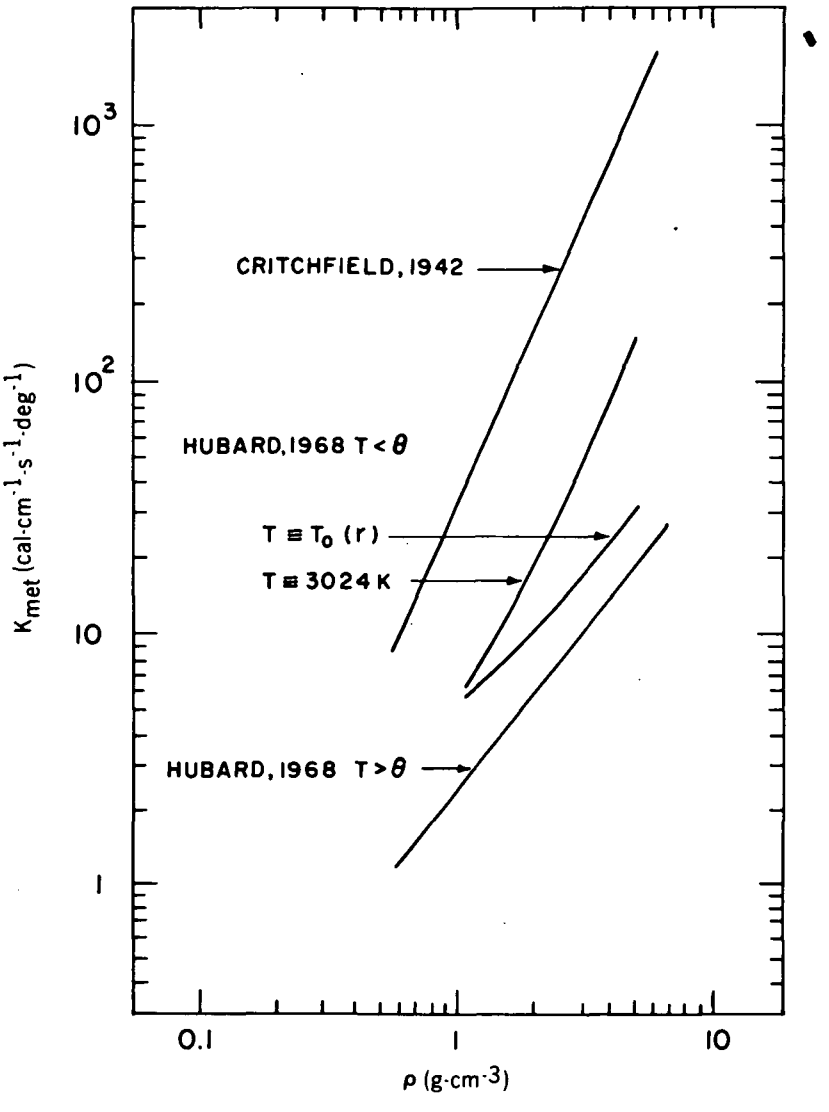


Figure 8.—Thermal conductivity of metallic hydrogen as a function of density, according to various theories (Bishop and De Marcus, 1970).

conductive, the observed rate of heat emission would require central temperatures of the order of 10^5 K; thus, there would be no solids. On the other hand, according to Bishop and De Marcus, a highly conductive core and a convective mantle would be compatible with observation. Figure 9 presents an example of cooling curves calculated by these authors. Hubbard's most recent model (curve *c* in Figure 5) yields a net heat flux that is marginally consistent with observation. The presence of an additional energy source may have to be considered.

The fifth possibility, a gradual gravitational shrinkage of the planet, has been proposed by many investigators (Gross and Rasool, 1964) and is based on the simple observation that a radial growth of the denser metallic-hydrogen layer at the expense of the surrounding, less dense molecular-hydrogen layer leads to a net shrinkage of the planet. The gravitational shrinkage produces a compression of the deeper layers of the planet and a consequent emission of heat. It turns out that this phase change, proceeding at the rate of only 1 mm per year, would release the correct amount of gravitational energy. It has been shown (Smoluchowski, 1967) that although part of the gravitational energy is stored as an elastic compression, there is enough energy released to account for the additional heat emission. Furthermore, a radial rate of phase change of solid hydrogen that is enough to account for the heat flux of Jupiter accounts equally well for the same phenomenon on Saturn (Smoluchowski, 1970a). Two questions remain: What controls the rate of phase change, and why is it the same on both planets? The answers to these questions are not known conclusively, although it has been suggested (Smoluchowski, 1967) that the rate is actually controlled by helium diffusion. As discussed above, solid molecular hydrogen undoubtedly contains an appreciable amount of dissolved helium, whereas no helium can be dissolved in the solid metallic-hydrogen phase at a pressure of about 2×10^{12} dyne-cm^{-2}. Thus, in order for the phase transition to proceed, helium has to be gradually eliminated from the interphase boundary layer by outward diffusion into the remainder of the molecular mantle. An estimate of this diffusion rate obtained by extrapolation from laboratory data is, within an order of magnitude, in agreement with the rate of 1 mm per year. In contrast to the considerable differences in the deeper interiors, the outer molecular layers of Jupiter and of Saturn are very likely similar (Peebles, 1964), so it is not unreasonable that the rates of diffusion and of phase change are also similar.

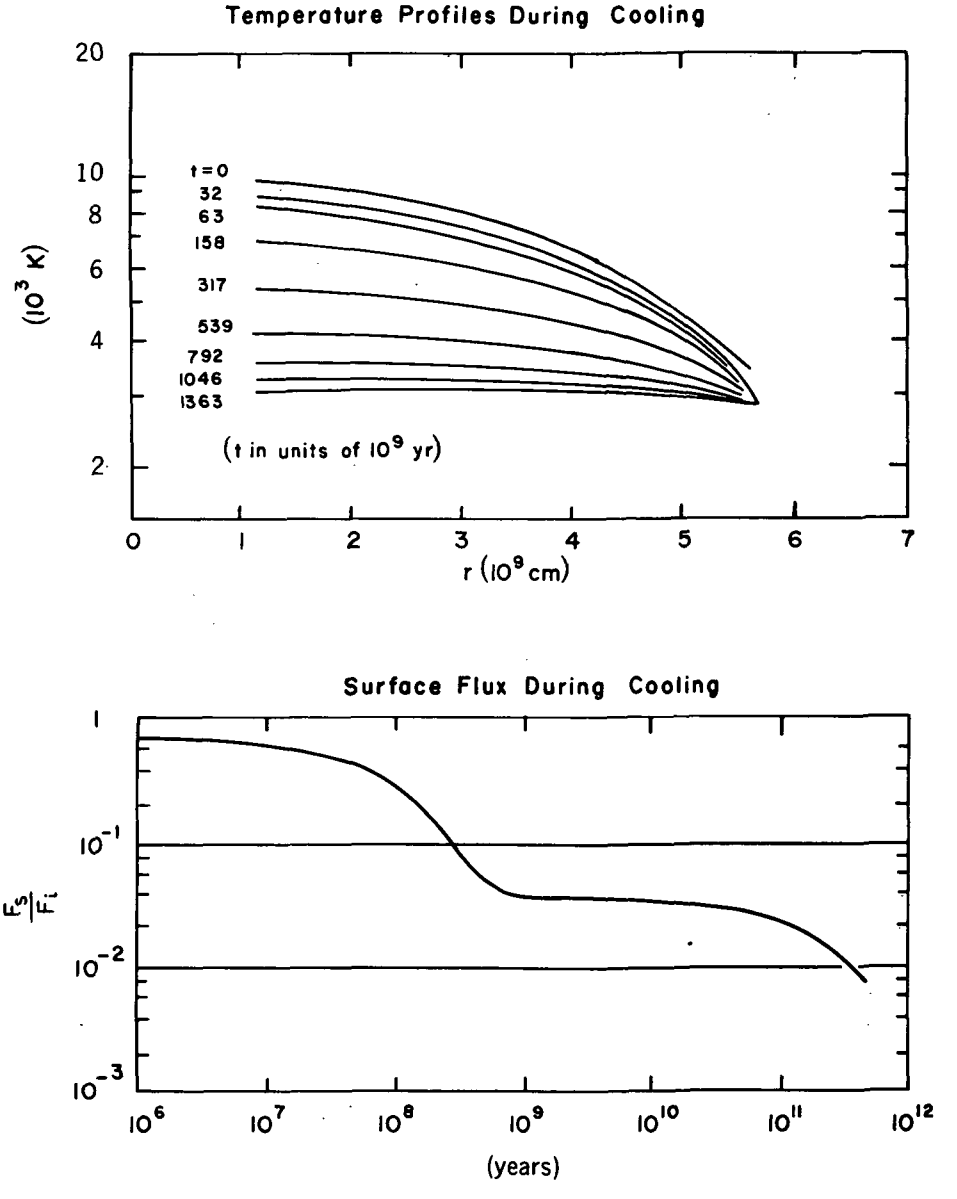


Figure 9.—A typical temperature profile and drop of the surface flux F_s of Jupiter as a function of time. F_i is surface flux for insolation equilibrium temperature T_{av} = 105 K (see Table 1 and Bishop and De Marcus, 1970, for details of other assumptions).

C. Convection

Convection in a rotating planet is controlled by the dimensionless Rayleigh number

$$\mathcal{R} = g\alpha\rho^2 C_p^4 \lambda^{-1}\eta^{-1} \operatorname{grad} T$$

and the Taylor number

$$\mathcal{T} = 4\Omega^2\rho^2 d^4\eta^{-2} ,$$

where g is gravitational acceleration, α is thermal expansion, ρ is density, C_p is specific heat, λ is thermal conductivity, η is viscosity, and Ω is angular velocity. Although $\mathcal{R}$ is independent of $\mathcal{T}$ at low values of $\mathcal{T}$, it increases as $\mathcal{T}^{2/3}$ for $\mathcal{T}$ greater than 10^2 to 10^3. The problem of the actual convective pattern and convective velocities on Jupiter is complicated by an only approximate knowledge of the transport coefficients λ and η and by the simplifying assumptions necessary to obtain an answer even when high-speed computers are used. It is known from the work of Herring (1950) and others that at sufficiently high temperatures and in a suitable range of stress, solids behave in a viscous manner, and the viscosity can be evaluated from the known self-diffusion coefficient. Turcotte and Oxburgh (1967) have shown that this kind of viscosity accounts very well for cellular convection in the Earth's mantle, for its heat flow, and for the observed continental drift velocity given by

$$u = 0.142 R^{2/3} d^{-1}\rho^{-1} C_p^{-1} ,$$

where d is the thickness of the convective layer. Smoluchowski (1970b and 1970c), using high-temperature and high-pressure extrapolations of experimentally determined self-diffusion in solid hydrogen, applied the same kind of argument to convection in the molecular-hydrogen layer of Jupiter. The resulting η is of the order 10^{18} stokes, which brings $\mathcal{T}$ down to the range where $\mathcal{R}$ is independent of $\mathcal{T}$, making the calculation more reliable. Thermal conductivity of molecular solid hydrogen was evaluated through the use of various formulae, among them the Leibfried-Schlömann formula

$$\lambda \approx 5 \times 10^{-8} A a \theta^3 T^{-1}\gamma^{-1} \text{ watt units} ,$$

where A is the mean atomic weight, a is the lattice constant in angstroms, γ is the Grüneisen constant, and θ is the Debye temperature. With $\alpha \approx 10^{-5}$, $C_p \approx 10^8$ to 10^9, $\rho \approx 1$, $d \approx 5 \times 10^7$, $\lambda \approx 10^7$ and $\eta \approx 10^{18}$ (in cgs units), one obtains $u = 10^{3 \pm 1}$ Å-s^{-1} in the solid molecular layer of hydrogen.

In contrast, in the liquid hydrogen layer, the dependence of $\mathcal{R}$ and $\mathcal{T}$ has to be taken into account if the Turcotte and Oxburgh analysis is to be used, and this introduces considerable uncertainties. Fortunately, good theories of so-called dense liquids exist, and they permit obtaining rather trustworthy values for the various pertinent physical quantities. With $\alpha \approx 1.5 \times 10^{-4}$, $\rho \approx 1$, $d \approx 10^7$, $\lambda \approx 3 \times 10^6$, and $\eta \approx 2 \times 10^{-2}$, one obtains $u \approx 10^{1 \pm 1}$ cm-s^{-1}. The corresponding heat fluxes in the solid and in the liquid hydrogen layer, calculated from the formula $0.167\, \lambda\, \mathcal{R}^{1/3}$ grad T, are (within a factor of 10) 10^4 and 10^5 erg-cm^{-2}-s^{-1}, respectively. This compares favorably with the observed flux of 10^3 to 10^4 erg-cm^{-2}-s^{-1}.

The fact that the H-He alloy in the deep metallic interior of Jupiter ($r/R_j < 0.4$) may be liquid, chemically homogeneous, and convective provides an attractive explanation (Smoluchowski, 1970a) for the origin and location of the huge magnetic field of this planet (10 to 50 gauss at the poles): The estimated electrical conductivity $\sigma \approx 5 \times 10^{-6}$ emu is of just the right order of magnitude to permit the operation of a hydromagnetic dynamo that is driven by convection. This is presently the favored mechanism for the generation of planetary magnetic fields.

D. The Red Spot

The Red Spot of Jupiter presents many puzzling problems: the nature of the anomaly, the stability of its zenographic latitude, the great variation of its zenographic longitude, the almost constant absolute magnitude of its eastward or westward velocity, and finally, the small, fairly periodic wobble of this velocity. Without going into the pros and cons of the various proposals (Sagan, 1971), it can be said that at present, a combination of the theories of Hide (1963 and 1969) and Streett (1969) seems to provide the best model. According to Hide, a large perturbation at the bottom of a deep atmosphere of a planet can be, under favorable circumstances, the foot of a so-called Taylor column. Such a column, which is a dynamic anomaly reaching to the top of a rotating fluid, is parallel to the axis of rotation and has been observed under suitable laboratory conditions. What we see on Jupiter, according to Hide, is the top of a Taylor column generated by some irregularity in the deeper layers. If this irregularity were attached to the solid part of the planet, the motion of the Red Spot would indicate a continuous and variable

exchange of angular momentum between the liquid core and the solid mantle. Runcorn's (1965) analysis of this mechanism suggests that in order to fit the observations, the radius of the liquid core would have to be a much larger fraction of the radius of the solid mantle than any of the present models permit. It seems, therefore, more likely that the irregularity is a perturbation on the top of the more liquid part of the supercritical atmosphere. This possibility is strongly supported by a recent physico-chemical study by Streett of various two-phase diagrams involving noble gas. He suggested that at very high pressures, solid molecular hydrogen containing some dissolved helium is in equilibrium with solid helium containing some dissolved hydrogen, as is shown in Figure 10. At lower pressures, however, helium becomes liquid, and there is an equilibrium between solid hydrogen

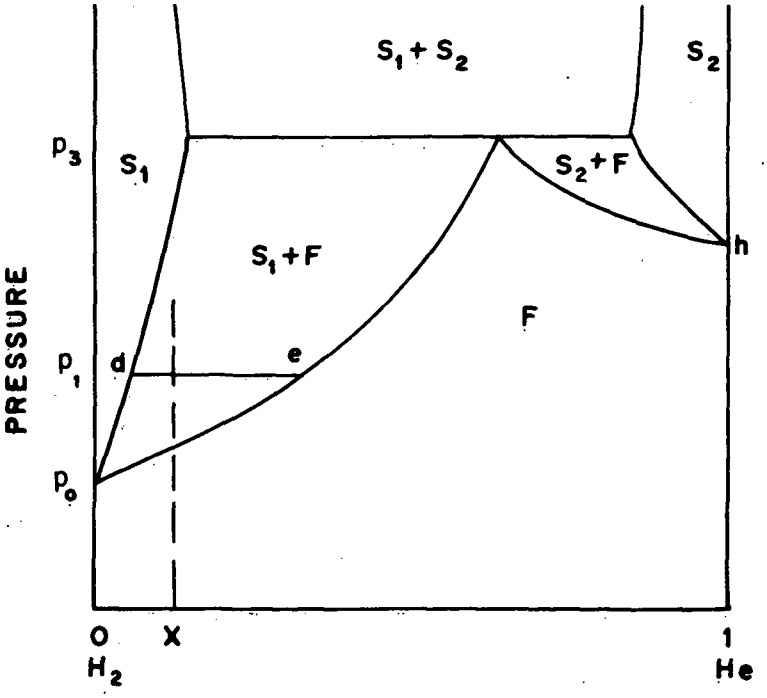


Figure 10.—The H_2-He phase diagram as a function of pressure; S_1 and S_2 are regions of solid solubility, and F is a region of fluid solubility. At composition X and pressure p_1, the solid solution d is in equilibrium with fluid e (Streett, 1969).

containing some dissolved helium and liquid hydrogen containing a large amount of dissolved helium. As a result, at an appropriate pressure level in a pressure gradient, the solid will float in the liquid. Such an island can thus act as the foot of Hide's Taylor column and still have considerable mobility. The motion of such an island would be controlled by convection patterns in the liquid hydrogen layer which are probably only weakly coupled to convective patterns in the more gaseous layer. The motion of the island may be affected, however, by a local anomaly of the huge central magnetic field of the planet or by a marginally probable field generated in the liquid molecular layer itself (Smoluchowski, 1970a). As mentioned above, no satisfactory theory of these motions has yet been obtained, although several equatorial or near-equatorial phenomena such as the belt of rapid eastward motion of clouds have been analyzed (Gierasch and Stone, 1968). Although the Hide-Streett model of the Red Spot still leaves many questions unanswered, it is reasonably compatible with some of the major features of this phenomenon. Actually, the convective velocities deduced above for the highly simplified model are indeed close (Smoluchowski, 1970b and 1970c) to the observed rate of longitudinal motion of the Red Spot of about 10^2 cm-s^{-1}. This may be fortuitous, however.

IV. SATURN

Although Saturn, the next largest planet, is in many ways quite similar to Jupiter, it is sufficiently different to provide very important additional checks on the various assumptions and theoretical models of the interiors of both planets. De Marcus (1959), Peebles (1964), and Hubbard (1970) included Saturn in some of their theoretical studies. Its lower density results from a mass more than three times smaller than that of Jupiter. This implies the existence of much lower pressures. According to these theories, Saturn contains much less hydrogen than Jupiter: about 70 percent less, according to De Marcus and Peebles, and about 30 percent less, according to Hubbard. It should be pointed out that although the gravitational multipole coefficients J and K are known for Saturn with much higher precision than for Jupiter (J is tenfold and K nearly a hundredfold more precise), they pertain to the gravitational field of Saturn including its ring. Thus, the analysis is less certain than desired.

A glance at the radial dependence of density, pressure, and fractional helium content shown in Figure 11 indicates that Saturn's interior is qualitatively similar to that of Jupiter, but it differs principally in two respects: First, the phase change

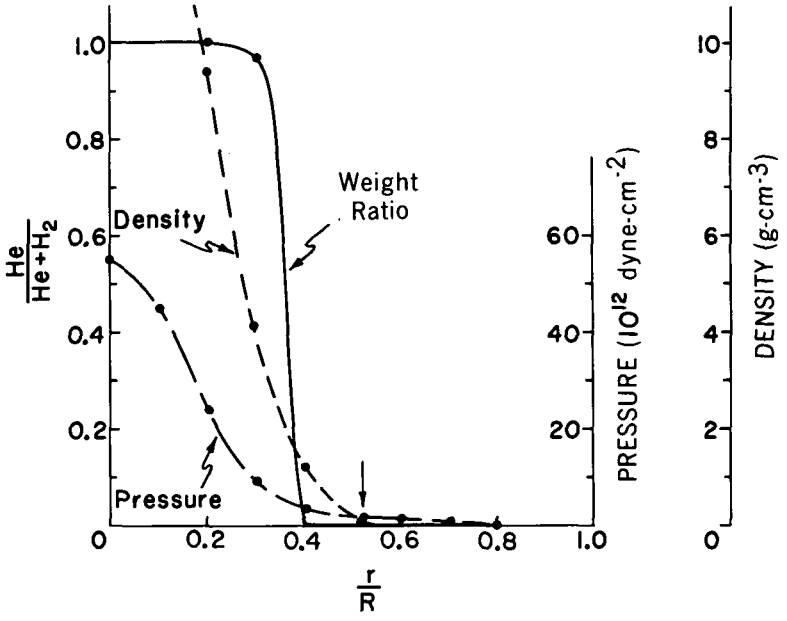


Figure 11.–De Marcus (1959) model of Saturn.

between molecular and metallic solid hydrogen occurs at a relatively (and absolutely) smaller radius ($r/R_s = 0.52$); and second, the helium-rich central core has a radius relatively (and absolutely) much larger than on Jupiter ($r/R_s \approx 0.35$). Peebles pointed out that the atmospheres of both planets are deep adiabatic, and probably quite similar.

The convective velocities in the interior of Saturn have not yet been theoretically evaluated. However, it is known from visual observation that, as on Jupiter, there are fast equatorial currents and a zonal cloud pattern. Thus, the convection may be qualitatively similar in the two planets. It should be pointed out, however, that recent observations (De Marcus, 1970) suggest that the angular velocity of the cloud system of Saturn decreases so rapidly with increasing latitude that an hour's difference in the rotation period exists between polar areas and the equator. If this enormous gradient of angular velocity is indeed confirmed, it would imply profoud differences in the convective pattern of Saturn and Jupiter.

As mentioned previously, Saturn emits 2.4 times more radiation than it receives from the Sun (Low, 1966 and 1969), and as in the case of Jupiter, this extra energy

can be accounted for by a 1-mm-per-year increase of the radius (Smoluchowski, 1970a) of the metallic-hydrogen layer at the expense of the molecular-hydrogen layer. A major difference between the two planets results from the much lower pressures and temperatures on Saturn: Helium on that planet is not soluble in the metallic hydrogen at any radius, and the temperatures are too low for either metallic helium or metallic hydrogen to be liquid. As a result, there is no chance for a hydromagnetic dynamo (Smoluchowski, 1970a) to operate in the Saturnian interior, and no strong magnetic field is generated there. This is not to say that a weak magnetic field may not be generated in the lower, liquid part of the atmosphere, although there seem to be considerable difficulties in postulating that liquid hydrogen, even in an impure state, can have sufficient electrical conductivity.

V. URANUS AND NEPTUNE

Our uncertainty concerning some of the salient features of the planets increases with increasing distance from the Sun. For instance, the radius of Neptune was believed to be 2.22×10^9 cm, and that of Uranus, 2.38×10^9 cm, which gave them densities of 2.25 and 1.60 gm-cm^{-3}, respectively. Recent, more precise occultation data (Dollfus, 1967; Kovalevsky and Link, 1969) indicate that the radii of both planets are near 2.5×10^9 cm, which lowers the average density of Neptune to 1.68 gm-cm^{-3} and that of Uranus to 1.44 gm-cm^{-3}. Clearly, such an enormous error in the average density affects considerably our ideas about the composition and structure of these planets. Since, at present, only preliminary results based on the newer density data are available, it is worth describing the earlier models developed by Reynolds and Summers (1965). These authors assumed a somewhat modified solar abundance of the elements and concluded that at temperatures expected in the planetary interiors, the only major constituents in elemental form are H_2, He, Ne, and A. Furthermore, Si and Mg occur as oxides, and the other main compounds are H_2O, CH_4, NH_3, and H_2S. The multiplicity of these elements and compounds makes a precise analysis of the planetary interior impossible, and it illustrates clearly the increasing complexity of the problem with the lower temperatures and pressures mentioned earlier. The authors divided these materials into three categories according to their melting points: H, He, and Ne; solids, or "ice" (H_2O, CH_4, NH_3, H_2S, and A); and "rock" (SiO_2, MgO, and metallic Fe and Ni, all taken in suitable proportion). Three models of the planetary interior were considered: (1) a

completely homogeneous model, (2) a two-shell model in which a core of proper mixture of rock and ice is surrounded by H, He, and Ne; and (3) a three-shell model consisting of an inner rock core, an ice mantle, and an outer H, He, and Ne layer. The calculated density profiles for the two planets are shown in Figure 12. Table 3 gives some of the calculated parameters, including the multipole gravitational coefficient J, and Table 4 shows the corresponding compositions. The large

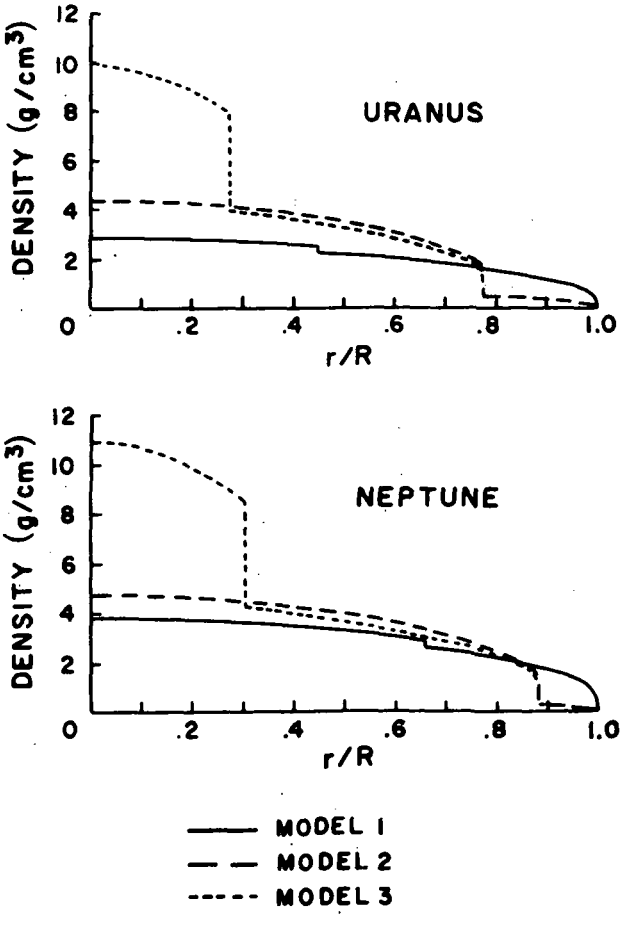


Figure 12.–Reynolds and Summers (1965) models of Uranus and Neptune. Models 1, 2, and 3 are described in the text.

Table 3.—Calculated physical parameters (p_c, central pressure in 10^{12} dyne-cm^{-2}; ϵ, surface ellipticity; and J, gravitational coefficient) based on older (in parentheses, Reynolds and Summers, 1965) and newer radii.*

Planet	Model	p_c	ϵ	J
Neptune	1	(5.51) 3.76	(0.0196) 0.0272	(0.0099) 0.0145
	2	(7.45) 8.0	(0.0165) 0.0195	(0.0068) 0.0068
	3	(13.38)	(0.0155)	(0.0058)
Uranus	1	(3.13)	(0.0584)	(0.0283)
	2	(5.90) 6.36	(0.0439) 0.0488	(0.0138) ≈0.001
	3	(10.28)	(0.0423)	(0.0122)

*R. T. Reynolds and A. L. Summers, private communication.

Table 4.—Calculated composition based on older (in parentheses, Reynolds and Summers, 1965) and on newer radii.* See the text for definitions.

Planet	Model	Ice	Rock	H, He, and Ne
Neptune	1	(0.807) 0.639	(0.108) 0.156	(0.085) 0.205
	2	(0.848) 0.702	(0.114) 0.171	(0.038) 0.127
	3	(0.848)	(0.114)	(0.038)
Uranus	1	(0.730)	(0.098)	(0.172)
	2	(0.790) 0.672	(0.106) 0.164	(0.104) 0.164
	3	(0.790)	(0.106)	(0.104)

*R. T. Reynolds and A. L. Summers, private communication.

difference between the densities of the two planets, which have essentially identical radii, is expressed as a difference in the proportions of H_2, He, and Ne. Since for Neptune $J = 0.074 \pm 0.0007$, one can conclude that model 2 is more satisfactory than the other two models. No data for Uranus exist that would permit a choice to

be made among the three models. The authors suggest that the most likely mechanism for the differentiation of the planetary composition is the loss of volatile materials during the accretion process.

The new, larger radii and lower densities alter the picture to the extent summarized in Table 3. Although these results, also obtained by Reynolds and Summers,* are preliminary and incomplete, it seems that again for Neptune, model 2 is in the best agreement with the observed value of J. For Uranus, the new calculation using model 2 yields $J \approx 0.001$. The principal consequence of the lower densities is an increase in the percentage of H, He, and Ne. The amount of free and combined hydrogen has been increased by about 10 percent by mass, as compared to the earlier models based on the higher densities.

ACKNOWLEDGMENTS

The author is indebted to Drs. C. W. Beckett, R. H. Dicke, W. B. Hubbard, R. T. Reynolds, C. Sagan, A. L. Summers, and R. C. Thompson for unpublished data.

REFERENCES

Alder, B. J., *Progress in Very High Pressure Research,* F. P. Bundy, ed., John Wiley and Sons, New York, 1961.

Bishop, E. V., and De Marcus, W. C., *Icarus* 12:317, 1970.

Buckingham, R. A., Davies, A. E., and Davies, A. R., *Proceedings of the Conference on Thermodynamic and Transport Properties of Fluids,* London, 1957.

Carr, W. J., Jr., Caldwell-Horsfall, R. A., and Fein, A. E., *Phys. Rev.* 124:747, 1961.

Carr, W. J., Jr., *Phys. Rev.* 128:120, 1962.

De Marcus, W. C., in *Handbuch der Physik,* D. Flügge, ed., J. Springer, Berlin, 1959, Vol. 52.

De Marcus, W. C., in *Magnetism and Cosmos,* Oliver & Boyd, Ltd., Edinburgh, 1965, p. 352.

De Marcus, W. C., *Proceedings of the XIVth General Assembly of the IAU,* Brighton, England, August 1970.

Dollfus, A., *Icarus* 12:101, 1967.

Fisher, B. B., U.S. Atomic Energy Commission Report LA-3364.

Gierasch, P. J., and Stone, P. H., *J. Atmos. Sci.* 25:1169, 1968.

Gordon, K. G., and Cashion, J. K., *J. Chem. Phys.* 44:1190, 1966.

Gross, S. H., and Rasool, S. I., *Icarus* 3:311, 1964.

Herring, C., *J. Appl. Phys.* 21:437, 1950.

*Private communication.

Hide, R., *Mem. Soc. Roy Sci. Liège* 7:481, 1963.
Hide, R., *J. Atmos. Sci.* 26:841, 1969.
Hubbard, W. B., *Astrophys. J.* 152:745, 1968a.
Hubbard, W. B., *Astron. J.* 73:S100, 1968b.
Hubbard, W. B., *Astrophys. J.* 155:333, 1969.
Hubbard, W. B., *Astrophys. J.* 162:687, 1970.
King-Hele, D. G., Cook, G. E., and Scott, Diana W., *Planet. Space Sci.* 13:1213, 1965.
Kovalevsky, J., and Link, F., *Astron. Astrophys.* 2:398, 1969.
Kronig, R., de Boer, J., and Korringa, J., *Physica* 12:245, 1946.
Low, F. J., *Astron. J.* 71:391, 1966.
Low, F. J., *Astrophys. J.* (*Letters*) 157:L69, 1969.
March, N. H., *Physica* 22:311, 1956.
Öpik, E. J., *Icarus* 1:200, 1962.
Peebles, P. J. E., *Astrophys. J.* 140:328, 1964.
Reynolds, R. T., and Summers, A. L., *J. Geophys. Res.* 70:199, 1965.
Runcorn, S. K., in *Magnetism and Cosmos,* Oliver & Boyd, Ltd., Edinburgh, 1965, p. 365.
Sagan, C., *Comments Astrophys. Space Phys.* 3:65, 1971.
Smoluchowski, R., *Nature* 215:691, 1967.
Smoluchowski, R., *Proceedings of the XIVth General Assembly of the IAU,* Brighton, England, August 1970a.
Smoluchowski, R., *Science* 168:1340, 1970b.
Smoluchowski, R., *Phys. Rev. Lett.* 25:693, 1970c.
Streett, W. B., *J. Atmos. Sci.* 26:924, 1969.
Sudbury, P. V., *Icarus* 10:116, 1969.
Turcotte, D. L., and Oxburgh, E. R., *J. Fluid Mech.* 28:29, 1967.
Wigner, E., and Huntington, H., *J. Chem. Phys.* 3:764, 1935.
Wildt, R., in *Planets and Satellites,* G. P. Kuiper and B. M. Middlehurst, eds., University of Chicago Press, Chicago, 1961, p. 159.

CHAPTER 9
RADAR AND RADIO EXPLORATION OF THE PLANETS

Arvydas J. Kliore
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

I. INTRODUCTION

In little more than one decade since the first radio detection of another planet, the science of radio and radar astronomy has progressed with unprecedented speed. The construction of new antennas and facilities, as well as the advent of radio occultation measurements from spacecraft, has provided great amounts of original information on the atmospheres and surfaces of the planets. It is clearly not possible to do justice to all of the radio- and radar-astronomical work of the past decade in one paper; therefore, this chapter is limited to a review of the more important results achieved by the passive radio, radar, and radio occultation methods. For brevity, the Sun and the Moon are excluded from discussion, and Jupiter is excluded because of the unique nature of its radio emission. A reasonably comprehensive reference list and bibliography of publications dealing with radio and radar astronomy are found at the end of the chapter; the reader is encouraged to fill in the gaps in the written material through independent reading.

II. PASSIVE RADIO OBSERVATIONS

Basically, the purpose of passive radio-astronomical work (Kraus, 1966) is to measure the temperature of an object. An antenna with a very narrow beam width is trained on a source, and the temperature of the source is deduced from the power that appears at the antenna terminals. One measures primarily the brightness B in $W\text{-}m^{-2}\text{-}Hz^{-1}\text{-}rad^{-2}$. If the sky has a brightness B at a position $d\Omega$ and a zenith angle

θ, the infinitesimal power incident on a surface area dA from a solid angle $d\Omega$ is given by

$$dW = B \cos\theta \, d\Omega \, dA \, d\nu ,$$

where $d\nu$ is the infinitesimal element of bandwidth. The total power received (in watts) is

$$W = \frac{1}{2} A_e \int_{\nu}^{\nu+\Delta\nu} \left[\int_{\Omega_s} B(\theta,\phi) Q(\theta,\phi) \, d\Omega \right] d\nu ,$$

where A_e is the effective aperture of the antenna (m^2), Ω_s is the solid angle subtended by the source (rad^2), Q is the nondimensional response of the antenna, or the antenna pattern function, and the quantity in brackets is the spectral power w, in $W\text{-}Hz^{-1}$.

By Planck's law, the brightness of a blackbody radiator at temperature T and frequency ν is given by

$$B = \frac{2h\nu^3}{c^2} \left(\frac{1}{e^{h\nu/kT} - 1} \right)$$

Integrating over all frequencies yields the total brightness B_T (in $W\text{-}m^{-2}\text{-}rad^{-2}$):

$$B_T = \int_0^{\infty} B \, d\nu = \sigma T^4 ,$$

which is the Stefan-Boltzmann relation, where

σ = Stefan-Boltzmann constant

and

T = temperature of the source.

At radio wavelengths, $h\nu \ll kT$, and B can be approximated by the Rayleigh-Jeans law because

$$e^{h\nu/kT} - 1 \approx h\nu/kT ;$$

i.e.,

$$B = (2h\nu^3/c^2)(kT/h\nu) = 2kT/\lambda^2 .$$

Thus, it can be seen that at radio wavelengths, the brightness of an object at a temperature T varies inversely as the wavelength squared. Therefore, from an observed brightness, one can determine the temperature.

For a source subtending a solid angle Ω_s, the source flux density (in $\text{W-m}^{-2}\text{-Hz}^{-1}$) is given by

$$S = 2kT\Omega_s/\lambda^2 .$$

If T varies over Ω_s, an integral must be used to express the source flux density.

Let us discuss the concept of antenna temperature. A resistor of resistance R will produce at its terminals a noise power per unit bandwidth (in W-Hz^{-1}) of

$$w = kT .$$

If the resistor is replaced by a lossless matched antenna of radiation resistance R placed inside a blackbody enclosure at a temperature T, the brightness B_c will be constant in all directions and will be given by the Rayleigh-Jeans law:

$$B_c = 2kT/\lambda^2 .$$

Then, the spectral noise power at the terminals will be

$$w = \frac{1}{2}A_e \int_{\Omega_s} BQ\, d\Omega = (kT/\lambda^2)A_e\Omega_A = kT ,$$

where Ω_A is the effective beam area of the antenna (rad^2). [By definition of the quantities A_e and Ω_A, $A_e\Omega_A = \lambda^2$ (Krauss, 1966).] This is seen to be the same spectral noise power as in the case of the resistor.

It is important to note that it is not the temperature of the antenna structure that determines the temperature of the antenna radiation resistance; rather, it is determined by the temperature of the region that the antenna "sees" in its reception pattern. Thus, if an antenna beam is pointed at a section of "sky" at a temperature T, its radiation resistance will be at a temperature T, and the spectral power at the antenna terminals will be $w = kT$. In this sense, the radio telescope, consisting of antenna and receiver, may be considered to be a radiometer for the measurement of the temperatures of distant objects which are coupled to the telescope system through the radiation resistance of the antenna. The temperature of this radiation resistance is called the antenna temperature—a very important concept in radio astronomy.

If a source at a temperature T_s does not extend over the entire beam of the antenna, the effective antenna temperature T_a is lower than T_s and is given by

$$T_A = \frac{A_e}{\lambda^2} \int_{\Omega_s} T_s(\theta,\phi) Q(\theta,\phi)\, d\Omega$$

$$= \frac{1}{\Omega_A} \int_{\Omega_s} TQ\, d\Omega\,.$$

If the source is very small compared to the beam area Ω_A, which is the case for planetary measurements, then Q is effectively equal to 1, and

$$T_A = \frac{1}{\Omega_A} \int_{\Omega_s} T_s\, d\Omega = \frac{\Omega_s}{\Omega_A} T_{\text{av}}\,,$$

where T_{av} is the average temperature of the source (K).

The temperature T_{av} thus obtained is the temperature that a blackbody radiator would need to have to produce the radiation observed at the wavelength λ, and it is commonly called the equivalent blackbody temperature. For thermal sources, the temperature is related to the actual thermal temperature of the source through the emissivity, as follows:

$$T_{\text{bb}} = T_{\text{av}} = \rho T_{\text{actual}}\,,$$

where ρ is the emissivity. The equivalent blackbody temperature closely approximates the thermal temperature. This is, of course, not true for nonthermal sources such as the decimetric and dekametric emissions from Jupiter.

In planetary measurements performed with existing radio-astronomical antennas, the ratio Ω_s/Ω_A is quite small, and hence T_A is much smaller than T_s. For example, in their observation of Mars in 1958, Mayer, McCullough, and Sloanaker (1958) obtained an antenna temperature $T_A = 0.24$ K at $\lambda = 3.15$ cm. Since the angle subtended by the disk of Mars was 18 arc seconds, and the antenna half-power beam width was 0.116 deg,

$$\Omega_s = \pi r^2 = \pi(9/3600)^2 = 2 \times 10^{-5} \text{ deg}^2$$

and

$$\Omega_A = \frac{4}{3}(0.116)^2 = 0.018 \text{ deg}^2 .$$

So, the apparent equivalent temperature of Mars is given by

$$T = T_A \frac{\Omega_A}{\Omega_s} = (0.24)(0.018/2 \times 10^{-5}) = 216 \text{ K} .$$

This example shows the difficulty of these measurements, as one must measure a T_A of a fraction of a degree in a noisy background. The physics is simple, but the technique and technology are sophisticated. Table 1 (Barrett, 1965), which lists the maximum expected antenna temperatures for a given antenna for planets in the solar system, indicates the difficulty of detection.

Let us begin by discussing the measurements of the brightness temperature of Venus, which have been made at wavelengths from 0.1 cm to 70 cm over a period of years. Figure 1 (Dickel, 1966) is a plot of brightness temperature versus wavelength, and we note that it rises and then falls, with a peak at about 13 cm, where $T_B \approx 712 \pm 80$ K. The explanation for this effect has to do with the very dense atmosphere of Venus. The falloff at short wavelengths is due to increasing opacity of the atmosphere with decreasing wavelength. The falloff at longer wavelengths is possibly due to a change in emissivity, a decrease in emissivity with wavelength, or the phenomenon that the radiation originates from a subsurface layer that is cooler than the surface region responsible for the emission in the middle wavelengths.

Through comparison of the computed microwave spectrum with that observed, these observations have served as a test of various models of the atmosphere of Venus. Thus, radio-astronomical observations have stimulated a large number of theoretical studies of the atmosphere of Venus.

Several efforts were made to find a relation, if one exists, between the phase angle of Venus and its brightness, in other words, to see if its sunlit side is warmer than its nightside. The main problem is that measurements cannot be made when Venus is in line with the Sun, which occurs near conjunction. Venus has such a dense atmosphere that one would expect very little, if any, variation in temperature with phase. Mayer, McCullough, and Sloanaker (1963) reported observing a minimum near inferior conjunction at 3.15 cm. However, Epstein et al. (1968)

Table 1.–Typical values of received power and antenna temperatures for planetary radiation at 3-cm wavelength.*

Planet	Solid Angle** (sr)	Mean Brightness Temperature $\overline{T}_B$ (K)	Antenna Temperature (K)	Received Power (W)
Mercury	2.2×10^{-9}	400	0.36	5.0×10^{-17}
Venus	7.0×10^{-8}	600	17.1	2.4×10^{-15}
Mars	5.9×10^{-9}	200	0.48	6.6×10^{-17}
Jupiter	4.1×10^{-8}	150	2.5	3.4×10^{-16}
Saturn	7.0×10^{-9}	100	0.28	3.9×10^{-17}
Uranus	2.4×10^{-10}	100†	0.0098	1.4×10^{-18}
Neptune	8.3×10^{-11}	100†	0.0034	4.7×10^{-19}
Pluto	5.0×10^{-12}	100†	0.0002	2.8×10^{-20}

*A parabolic antenna, 30.5 m (100 ft) in diameter, with an aperture efficiency of 0.50 at $\lambda = 3$ cm and a receiver bandwidth of 10 MHz is assumed.

**At mean conjunction or opposition.

†Estimated temperature for purposes of comparison.

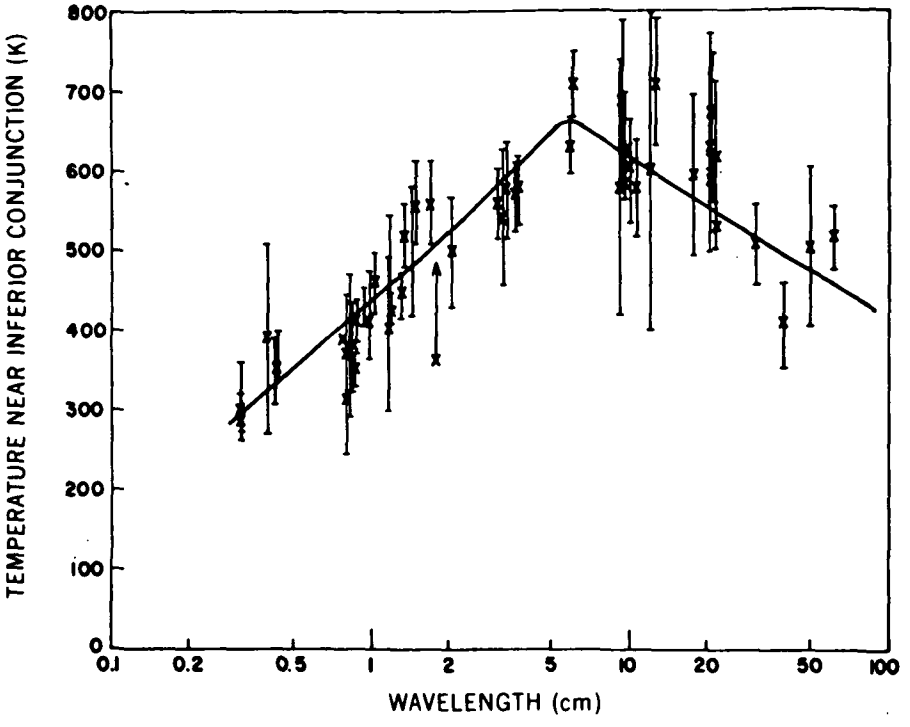


Figure 1.–The microwave spectrum of Venus near inferior conjunction.

reported the opposite (i.e., a maximum temperature at or near inferior conjunction at 3.4 mm). However, this is probably an observation of the top of the cloud layer, where there might be an inverse phase effect due to circulation from the subsolar to the antisolar side. At longer wavelengths, one might view the surface, where the variation may be nil (Drake, 1964; Dickel, 1966; Dickel et al., 1968). Sinclair et al. (1970) performed interferometric measurements designed to determine equator-to-pole variations in temperature. Their data suggested some variation in temperature between sunset and sunrise terminators, and also from pole to equator. The variation, however, is quite small, of the order of 10 K.

Now, let us consider Mars. Mars was first detected by radio-astronomical techniques in 1958 by Mayer, McCullough, and Sloanaker (1958). They found an effective temperature of about 218 ± 76 K. Since then, there have been many measurements made that are not, basically, in conflict with one another. Table 2 (Dent et al., 1965) summarizes brightness temperature measurements of Mars for

several wavelengths; if the uncertainties are taken into account, the measurements all fall near 200 K, which is lower than the observed IR temperature of about 250 K. This can be explained in several ways. One possibility is that the radio emission originates below the surface, where it is cooler. The decreasing temperatures with longer wavelengths indicate what one would expect, that longer wavelength emissions originate from further below the surface.

Mercury, our next subject, presents a special observational problem in that at maximum elongation it is only 28 deg from the Sun. This makes it necessary to separate the contribution of Mercury from that of the Sun. Figure 2 presents an example of how this was done through averaging to obtain a 0.04-K antenna temperature rise (Howard, Barrett, and Haddock, 1962).

Figure 3 shows a large amount of data on the Mercury brightness temperature, compiled by Klein (1970) at 3.75 cm. There is a definite variation with phase from about 300 to 450 K, which is in rough agreement with some of Epstein's work (Epstein et al., 1970). There seem to be peaks at superior conjunction and minima at

Table 2.–Summary of Martian temperature measurements.

Wavelength (cm)	Disk Temperature T_D (K)	$T_D(R/\bar{R})^{1/2}$ * (K)	Reference
0.0012	242**	234	Sinton (1964)
0.12	165 ± 17	169 ± 17	Low (1965)
3.14	211 ± 28	211 ± 28	Giordmaine, Alsop, Townes, and Mayer (1959)
3.15	218 ± 50	214 ± 49	Mayer et al. (1958)
3.75	182 ± 11	190 ± 12	Dent, Klein, and Aller (1965)
6.0	192 ± 26	200 ± 27	Kellerman (1965)
10.0	177 ± 17	184 ± 18	Heeschen (1963)
11.3	162 ± 18	169 ± 19	Kellerman (1965)
21.3	190 ± 41	198 ± 43	Kellerman (1965)
	(169 ± 32)†	(177 ± 34)†	

*R = Mars solar distance at the time of measurement; $\bar{R}$ = mean solar distance (1.524 AU).
**Average of Sinton's reported measurements.
†Value obtained by Kellermann (1965) by deleting one extreme measurement.

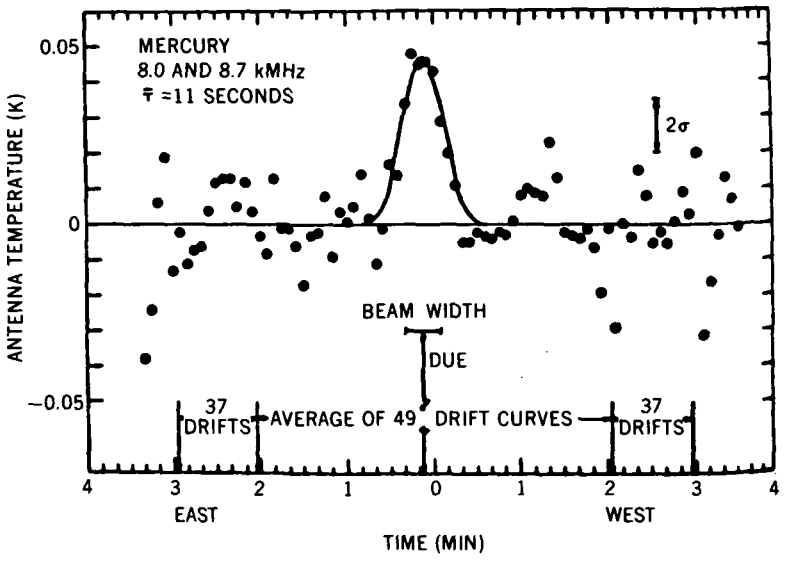


Figure 2.–The first detection of Mercury. Note that it was necessary to average many curves before the Mercury signal was apparent.

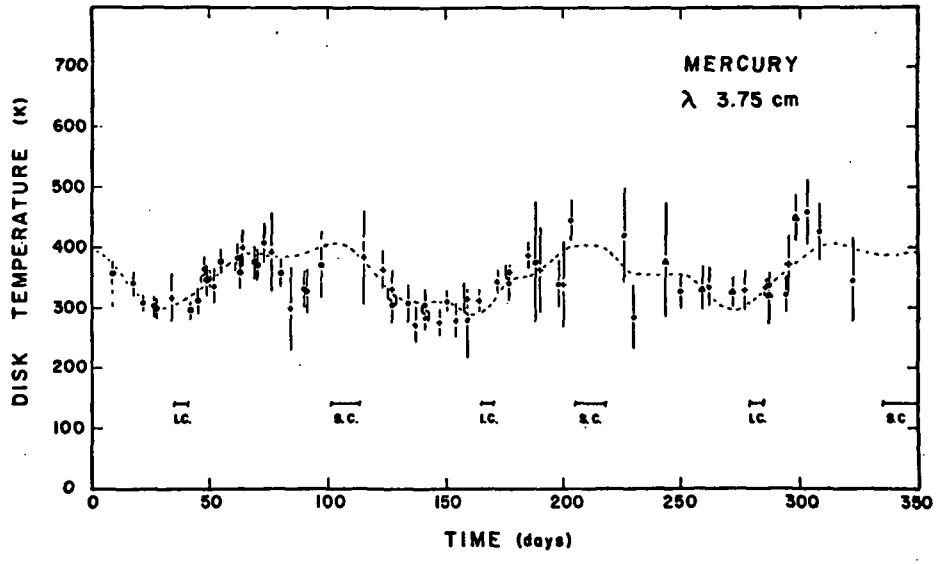


Figure 3.–The 3.75-cm disk temperature of Mercury from October 1965 through April 1968. The data are superposed assuming a 350-day (3-synodic-period) quasi-periodicity. The measurements made in 1965, 1966, 1967, and 1968 are represented by triangles, crosses, filled circles, and open circles, respectively.

inferior conjunction, but they are probably due to changes in both the phase (the illuminated portion of the surface) and the reflectivity of the region facing the Earth. When one is looking at the entire planet, one obtains an integrated average temperature, of course. Morrison* discusses the problem of thermal physics of the surface of Mercury and comes to the conclusion that it acts much as the Moon would at the distance of Mercury. Table 3 (Morrison and Klein, 1970) provides a summary of measurements of the brightness temperature of Mercury.

The thermal radio emission from the outer planets has been measured at wavelengths ranging from millimeter to decimeter (Kellerman, 1970). The measured brightness temperatures usually exceed the expected equilibrium temperatures based on the albedo and solar heating, and, at least for Jupiter and Saturn, the existence of an internal source of heat comparable to solar heating is indicated.

The observed brightness temperature of Jupiter does increase somewhat with increasing wavelengths, which indicates there is an absorbing atmosphere and that at the shorter wavelengths one is looking at the cooler upper atmosphere, whereas the longer wavelengths represent areas deeper in the atmosphere where the temperature is higher. At millimeter wavelengths (Epstein, 1968), for instance, the brightness temperature is about 140 K, representing a region in the atmosphere somewhere above the clouds.

Saturn was first detected by radio astronomy in 1960 (Cook et al., 1960). The IR temperature of Saturn as determined by Low (1966) is about 93 ± 3 K. From Table 4 (Gulkis et al., 1969), we see that at 3.45 cm, the microwave brightness temperature is about 106 K, and the temperature rises to about 170 to 200 K at wavelengths of about 10 cm. Some investigators had concluded that there are radiation belts around Saturn, as in the case of Jupiter, and therefore one should expect an increase in microwave emission at longer wavelengths. However, to date there has been no nonthermal radiation detected from Saturn. There have been attempts to measure the polarization of the radiation received from Saturn, as had been done for Jupiter, which would be caused by the orientation of a magnetic field supporting a radiation belt. Since interferometric methods have found no evidence of polarization, we conclude there is little, if any, exterior magnetic field. Moreover, Saturn's radiation seems to originate from the disk, whereas Jovian decimetric emissions appear to come from outside the disk of the planet.

To summarize, then, Saturn has a thick atmosphere, and the increase in microwave brightness temperature with increasing wavelength is due to a deep

*To be published in the *Astrophysical Journal.*

Table 3.—Observed average temperatures of Mercury.

Wavelength (cm)	Date	No. of Days	T_B (K)	Reference
0.34	1965 to 1966	155	277 ± 30	Epstein et al. (1967)
1.95	Feb. to Mar. 1966	11	288 ± 30	Kaftan-Kassim and Kellerman (1967)
1.95	June 1967 to Mar. 1968	12	350 ± 30	Morrison and Klein (1970)
2.82	May 24, 1968	1	375 ± 40	Medd (1968)
3.75	1965 to 1968	84	380 ± 20	Klein (1969)
6.00	Jan. to Mar. 1969	8	385 ± 20	Morrison and Klein (1970)
11.3	May to June 1964	14	300 ± 40	Kellerman (1965)

atmosphere whose opacity changes with wavelength. Gulkis et al. (1969) have attributed the opacity to ammonia, which need be present only in cosmic abundance.

Table 4.—Observed microwave spectrum of Saturn.

Wavelength (cm)	Disk Temperature* (K)	Reference
0.12	140 ± 15	Low and Davidson (1965)
0.32	97^{+52}_{-42}	Tolbert (1966)
0.34	130 ± 15	Epstein (1968)
0.43	103^{+70}_{-64}	Tolbert (1966)
0.80	132 ± 9	Salomonovich (1965)
0.86	116 ± 30	Tolbert (1966)
0.955	118 ± 20	Hobbs (reported by Epstein, 1968)
1.53	146 ± 23	Welch and Thornton (1965)
1.53	141 ± 15	Welch, Thornton, and Lohman (1966)
1.90	200 ± 30	Kellerman and Pauliny-Toth (1966)
3.12	123 ± 16	Berge (reported by Berge and Read, 1968)
3.45	106 ± 21	Cook, Cross, Bair, and Arnold (1960)
6.0	179 ± 19	Kellerman (1966)
6.0	190 ± 45	Hughes (1966)
9.0	165 ± 25	Berge and Read (1968)
9.4	177 ± 30	Rose, Bologna, and Sloanaker (1963)
10.0	196 ± 44	Drake (1962)
10.7	172 ± 20	Berge and Read (1968)
11.3	196 ± 20	Kellerman (1966)
11.3	182 ± 20	Davies, Beard, and Cooper (1964)
21.2	286 ± 37	Davies and Williams (1966)
21.3	303 ± 50	Kellerman (1966)
70.0	<1250	Gulkis et al. (1969)

*Errors quoted are believed to be one standard deviation.

Figure 4 (Gulkis et al., 1969) shows the microwave spectrum of Saturn, with theoretical spectra due to an atmosphere containing ammonia. The solid lines represent varying amounts of ammonia, and a mixing ratio of $\alpha = 5 \times 10^{-4}$ seems to fit best. The increasing brightness temperatures at longer wavelengths indicate that the emission is coming from deeper within the atmosphere, but it is still thermal.

For Uranus, the situation is analogous although more difficult since the planet is about 19 AU from the Sun. Given an albedo of about 0.45, one would expect a temperature of about 67 K. Menzel and others in 1926 found an upper limit on the IR temperature of about 100 K. Radio measurements at 11 cm by Slee (1964) indicated a brightness temperature of 300 K with a large uncertainty. Kellerman (1966) with the 210-foot antenna at Parks, Australia, obtained a measurement of 130 ± 40 K for the temperature of Uranus, which is the most reliable to date. If one

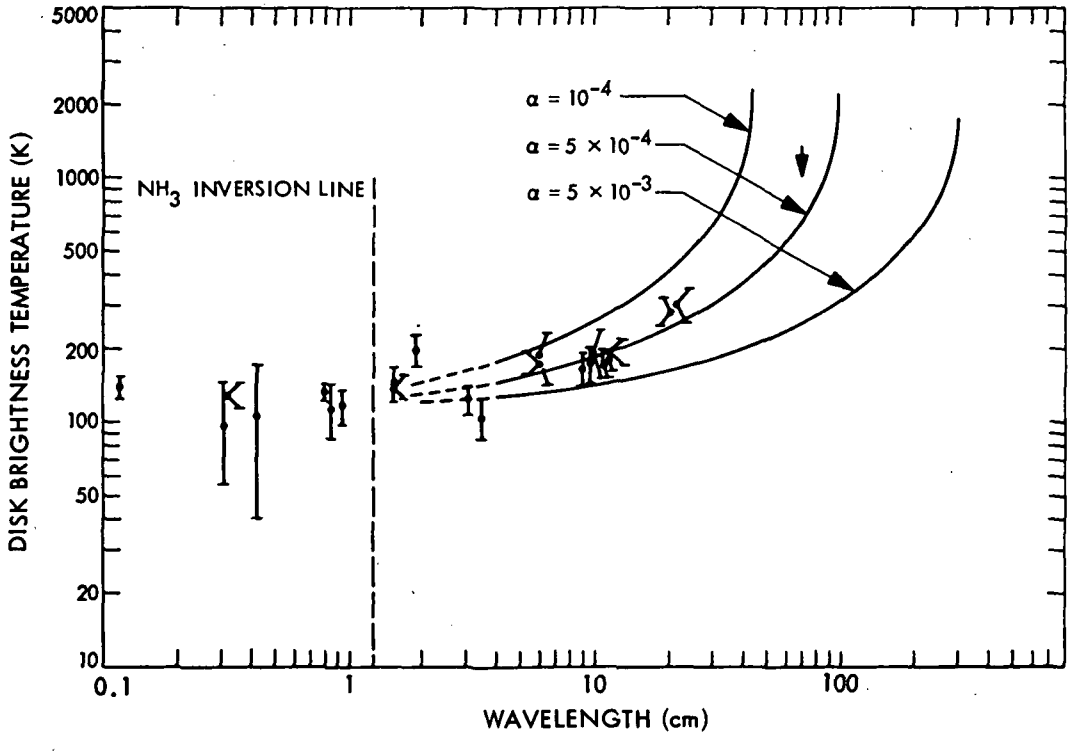


Figure 4.—The microwave spectrum of Saturn. The solid lines are theoretical calculations for different values of the ammonia mixing ratio α.

compares the measured microwave temperature with the expected temperature (on the basis of insolation and albedo), the microwave temperature is considerably higher, which suggests a greenhouse effect and thick atmosphere (as on Saturn), an increase in temperature with depth, and the presence of a microwave absorber, such as ammonia.

Neptune and Pluto are too cool and too far away to be investigated with current radio-astronomical techniques. One will probably have to wait until the "Grand Tour" missions in the late 1970's to obtain spacecraft measurements of the temperatures of these planets.

III. RADAR OBSERVATIONS

Radar technology, like many other fields, owes its development as a practical method to World War II. Since those military beginnings, both the technology and the application have made great strides, with a happy result for astronomy.

Radar transmits a very powerful radio-frequency wave, which is reflected by some discontinuity in its path (such as an airplane or a planet's surface); the reflected signals, greatly diminished in power by their round-trip journey, are received by the antenna and can be used to interpret the nature of the discontinuity. While military radar need only detect a large airplane at a range of a few kilometers, planetary radar must be capable of detecting Mercury, for example, which has at best the radar reflectivity of a dime at 10 000 miles (Goldstein, 1970). One obviously needs a high-powered transmitter, a large antenna, and sophisticated receiving equipment. The radar equation describes the power returned in the echo (Evans and Hagfors, 1968) as follows:

$$P_r = P_t G A \sigma / (4\pi R^2)^2 ,$$

where

P_r = echo power,

P_t = transmitter power,

σ = radar cross section of the target.

R = distance to the target,

G = gain of transmitting antenna,

and

A = aperture of the receiving antenna.

The echo power falls off, then, as the fourth power of the distance. If the target is a sphere of radius a, then $\sigma = \pi a^2 \rho G_m$, where ρ is the reflectivity and G_m is called directivity. From planetary echoes, one cannot separately estimate ρ and G_m; they are lumped into one parameter which may depend on wavelength. Note that there are two parameters which appear to be similar, G_m and G. They are defined similarly; however, one is a property of the antenna, the other of the reflecting surface. In planetary radar work, one is dealing with very distant targets which subtend very small solid angles; hence, only the gain on the axis of the antenna is important. This is the maximum gain of the antenna, and it is equal to the ratio of the power actually received to the power that would be received with an omnidirectional antenna of unity gain.

As an example, let us consider the case of Venus near inferior conjunction (closest to the Earth) being ranged by the NASA/JPL 210-foot-diameter Goldstone facility. Near inferior conjunction, the distance to Venus is about 60 million kilometers. The gain of the antenna is about 60 dB, and the cross-sectional area of Venus is

$$\pi a^2 = \pi(6.05 \times 10^6)^2 = 1.5 \times 10^{14} m^2 .$$

For ρG_m of 0.1, the total received echo power for a transmitted power of 400 kW is about 10^{-17} W, which is still several orders of magnitude above the detection threshold. It should be recalled, however, that this represents the most favorable situation in planetary radar ranging and that other targets return far less power. The relative detectabilities of other planetary targets are shown in Figure 5 (Pettengill, 1965). It is assumed that the surface reflectivities are equal (which is not true, especially for the outer planets, which have heavy absorbing atmospheres). However, one can see that Mercury and Mars require a sensitivity about 10^4 greater than that required to detect Venus under the most favorable conditions.

There are basically two modes of operation of planetary radar. One is to transmit a continuous-wave monochromatic signal and receive the echo. This echo can be analyzed for orbital motion, relative velocity, surface roughness, and the rotation rate of the planet. Note, however, that there is a large amount of relative motion (e.g., rotation of the Earth, orbital motions of the target, and so forth) which must be considered. The frequency of the receiver is usually controlled by means of a computer-generated ephemeris drive (which computes the received frequency one should expect due to the relative motion and adjusts the receiver accordingly).

The second mode of operation is modulation, or pulsing, of the continuous wave. Then, one can obtain time-of-flight measurements by comparing times of transmission and reception of a certain code, and from this time (propagation delay), the range to the target can be calculated. Propagation delay is also related to the orbital motion of the target and to its radius.

Table 5 (Evans, 1969a) gives delay depths and Doppler spreads for the four nearest planets. The delay depth is the difference between the delay time for reflections from the front face of the target (nearest the Earth) and the delay time obtained from diffuse scattering centers near the limb. The delay depth is thus a function of the radius of the planet. Similarly, the limb-to-limb Doppler spread is dependent upon the rotation rate of the planet.

Most of the work described below has been performed with the Goldstone radar of JPL, the Millstone and Haystack facilities of the MIT Lincoln Laboratory, and Cornell University's Arecibo Observatory in Puerto Rico. Also, some radar work has been performed in the Soviet Union.

Venus was the third extraterrestrial target (after the Sun and Moon) from which a reflection was received. The first unambiguous evidence of a Venus return was obtained by JPL (Victor and Stevens, 1961) with an 85-foot-diameter antenna operated at about 100 kW transmitter power. That same year, the MIT Millstone radar recorded echoes from Venus (Pettengill et al., 1962), and the era of planetary radar astronomy had begun.

By simultaneously measuring Doppler shift and delay, one can locate zones from which the signal is being reflected. The process is analogous to a plane wave's hitting the planet's surface; the wave would be reflected first from the subradar point. After this, subsequent portions of the echo would come from annular rings of successively larger radii. Figure 6 (Evans, 1969a) explains the delay-Doppler method of locating the zones of origin of the echo signals. This technique is used to obtain radar reflectivity maps of planetary surfaces, which is discussed later in this section.

With a continuous-wave signal, a Doppler spectrum can be obtained by plotting the received power spectral density as a function of the Doppler shift. At zero Doppler shift, the power is maximum, and it decreases rapidly with Doppler shift, in the case of Venus by some 30 dB, or by three orders of magnitude. This is due to decreased reflectivity at the lower angle of incidence near the limb. Such spectra have been used to obtain the scattering properties of the surface of Venus (Muhleman, 1966; Ingalls and Evans, 1969). By the use of range gating, such spectra can be obtained for each separate annular section of the target corresponding to a given delay time. The spectra obtained from the limb are much weaker than those from the subradar point, for reasons discussed above.

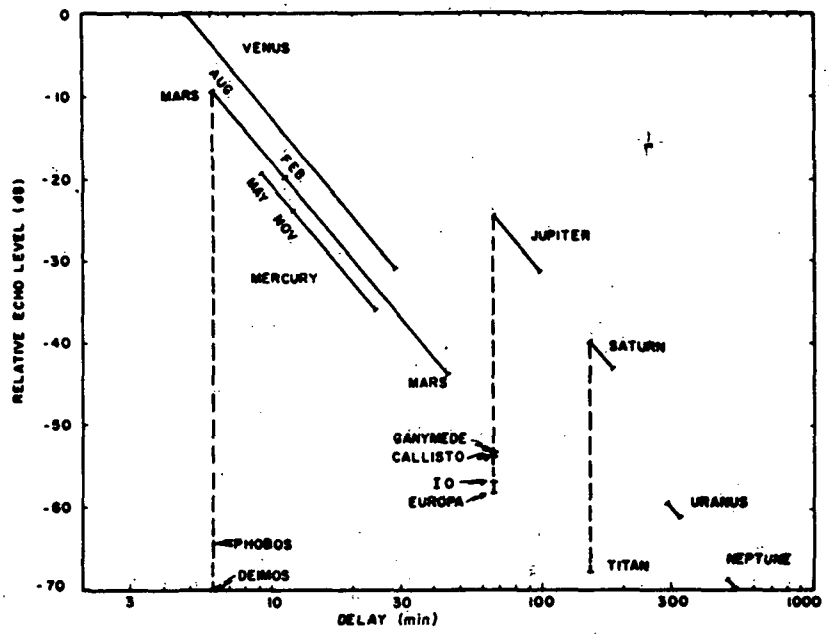


Figure 5.—Relative detectability and echo delay for the planets and their moons by radar, with equal surface reflectivity assumed. Months listed by tick marks for Mars and Mercury show detectability when closest approach occurs on that date.

Table 5.—Delay depth and Doppler spread* for radar-astronomy targets.

Target	Two-Way Delay Depth to Limb (ms)	Limb-to-Limb Doppler Spread	
		Maximum (Hz)	Minimum (Hz)
Mercury	16.2	≈43	≈18
Venus	40.6	44	13.6
Mars	22.6	3 200	
Jupiter	≈475	167 000	

*For f = 1000 MHz.

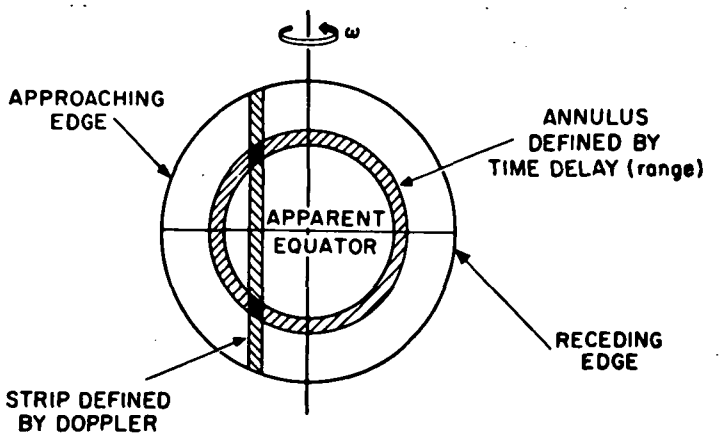


Figure 6.—Diagram of the disk of a planet as seen from the Earth. By transmitting a short pulse, it is possible to resolve a narrow annulus as shown. Alternatively, if continuous-wave transmissions are employed, it is possible by spectrum analysis of the echoes to select strips parallel to the projected axis of rotation. The small heavily shaded regions may be isolated by combining the two techniques (delay-Doppler mapping).

Still another method for studying planetary surfaces is the use of depolarized spectra (Goldstein, 1965; Carpenter, 1966). The signal is sent out with circular polarization, for example. When the signal encounters boulders, rocks, mountains, or depressions, part of the echo is depolarized, and it arrives at the receiver polarized in the opposite sense. By virtue of the fact it is produced by rough terrain, the depolarized spectrum makes it possible to detect surface features. The depolarized component is usually much weaker than the polarized component of the echo signal. By observing the position of spectral salients in spectra taken on different dates and correlating this information with planetary geometry, one can get information on the rotation rate of a planet.

By analyzing the received spectra at different wavelengths, one can obtain the reflectivity (or the radar cross section) since the power transmitted is known, and the power received from the entire planet can be measured by integrating the Doppler spectrum. Figure 7 (Evans, 1969b) is a summary of such data, given as a percentage of the expected return from a perfectly reflecting sphere having the radius of Venus. We note that at short wavelengths the cross section is low, due to a

large attenuation of the signal caused by absorption in the atmosphere. The longer wavelengths lead to a higher cross section (about 16 percent, which is greater than that of either Mars or the Moon) and seem to pass through the atmosphere relatively unattenuated. A combined study of the radar cross section and the variation of brightness temperature with wavelength provides a test for models of the

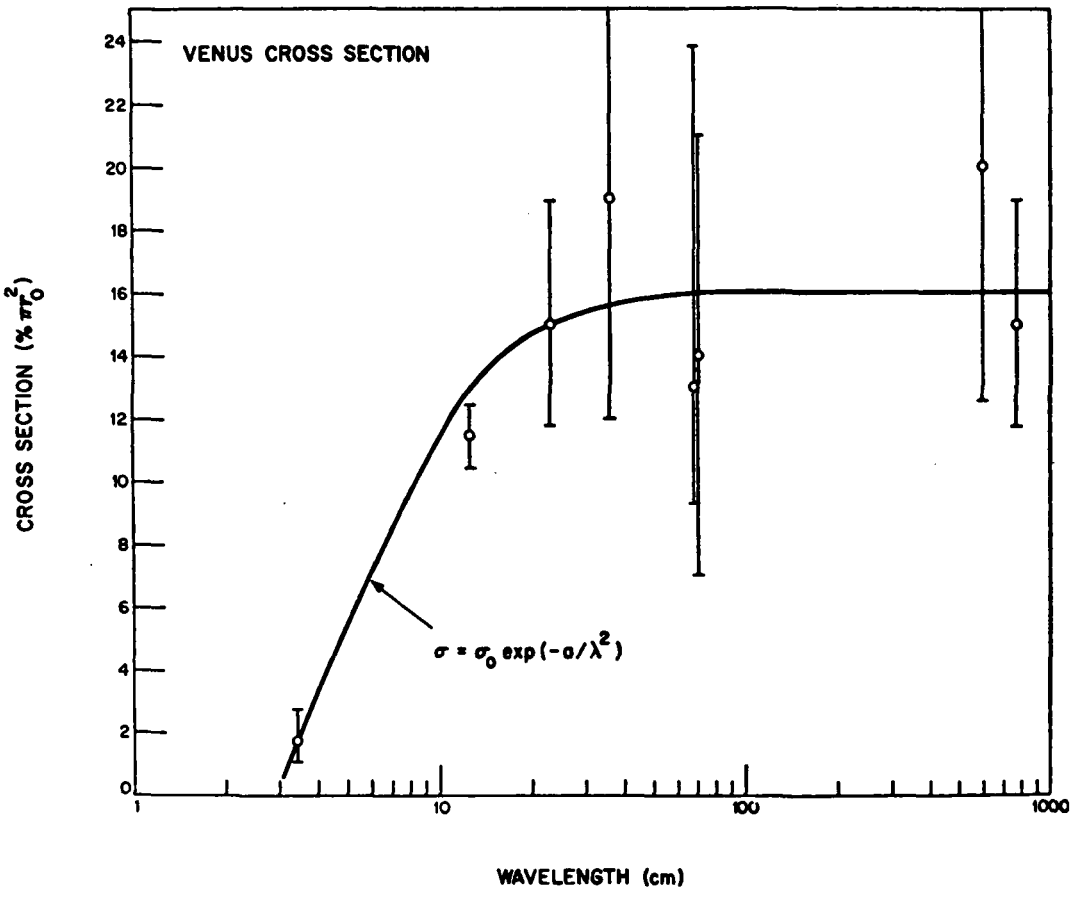


Figure 7.—Variation of the radar cross section of Venus with wavelength according to various observers. The curve is theoretical and is based upon the assumption that the intrinsic cross section is 16 percent πr^2 and that atmospheric absorption is responsible for lowering the cross section; the optical depth depends on the square of the frequency.

atmosphere of Venus. The decrease in the cross section at shorter wavelengths is predicted by the model shown in Figure 7.

A valuable contribution of radar astronomy has been the determination of the rotation rate of Venus. Recent radar determinations of the rotation rate are presented in Table 6 (Evans, 1969a). There are basically three methods for determining rotation:

(1) The base bandwidth method makes use of the fact that in a spectrum of a continuous-wave echo return, the Doppler shift of the wings of the spectrum is proportional to the rotation rate times the radius of the planet (Carpenter, 1966). When such measurements are performed, over a long interval of time, a curve such as Figure 8 (Pettengill, 1965) is obtained. After the orbital geometry and rotation of the Earth are included, a best-fit curve corresponding to a certain value of rotation rate can be found. This method suffers from the drawback that the limb components in an echo are very weak and very difficult to recognize.

(2) The delay-Doppler method (Dyce, Pettengill, and Shapiro, 1967; Shapiro, 1967) avoids this difficulty of weak limb components by employing range-gated pulses which provide spectra of echoes from discrete annuli of the target. In these spectra, the power is greatest where the Doppler contours become tangential to the annulus, and, hence, these frequencies can be measured more easily.

(3) The feature method, the simplest of the three, is based on the observation of the motion of features with time on depolarized echo spectra. This was one of the early methods used to investigate the rotation rate of Venus (Goldstein and Carpenter, 1963).

The radar rotation period of Venus, about 243 days retrograde, is surprisingly close to the period required for Venus to present the same side toward the Earth at each inferior conjunction. This may be a consequence of "Earth lock," in which case a rather large equatorial asymmetry in the moments of inertia of Venus must exist (Goldreich and Peale, 1966). However, no such asymmetry was immediately evident in the Mariner 5 celestial mechanics experiment results.

Because the surface of Venus is totally obscured at optical wavelengths by dense clouds, radar methods provide the only possibility of observing its nature. Ever since Goldstein's and Carpenter's (1963) early work, radar astronomers have been aware of areas of different radar reflectivity on the surface of Venus. Recently, with new improvements in planetary radar facilities, it became possible to obtain radar reflectivity maps of the surface of Venus. Rogers and Ingalls (1969), using the Haystack and Westford facilities of Lincoln Laboratory as a radar interferometer, obtained a radar map of Venus at a maximal resolution of about 110 km, covering a

Table 6.–Estimates of the rotation period of Venus.

Date of Observation	Period (days)	Pole a	Position δ	Method*	Reference
1961	≈225(?)	(?)	(?)	1	Victor et al. (1961)
1962	- 230 +40, - 50	(?)	(?)	1, 3	Goldstein and Carpenter (1963)
1964	- 249 ± 7	75 +10, - 4	68 ± 4	1, 3	Carpenter (1966)
1964	- 244 ± 2	90.9 ± 1	66.4 ± 1	2	Dyce et al. (1967)
1964	- 241 ± 1	94 ± 3	- 64 ± 2	2	Shapiro (1967a)
1966	- 242.6 ± 0.6	(?)	(?)	3	Goldstein (1967)
1966	- 243.09 ± 0.18	84.7 ± 1.8	65.8 ± 1.2	2, 3	Shapiro (1967b)

*Methods (discussed in text): (1) base bandwidth, (2) delay-Doppler, (3) feature.

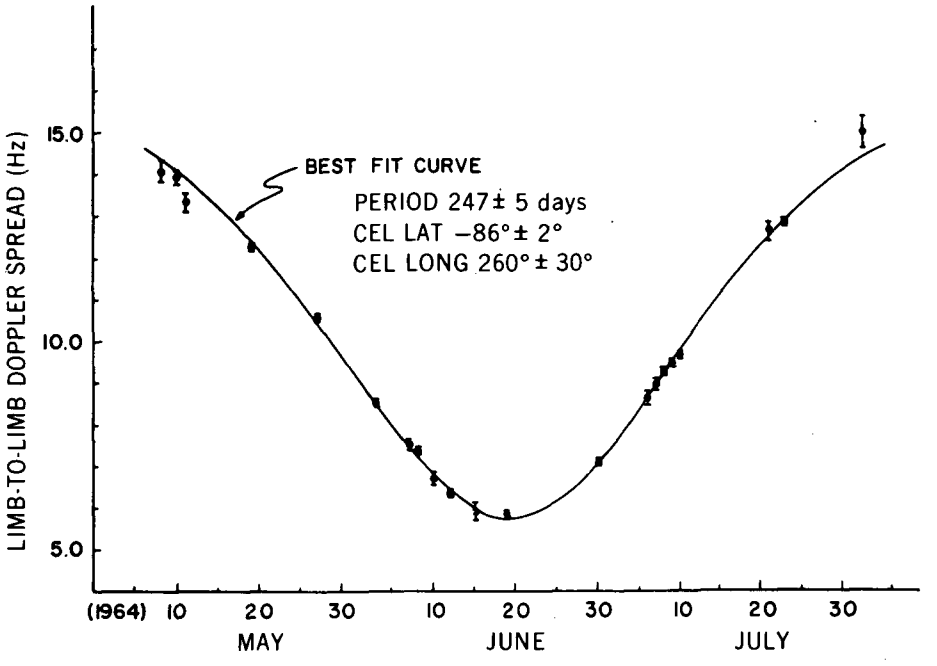


Figure 8.—Plot of limb-to-limb Doppler spread versus date observed for Venus during the 1964 inferior conjunction. The solid curve represents the least-mean-squares fit to the data and corresponds to the rotation axis specified. (Data taken at the Arecibo Ionospheric Observatory.)

latitude range of −50 to 40 deg. The use of interferometric techniques enabled the north-south ambiguity inherent in the delay-Doppler mapping technique to be resolved.

Goldstein and Rumsey (1970), using the Goldstone facility of JPL, obtained a radar map of Venus having a best resolution of about 50 km, which is approximately the same as viewing the Moon with the unaided eye; a reproduction of this map is shown in Figure 9. Many radar-bright and radar-dark features are visible, including the feature Alpha (Zohar and Goldstein, 1968), which appears to be roundish, with a diameter of about 1000 km. A delay-Doppler technique employing range gating to obtain time-delay data was used. Range gating is accomplished by modulating the transmitted signal and applying the inverse modulation to the received signals so that only signals with the proper time delay can pass through the gate. In both of these methods, an empirical scattering law was applied to compensate for the dramatic

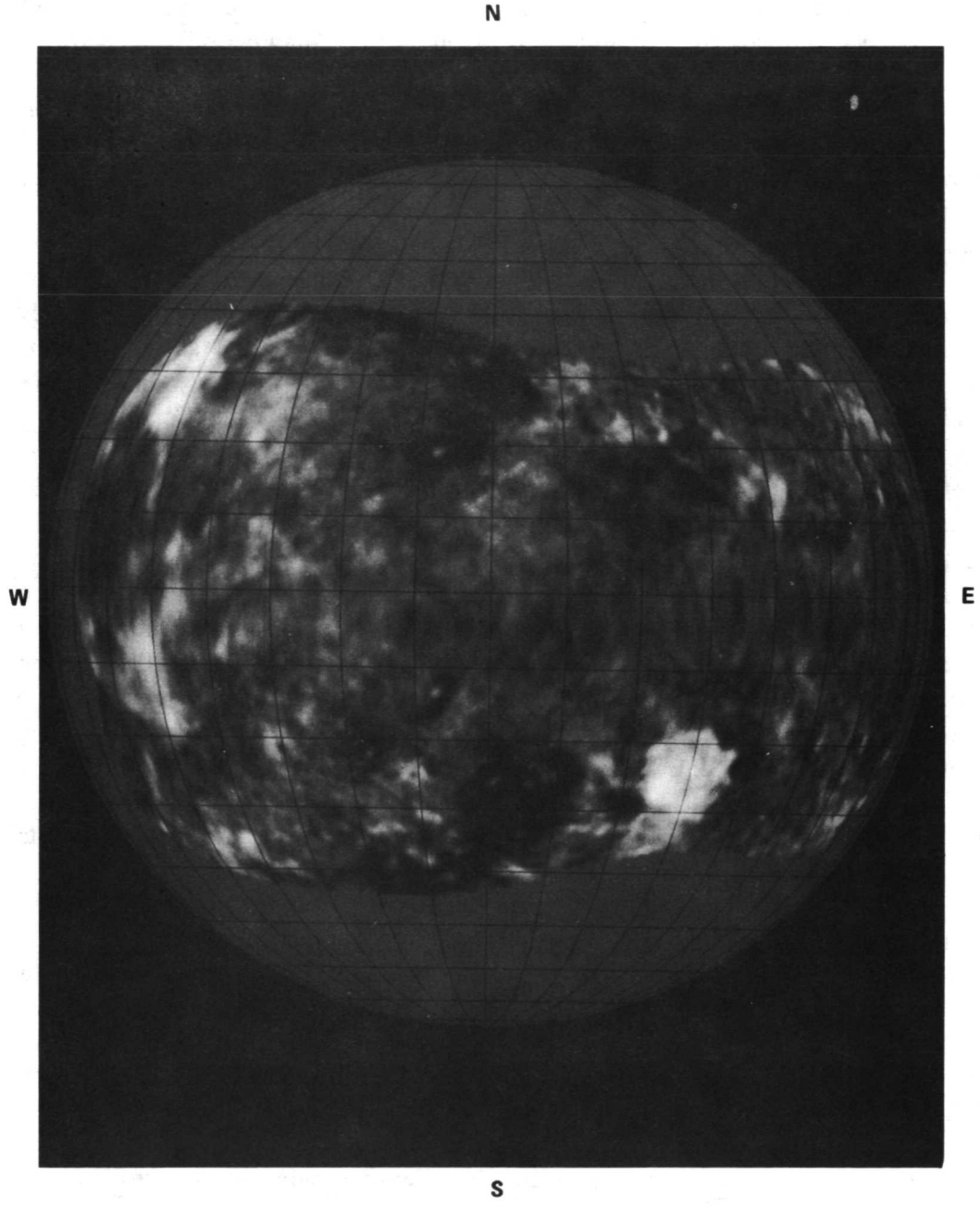


Figure 9.—Radar reflectivity map of Venus.

decrease of return echo power from the subradar point to the limb. The north-south ambiguity was resolved by combining data taken over a 17-day period, during which the geometry changed appreciably.

The variation of surface height in the equatorial regions of Venus has also been studied by means of the delay-Doppler method (Smith et al., 1970). An elevated region about 2 km in height was observed to extend for about 150 km in longitude, and the equatorial asymmetry was found to be about 1 km, with the center of figure displaced from the center of mass by about 1.5 km.

Some of the radar-reflecting surface features of Venus (which may be fields of boulders, craters, lava flows, mountain ranges, and so forth) have been given names of pioneers in electrical and radio science, such as Hertz, Gauss, Maxwell, Edison, and Faraday. One may be certain that ultimately these will be joined by the pioneers in radar astronomy, such as Goldstein, Pettengill, Evans, and Carpenter.

With continued observations of the planets, one can also obtain such celestial mechanics information as a planet's radius and the geometric distance to the planet at any given time. One can fit the radar data to an elaborate model and obtain residuals. This was done by Ash et al. (1968) in order to investigate the discrepancy in the value of the radius of Venus between the radar and Venera 4 results. When the residuals (the differences between observed and theoretically predicted return times) are plotted using a 6075-km radius for Venus, the fit is poor (Figure 10); yet, the Venera 4 probe ceased to transmit at a point in the atmosphere which Mariner 5 occultation results showed to be 6075 km from the center of Venus. Using the same radar data, but with a radius of 6048 km, Ash et al. obtained an excellent fit (Figure 11). The radius of 6048 was determined by radar measurements. This result indicated that Venera 4 failed some 25 km above the surface, as Soviet scientists eventually admitted.

Mars is a much more difficult target than Venus: It is further away, and the return spectra are much more spread out due to its faster rotation. In 1963, Goldstein and Gillmore (1963) showed the first spectrograms in which Mars could be distinguished from the background. A composite averaged spectrum of Mars (Pettengill et al., 1969) is shown in Figure 12; the Doppler shift is measured in kilohertz instead of hertz as in the case of Venus. Different surface features provide different reflectivities at different longitudes. Spectrograms from Mars show a much sharper, higher-power spectrum at the reflecting subradar point. This indicates that the surface of Mars is, in general, smoother than Venus at radar wavelengths, and the radar reflection is more specular in nature.

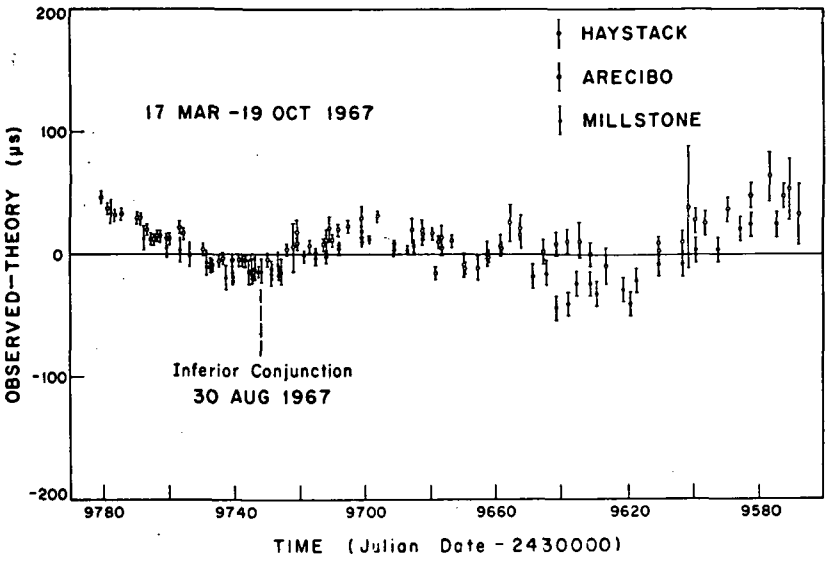


Figure 10.—Sample of the post-fit residuals of Earth-Venus time-delay measurements obtained by assuming the radius of Venus to be 6075 km. Fixing the radius at a larger value leads to an even poorer agreement with the radar data.

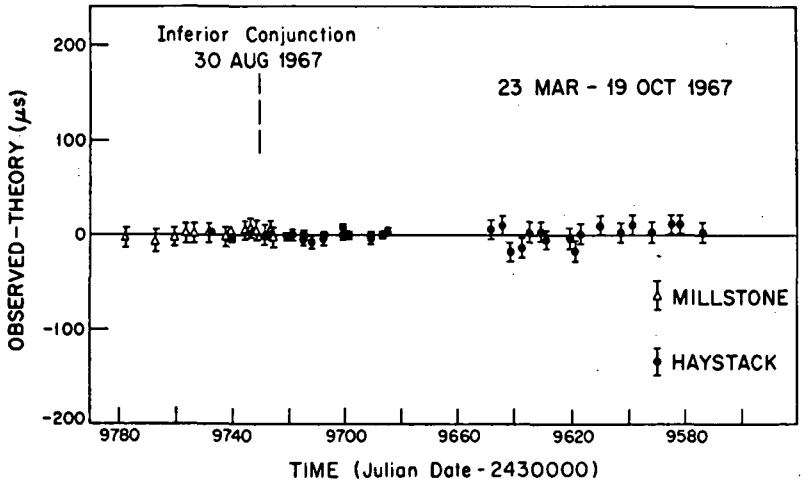


Figure 11.—Representative sample of the post-fit residuals of Earth-Venus time-delay measurements obtained from the Lincoln Laboratory radar data. The estimated radius of Venus was 6048 km.

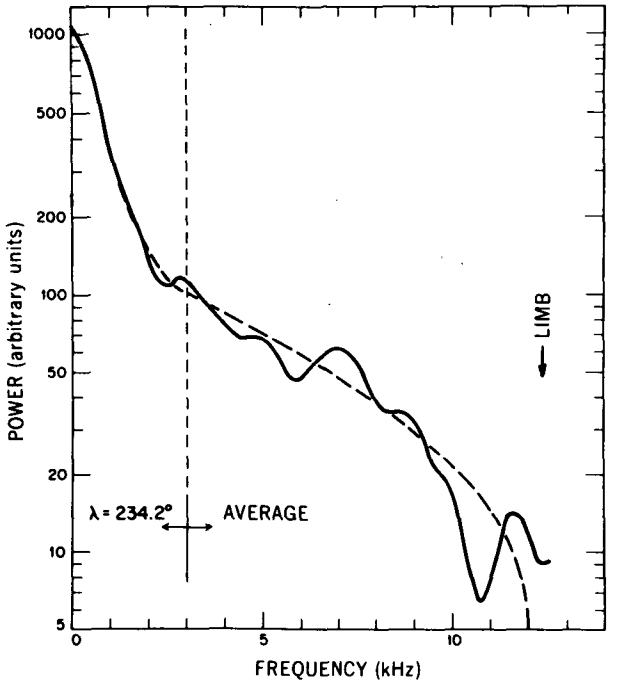


Figure 12.—Frequency power spectrum of Mars constructed by Pettengill et al. (1969).

Because most of the power in the echo from Mars is reflected near the subradar point, the measurement of delay time and received power as Mars rotates under the radar can be used to determine topographical features and the radar reflectivity (cross section) along the path of the subradar point. Such measurements were made in 1967 at Lincoln Laboratory (Pettengill et al., 1969) at a wavelength of 3.8 cm. The subradar point that year described a swath at about 22° north latitude, and the results showed a range of topographical variation of about 12 km. These results are shown in Figure 13. The results also indicated a wide variation in reflectivity with longitude (0.006 to 0.05 πa^2), and little or no correlation of topographical elevations with light or dark visible surface markings. Similarly, there was no correlation between radar reflectivity and topographical height.

During the 1969 opposition of Mars, many additional measurements were made by both Lincoln Laboratory (Rogers et al., 1970) at 3.8 cm and JPL (Goldstein et al., 1970) at 12.5 cm. This time, the subradar point ranged between 3° and 12° north latitude. It was found that the gross topographical features were quite similar,

with maximum relative elevation occurring in the region of about 90° west longitude (in the Tharsis area), and the total variation was around 12 km. After correlating the radar reflectivity results with visual luminance, it was decided that the optically dark regions had a significantly higher radar reflectivity and appeared to be smoother on the scale of 3.8-cm wavelength. Also, optically dark areas were found to preferentially lie on eastern slopes of topographical elevations (Lincoln Laboratory, 1970; Rogers et al., 1970). No other correlations appeared to be significant.

One of the measurements of the Mariner 6 radio occultation experiment (Kliore et al., 1970) was made at a point (4°N, 355°E) covered by the radar measurements. By comparing the pressure measured at this point (5.5 mb) to the relative elevation and extrapolating over the range of topographical variation, the surface pressure in the north equatorial regions of Mars can be inferred to vary from about 3 to 8 mb.

Mercury has been under planetary radar observation since 1962 (Kotelnikov et al., 1962). Recently, several large continent-sized surface features have been detected by means of depolarized spectrograms (Goldstein, 1970), which are shown in Figure 14.

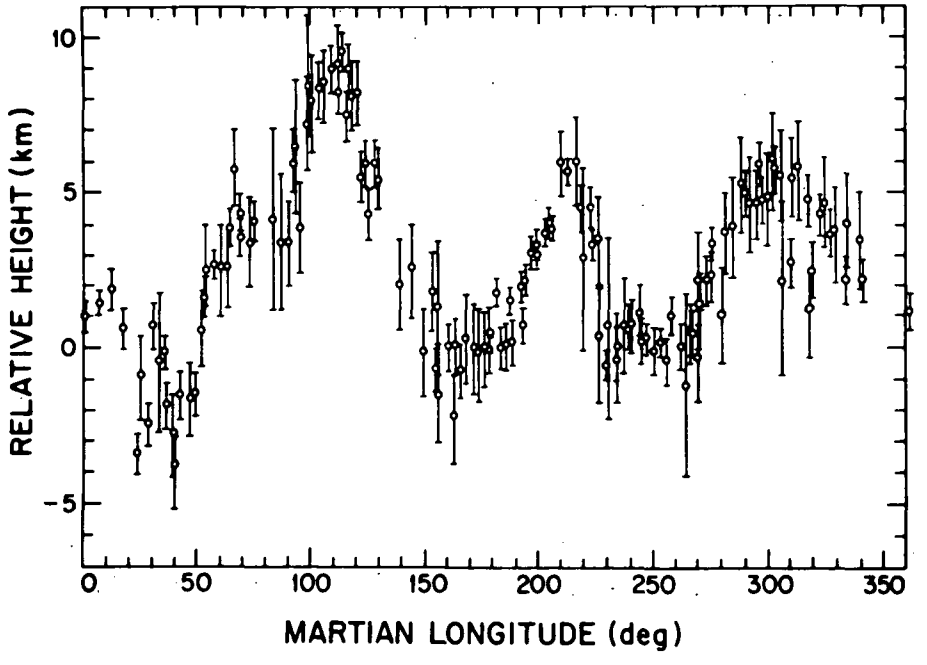


Figure 13.–Martian topography at 21° north latitude.

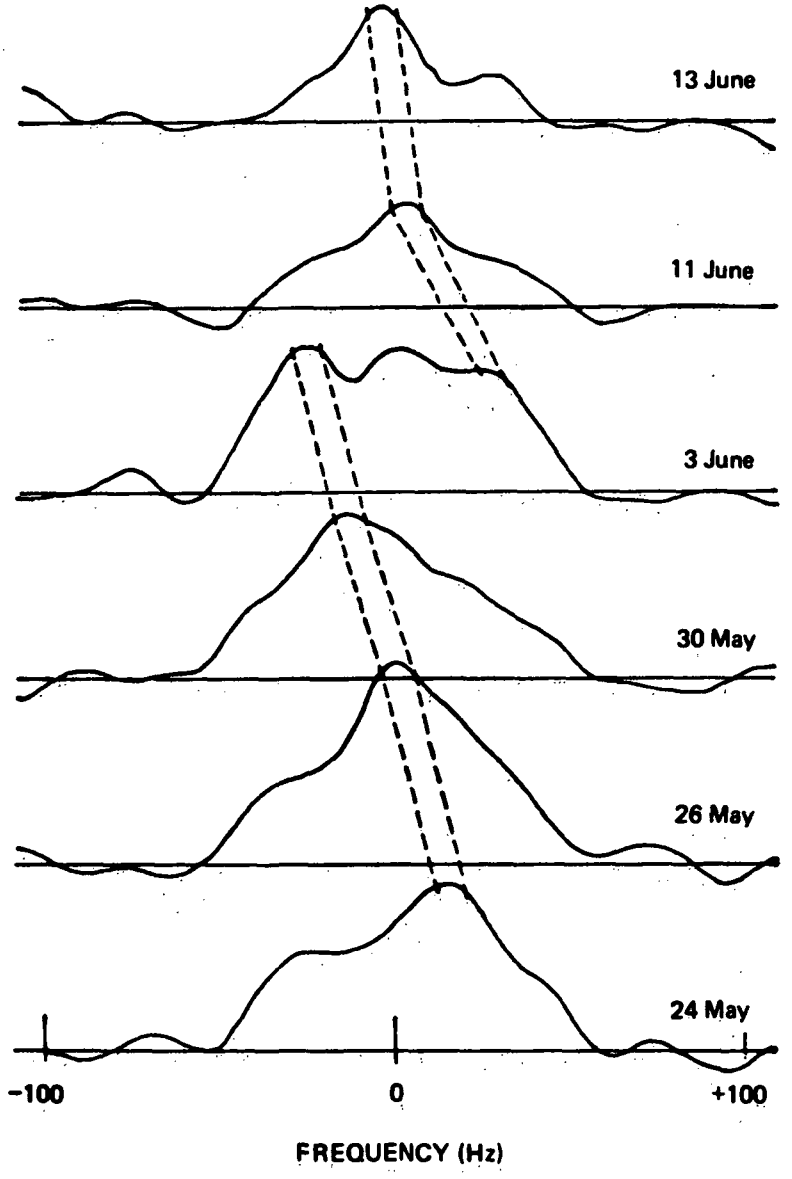


Figure 14.–Spectrograms of depolarized radar echoes from Mercury. Spectral salients, corresponding to surface features, are shown by dashed lines.

Prior to the radar investigations, the rotation period of Mercury was thought to be the same as its orbital period, about 88 days. However, Pettengill and Dyce (1965) obtained limb-to-limb Doppler spread data which contradicted the old value and led to the conclusion that the rotation period was about 59 days (Figure 15). This rotation period is very close to that corresponding to 3/2 synchronism with orbital motion (3 revolutions in 2 orbits) (Colombo and Shapiro, 1966). Subsequently, it has been established that the new rotation period is consistent with the results of visual observations (McGovern, Gross, and Rasool, 1965).

Measurements of surface-height variation by round-trip delay measurements to the subradar point (Smith et al., 1970) have revealed no asymmetry in the equatorial regions greater than the standard error (about 0.4 km). This is puzzling in view of the 3/2 orbital synchronism of the rotational and orbital periods.

Radar astronomy also offers a very valuable tool for the investigation of general relativity (Shapiro et al., 1968). For Mercury radar data, the excess delay time with respect to the Newtonian solution is seen to be in excellent agreement with the predictions based on general relativity (solid lines in Figure 16). Future measurements will make it possible to test competing theories of relativity.

Jupiter is also a very difficult radar target. Under ideal conditions, its echo power would be three times lower than that for Mars. However, because of its rapid rotation, the bandwidth is about 50 times that for Mars (11 000 times that for Venus), which greatly compounds the difficulties. The first attempts to detect Jupiter by radar occurred in 1963 (Goldstein, 1964; Kotelnikov et al., 1964). The JPL measurement at the 12.5-cm wavelength yielded a statistically significant return from a section of Jupiter's surface centered at 32° longitude (System I) only. Subsequent attempts to detect Jupiter at Arecibo at a wavelength of 70 cm (Dyce, Pettengill, and Sanchez, 1967) were unsuccessful, and it appears that the 1963 contact was a result either of fortuitous Jovian meteorological conditions or of a frequency-dependent reflection mechanism. Alternatively, the detection could have been spurious, for another attempt in 1964 at 12.5 cm yielded negative results.

In addition to planets, the asteroid Icarus was detected by radar as it passed at a distance of 6.5 million kilometers from the Earth on June 14, 1968. Detection was made at MIT's Haystack facility as well as at JPL's Goldstone radar (Goldstein, 1968). According to the Goldstone results, Icarus has a radius of about 0.5 km and a radar reflectivity cross section of about 13 percent.

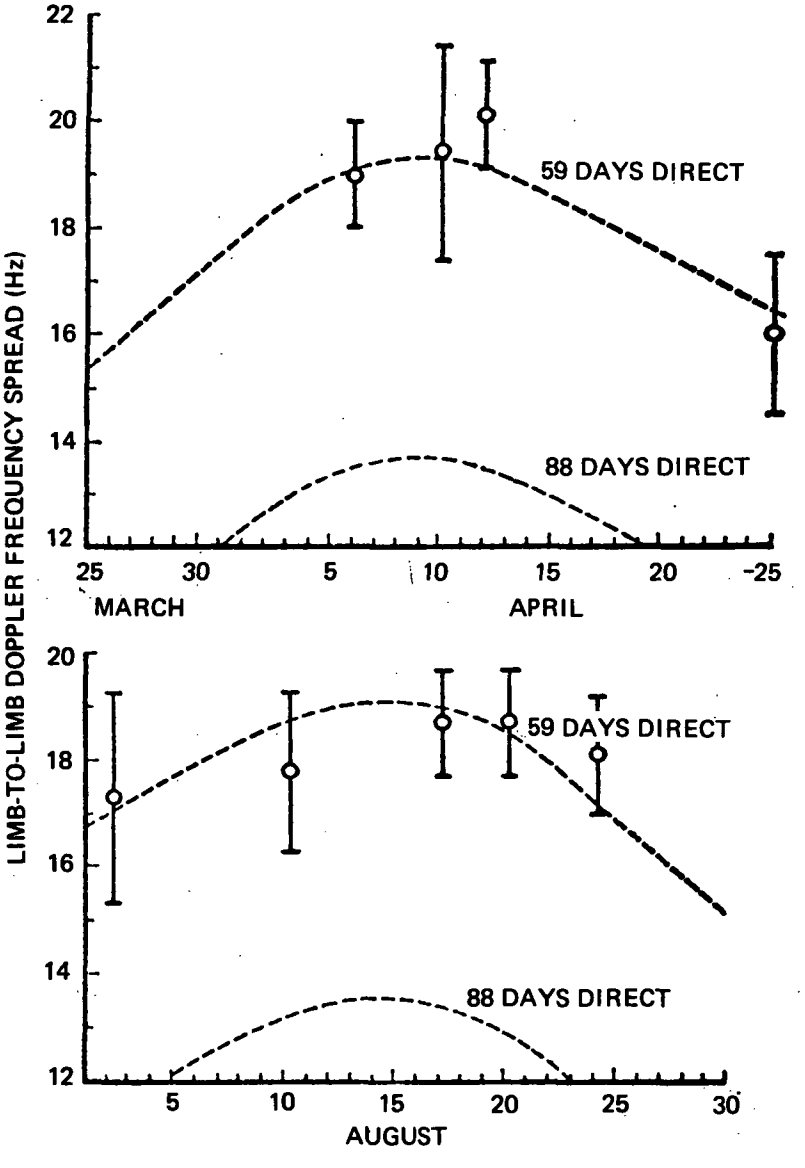


Figure 15.—Inferred limb-to-limb Doppler spread vs. date for observations of Mercury taken at the Arecibo Ionospheric Observatory during 1965. The *dotted curves show the theoretical variation for direct sidereal periods of* rotation of 59 and 88 days, under the assumption that the axis of rotation is normal to the orbital plane of Mercury.

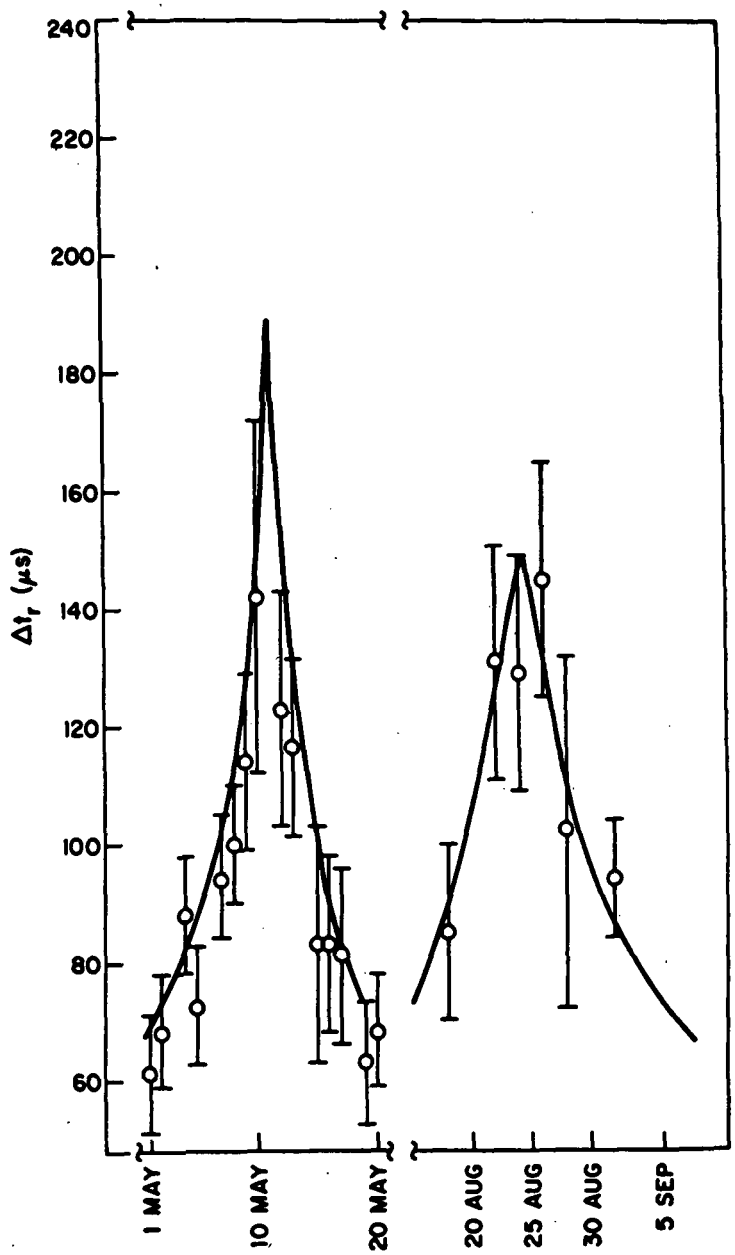


Figure 16.—Comparison of radar delay residuals for Mercury (in 1967) with respect to the Newtonian solution. Solid lines are predictions of the excess delay introduced by general relativity.

IV. RADIO OCCULTATION MEASUREMENTS

The first microwave measurements of a planetary brightness temperature from a spacecraft were performed in 1962 with the microwave radiometer on Mariner 2 (Barath et al., 1964), which confirmed the high temperatures previously observed with Earth-based radio telescopes. Since then, there have been no further spacecraft radiometry experiments; however, the microwave spectrum was gainfully exploited through the radio occultation technique (Fjeldbo, 1964; Kliore, Cain, and Hamilton, 1964). The radio technique is similar in principle to the stellar occultation technique, familiar to optical astronomy, in the sense that both use observations of a source as it is being occulted by the disk of a planet in order to obtain a measurement of the planet's atmosphere. However, there are major differences which make the radio technique far more valuable. First, in the case of stellar occultation, the source is noncoherent, and measurements of only its intensity can be made. In contrast, a spacecraft radio transmitter, which serves as the source in the radio occultation technique, emits a monochromatic signal of very high frequency and phase stability; thus, minute changes in phase due to the effects of a planetary atmosphere can be measured (less than 0.1λ, or about 1 cm). Second, the refractive defocusing attenuation is given by

$$A(h) = 1 - \frac{R_A R_E}{R_A + R_E} \frac{d\epsilon(h)}{dh},$$

where

R_A = distance from the source to the occulting limb,
R_E = distance from the Earth to the occulting limb,

and

$\epsilon(h)$ = angle of refraction at altitude h;

hence, for a stellar occultation, in which $R_A \gg R_E$, the attenuation varies as $R_E \epsilon'$ where R_E is typically about 10^8 km. In the case of radio occultation, the source is near the planet; thus, $R_E \gg R_A$, and the attenuation varies as $R_A \epsilon'$, where R_A is typically about 10^4 to 10^5 km. Obviously, if the dynamic ranges of both measurement techniques are equal, the radio occultation method is 10^3 to 10^4 times less sensitive to defocusing attenuation, which enables observations to be made far deeper into a planetary atmosphere than with the stellar method. Finally, the

deep-space tracking receivers used to acquire the occultation data have a signal-to-noise threshold at a received power level of about 10^{-21} W, which provides extremely accurate measurements of phase and frequency over a very large dynamic range. For example, in the case of the Mariner 5 Venus experiment (Kliore, Cain, Levy, Fjeldbo, and Rasool, 1967 and 1969) precise measurements were performed over a dynamic range of received power of four orders of magnitude, starting with 10^{-17} W.

The radio occultation method is based on the measurements of the changes in phase, frequency, and amplitude of the microwave radio signal from a spacecraft transmitter introduced by the effects of refraction in a planetary atmosphere. These changes are detected through comparison of the properties of the total received signal with predictions in the absence of a planetary atmosphere based on the orbit of the spacecraft. Because the changes in phase during the occultation period due to the motion of the spacecraft, the rotation of the Earth, and other dynamical effects amount to some 10^7 to 10^8 cycles at S-band (2.3 GHz), compared to a maximum planetary contribution of about 30 cycles, the dynamical effects must be known and removed with a very high accuracy.

The changes introduced in the signal depend upon the variation of the refractive index with height in the planetary atmosphere. The index of refraction is defined as

$$n(h) = c/v(h) ,$$

where

c = velocity of propagation in free space

and

$v(h)$ = velocity of propagation in the medium in question.

It is often convenient to use the refractivity, which is defined as

$$N(h) = 10^6 \times [n(h) - 1] .$$

For example, the refractivity at sea level on the Earth is about 300, whereas at the surface of Mars it is about 3.5.

In the neutral atmosphere, the phase velocity of radio waves is reduced, and, hence, $n > 1$. In the ionosphere, the phase velocity is advanced, and $n < 1$. Although in general the index of refraction in a planetary atmosphere and ionosphere varies with latitude, longitude, and time, as well as height, for the purposes of analysis of radio occultation data, spherical symmetry is assumed.

The total phase delay $\Delta\Phi$ is given by the difference between the straight-line path from the spacecraft to the Earth and the actual phase path through the planetary atmosphere. As an approximation valid for small refractive bending angles (at Mars, the maximum bending angle is less than 1 mrad), the phase delay (in cycles) is

$$\Delta\Phi = (f/c)(\Delta r + \Delta l) ,$$

where

f = frequency,

Δr = apparent phase-path increase due to propagation delay,

and

Δl = phase-path increase due to refractive bending.

The phase-path increase due to propagation delay is given by

$$\Delta r = 2 \int_S (n - 1)\, ds .$$

For the case of negligible bending, such as is encountered at Mars, the propagation path can be approximated by a straight line, resulting in

$$\Delta r = 2 \int_R^\infty \frac{(n - 1)\, dr}{\sqrt{1 - (R/r)^2}} ,$$

where R is the closest approach radius of the radio ray. The contribution of the refractive bending can likewise be approximated by

$$\Delta l = R_A(1 - \cos\epsilon) \approx R_A(\epsilon^2/2) ,$$

where R_A is the distance from the spacecraft to the occulting limb. The bending angle itself is given by

$$\epsilon = -2n(R)R \int_R^\infty \frac{n'(r)\, dr}{n(r)\sqrt{n^2r^2 - n(R)R^2}} .$$

The change in frequency that is observed is simply the time derivative of the phase change:

$$\Delta f = \frac{d}{dt}\Delta\Phi .$$

In dense atmospheres, such as those of Venus and Jupiter, the approximate relationships given above are not valid, and the rigorous expressions which take into account significant refractive bending must be used. (The maximum refractive bending observed in the atmosphere of Venus just prior to the extinction of the signal was about 17 deg.) It should be pointed out that in the case of Venus, extinction of the signal was not caused by occultation by the solid limb but by excessive defocusing attenuation just above the level of critical refraction. Critical refraction occurs when the radius of curvature of the refracted ray is equal to the distance to the center of the planet. This condition occurs when

$$rn'(r) + n(r) = 0 .$$

For Venus, it was found that critical refraction occurs at an altitude of about 35 km, when the atmospheric pressure is about 7 atm. On Jupiter, because of its very large radius, critical refraction can be expected to occur at a pressure level of 1 to 3 atm, depending on the H_2/He ratio, and, therefore, no further tangential penetration of a radio ray is possible.

Figure 17 is a simplified diagram of a radio occultation data acquisition system. A signal whose frequency is referenced to a rubidium standard (soon to be replaced with a hydrogen maser) is transmitted to the spacecraft. The spacecraft transponder coherently retransmits the received signal after changing its frequency slightly in order to avoid confusion. In this configuration, the radio beam traverses the planetary atmosphere twice (two way). If there is no signal transmitted to the spacecraft, it transmits a signal referenced to its own oscillator, and only the downlink signal passes through the atmosphere (one way).

On the ground, the signal is received by a Deep Space Network tracking station. (The Goldstone Mars station has an antenna 210 feet in diameter.) The receiver employs a maser front end cooled by a closed-cycle liquid-helium cryostat to obtain extremely low-noise amplification. From there, the signal passes through a phase-locked-loop receiver and a tracking-data handling system that provides a record of nondestructively counted biased Doppler cycle counts at time intervals of 0.1 or 1.0 s. Because the phase-locked receiver requires a finite interval of time to

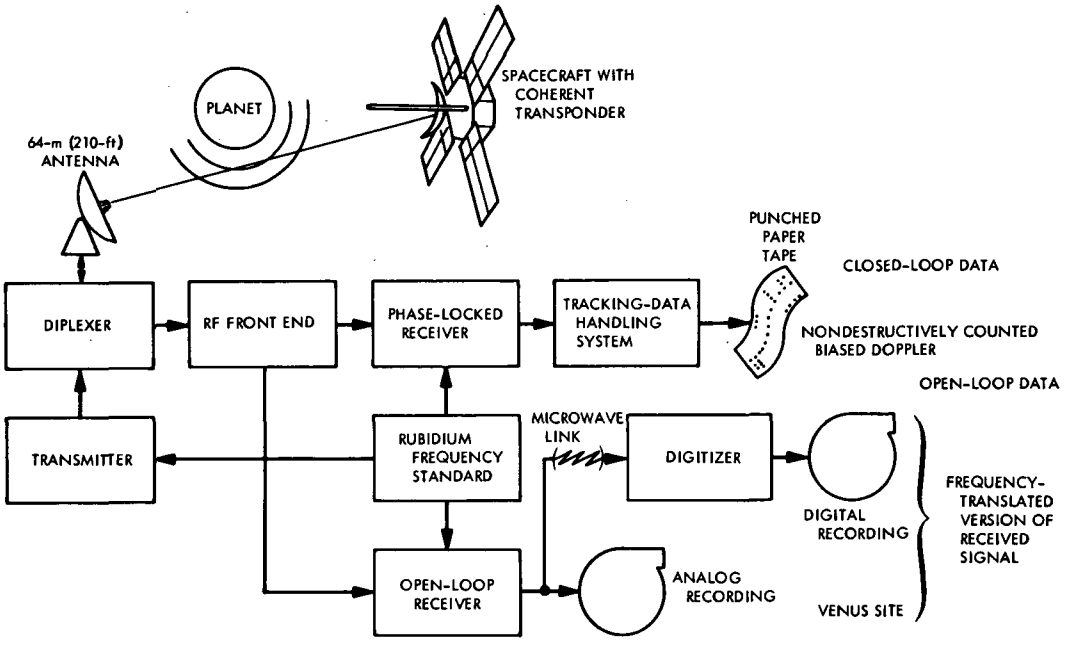


Figure 17.–Occultation data acquisition system.

reacquire lock following reappearance of the signal after occultation, an open-loop receiver also is used to record the data. This receiver produces a signal that is frequency translated to the audio range while maintaining phase integrity with the original signal. This allows one to preserve all of the phase variations present in the 2.3-GHz signal in the audio-frequency tape recording. These analog records are then digitized and preserved on digital magnetic tape (Levy, Otoshi, and Seidel, 1966; Kliore and Seidel, 1969).

The closed-loop data are then processed by the tracking data reduction programs of the JPL navigation system. In the final step of this process, the orbit determination program compares the obtained Doppler measurements with predictions based on the motion of the spacecraft and produces residuals. In the case of the open-loop data, more preprocessing is necessary, including passage of the digitized signal through a digital phase-locked receiver program that determines the frequency and strength of the signal. The frequency is then compared with predictions to yield the open-loop residuals. The signal-strength data are used to determine the precise times of loss and reappearance of signal during occultation and to study diffraction, attenuation, and absorption phenomena.

The residuals are then processed to remove any bias or rate terms present due to inaccuracies in the orbit, and this leads to data similar to that shown in Figure 18 (Kliore, Cain, and Levy, 1967). The S-shaped feature at about 02:30:10 is due to the ionosphere of Mars, and the sharp upswing near the end of the data represents the effect of the neutral atmosphere.

The frequency residuals are then integrated to produce the total phase difference $\Delta\Phi$ (Figure 19). Note the clearly visible effect of a secondary ionospheric maximum at about 02:30:25. The smoothness of the curve in Figure 19 indicates the stability and accuracy of the data acquisition systems.

In further processing, the knowledge of the spacecraft trajectory must be brought into play. One may use it together with a simple model of the planetary atmosphere in order to determine the parameters of the atmosphere (Kliore et al., 1965 and 1966). Alternatively, the more unambiguous method of direct integral inversion may be used (Fjeldbo and Eshleman, 1968; Phinney and Anderson, 1968; Kliore, 1971). In this method, the bending angle is first determined as follows. The

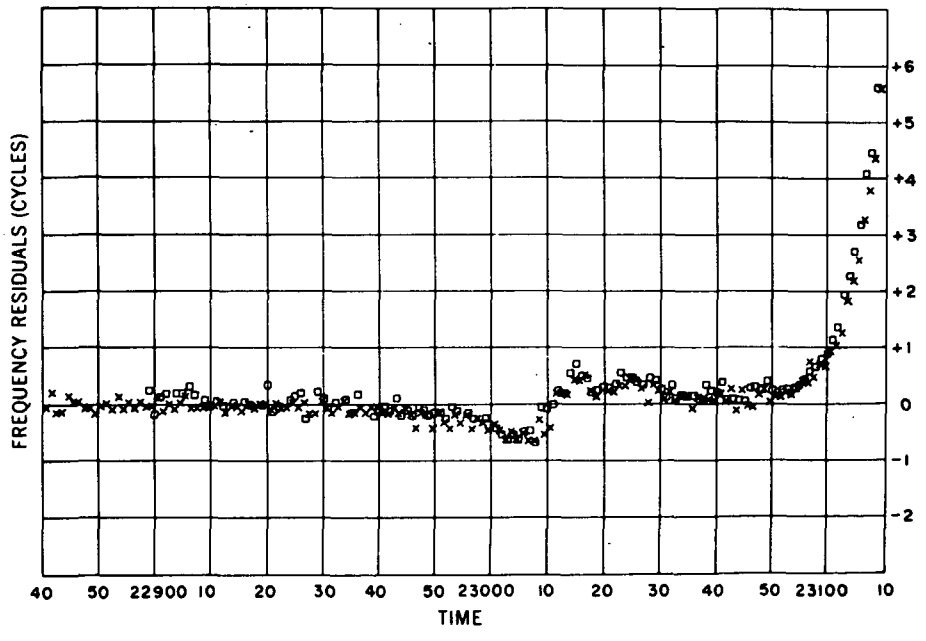


Figure 18.—Mariner 4 frequency residuals. Entrance occultation data points are denoted by squares for open loop, X's for closed loop.

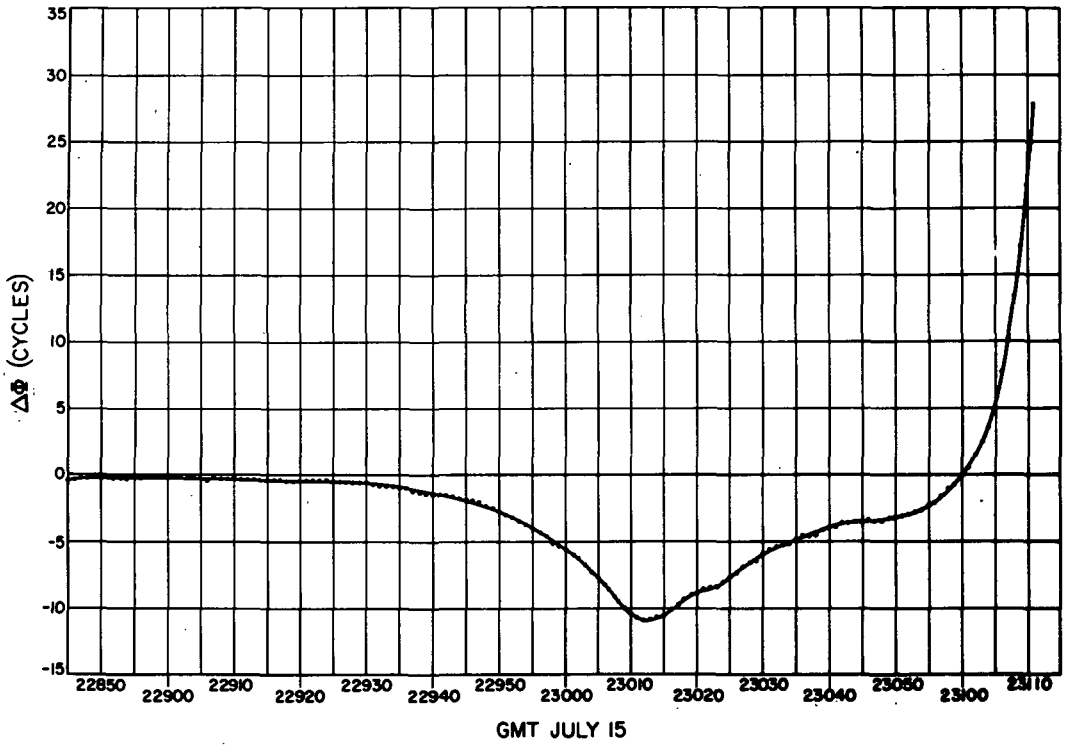


Figure 19.—Mariner 4 phase-difference data.

component of the planet-centered spacecraft velocity in the direction of the Earth is

$$V_E = \mathbf{V} \cdot \mathbf{E} ,$$

where

V = planet-centered velocity in the plane containing the spacecraft and the centers of the Earth and the planet

and

E = unit vector in the direction of the Earth.

The Doppler frequency expected to be seen at the Earth is then

$$\Delta f_E = (f/c)|V_E| .$$

Finally, the bending angle is given by

$$\epsilon = \psi_E - \psi ,$$

where

$$\psi = \cos^{-1}[(c/f|\mathbf{V}|)(\Delta f_E + \Delta f)]$$

and

$$\psi_E = \cos^{-1}(V_E/|\mathbf{V}|) ,$$

and Δf is the observed frequency residual.

The asymptotic distance, or miss parameter, of the ray is given by

$$p_i = |\mathbf{M}_i| \sin\alpha ,$$

where $\mathbf{M}_i$ is the planet-centered position vector of the spacecraft and α is ϵ plus the angle subtended by the Earth and the center of the planet at the spacecraft.

The quantities ϵ and p are necessary in order to separate as follows the phase change due to retardation from the total phase change:

$$\Delta r = (c/f)\Delta\Phi - \frac{1}{2}\sqrt{|\mathbf{M}|^2 - p^2\epsilon^2} .$$

The integral relating ΔR to $R(r)$ can now be inverted through the use of the Abel integral transform to yield

$$N(r) = 10^6[n(r) - 1] = \frac{10^6}{\pi r^2}\int_{r_M}^{r} \frac{p[\Delta r(p) + p\Delta r'(p)]}{\sqrt{p^2 - r^2}}\,dp ,$$

where r_M equals p_{max}.

The profiles of refractivity obtained in this manner from the Mariner 6 and Mariner 7 data are shown in Figure 20 (Kliore et al., 1970). One can clearly see the negatively refractive ionosphere with its secondary peak at about 115 km and the

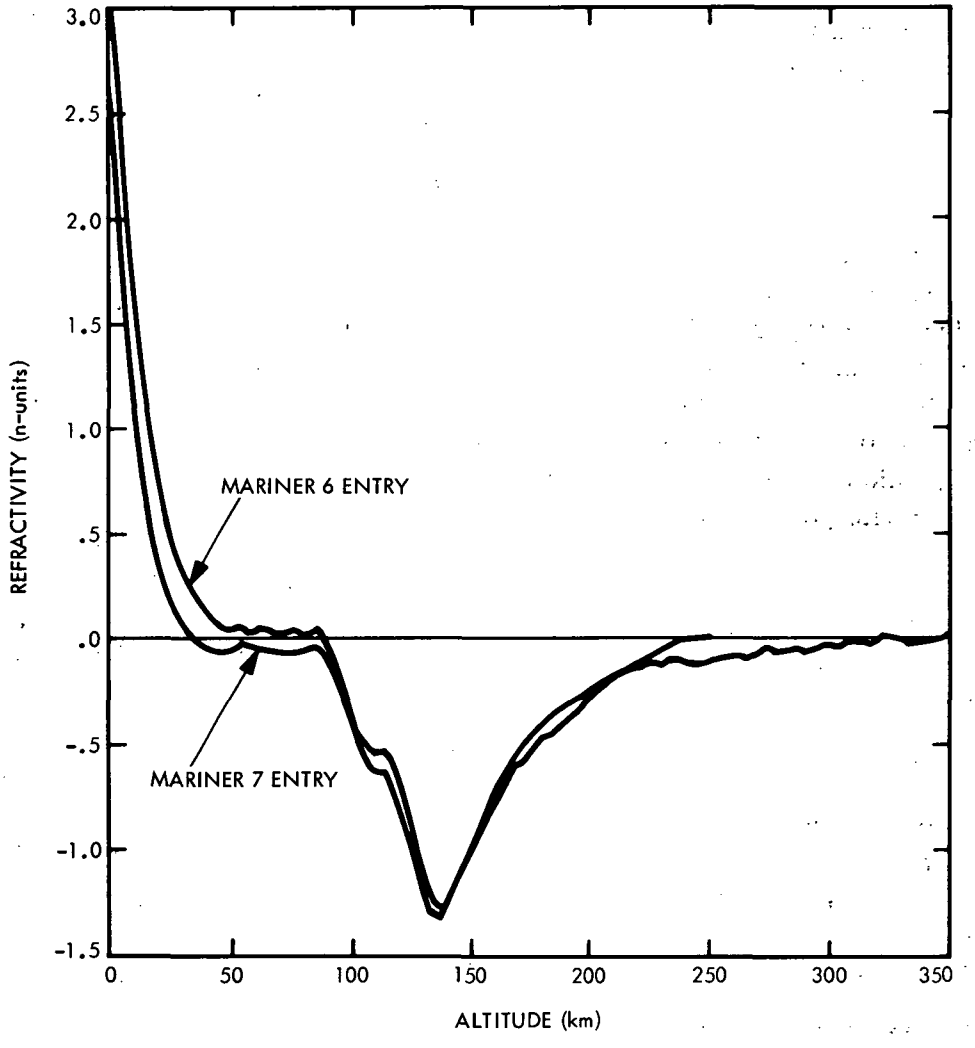


Figure 20.–Mariner 6 and 7 entry-refractivity profiles.

positive-refractivity regions near the surface, representing the neutral atmosphere. The nonzero refractivity in the interval between the bottom of the ionosphere and the top of the neutral atmosphere is an artifact due to noise.

The negative refractivities obtained for the ionosphere can be converted immediately to electron density as follows:

$$N_e(r) = \frac{-f^2 N(r)}{4.03 \times 10^{13}} \text{ (cm}^{-3}\text{)},$$

where f is the frequency (Hz) and N is the refractivity (n-units).

The electron density profiles obtained from the Mariner 6 and 7 data are shown in Figure 21 (Fjeldbo, Kliore, and Seidel, 1970). The two profiles are remarkably similar, as they were obtained under nearly equal solar illumination conditions.

In the neutral atmosphere, an assumption of composition must be made in order to obtain the pressure and temperature from refractivity data. When the composition is known, refractivity can be converted to mass density and the atmospheric pressure can be obtained by downward integration of the hydrostatic equation:

$$P(r) = -\bar{m}/RQ \int_{r_M}^{r} N(r) g(r)\, dr ,$$

where

$g(r) = g_0 (r_0/r)^2$ (g_0 is the acceleration of gravity at $r = r_0$),

R = universal gas constant,

Q = specific refractivity of assumed gas mixture (refractivity at 760 mm Hg and 0° C),

and

$\bar{m}$ = mean molecular weight.

The temperature is then given by

$$T = PQ/N .$$

These relationships are valid if there is no condensation of volatiles. If the temperature obtained above falls below, for instance, the saturation temperature of carbon dioxide at a given pressure, the pressure and temperature computations must be iterated until the saturation law is satisfied. Examples of atmospheric pressure profiles obtained in this manner from the Mariner 6 and 7 data are shown in Figure

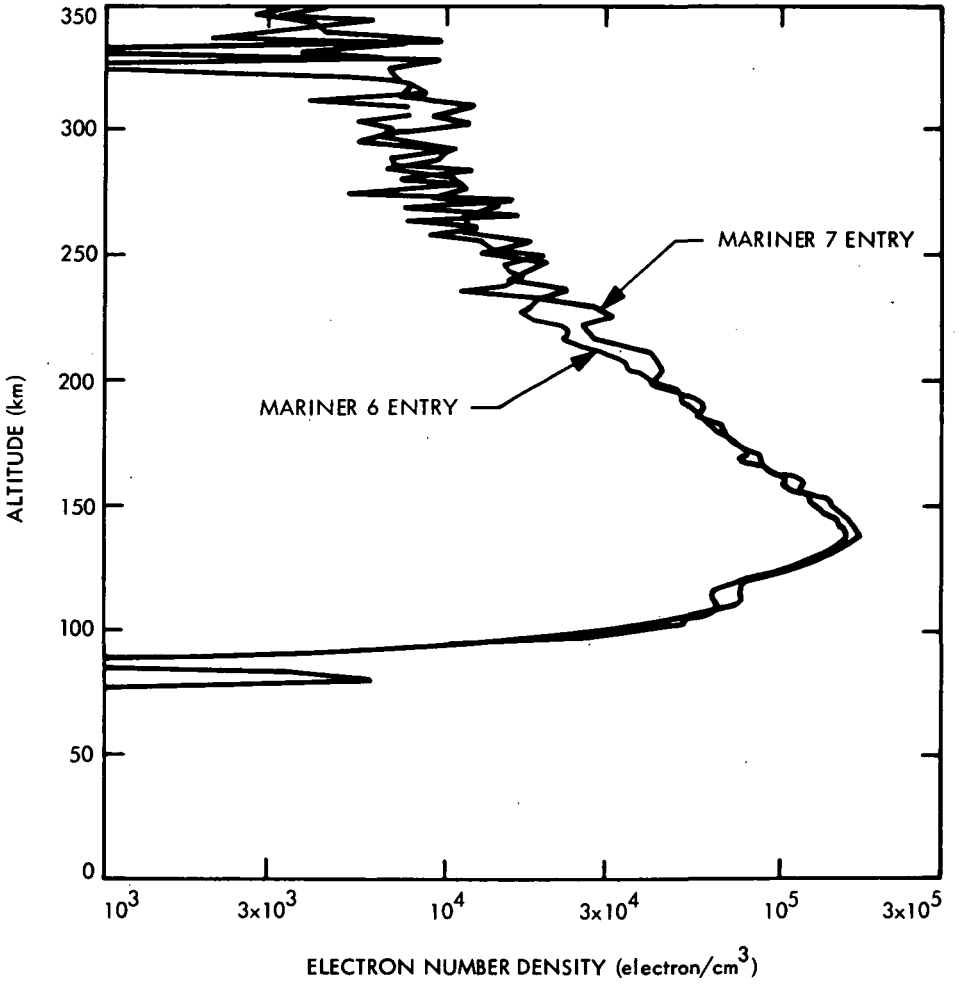


Figure 21.—Electron density profiles in the Martian ionosphere.

22. The observed differences in the surface pressure are most likely a result of local elevation differences relative to the mean gravitational equipotential surface of Mars (consistent with the topographical variations measured with Earth-based radar as presented in Section III). The differing slopes of the pressure profiles indicate different scale-height variations, which are a consequence of the wide variety of temperatures found in the measurements. Since the occultation locations ranged from 4° north latitude in the daytime (Mariner 6 entry) to 79° north latitude at

night (Mariner 6 exit), the observed temperatures are consistent with theoretical predictions and with the measurements of the surface temperatures performed with the Mariner 6 and 7 infrared radiometer instruments. Figure 23 shows a cross plot of the pressures and temperatures, indicating that condensation of carbon dioxide can be expected at altitudes from about 15 to about 30 km on Mars.

During the Mariner 5 mission to Venus in 1967, a dual-frequency radio propagation experiment (Stanford Mariner Group, 1967; Fjeldbo and Eshleman, 1969) was conducted in addition to the S-band (2.3 GHz) experiment. This experiment, carried out by investigators from Stanford University, utilized an onboard receiver operated at frequencies of about 423 and 50 MHz. Signals at those

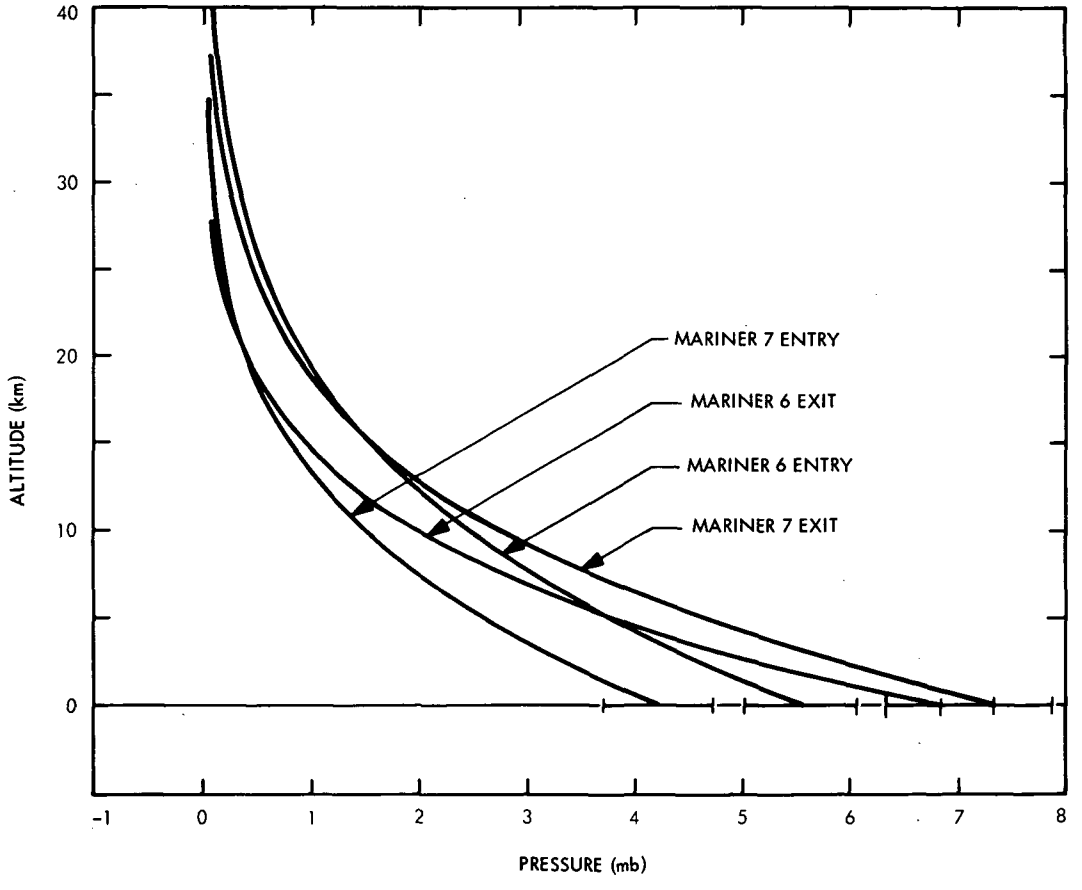


Figure 22.—Atmospheric pressure profiles derived from Mariner 6 and 7 data.

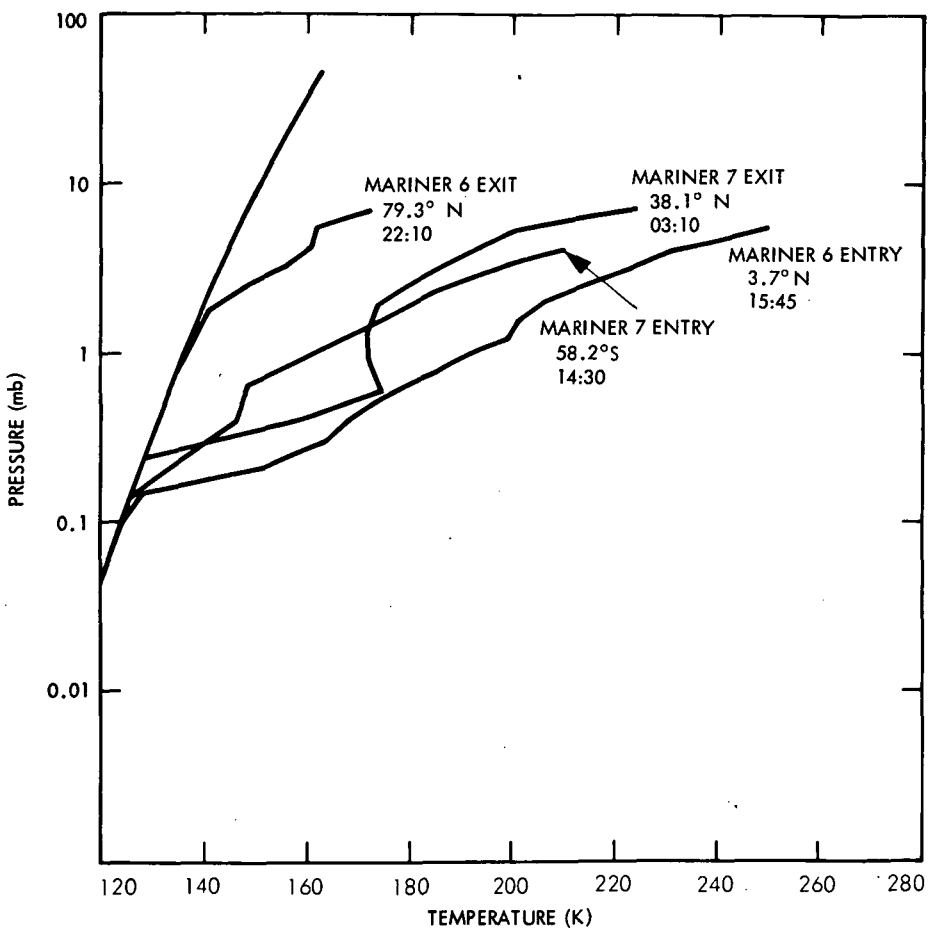


Figure 23.—Pressure-temperature plots from Mariner 6 and 7 data. The smooth solid curve at the left represents the carbon-dioxide saturation boundary.

two frequencies were transmitted to the dual-frequency receiver from the 150-foot-diameter radio telescope antenna at Stanford, and the amplitudes, frequencies, and differential Doppler frequencies were encoded and telemetered to the tracking stations on Earth along with other scientific telemetry data. Although the gradients in both the nighttime and daytime ionosphere of Venus caused multipath effects which caused both receivers to lose lock, much valuable information was received on the nighttime ionosphere and the daytime plasmapause of Venus. A serious limitation of such uplink experiments is the necessity to transmit information over the spacecraft telemetry link, which limits the amount of data that can be recovered.

In the future, S-band experiments will be conducted with the Mariner 1971 Mars orbiters, which will provide roughly 100 occultation measurements during the mission. Also, the Pioneer F Jupiter probe due to be launched in 1972 will carry out S-band radio occultation measurements of that planet's atmosphere. The Mariner mission in 1973 to Mercury, with a gravity assist at Venus, will have an X-band downlink system (3.8 cm) coherent with the S-band signal. Such a dual-frequency system will greatly expand the capability of the radio occultation method to investigate the atmospheres and ionospheres of Venus and Mercury, as well as provide spacecraft tracking data corrected for the effect of interplanetary charged particles. Although it currently is only in the planning stages, the proposed "Grand Tour" missions to be flown in the late 1970's to Jupiter, Saturn, Uranus, Neptune, and Pluto will most likely use a two-frequency downlink communications system to provide fascinating information on the radio propagation characteristics of their atmospheres.

ACKNOWLEDGMENTS

The author is grateful to G. H. Pettengill, J. V. Evans, T. Hagfors, D. D. Morrison, J. W. Warwick, A. D. Kuzmin, D. O. Muhleman, S. Gulkis, R. M. Goldstein, R. L. Carpenter, A. E. Rogers, I. I. Shapiro, W. G. Melbourne, and D. A. O'Handley for kindly supplying reprints of their work and other materials which were used in the compilation of this paper.

REFERENCES

Ash, M. E., Ingalls, R. P., Pettengill, G. H., Shapiro, I. I., Smith, W. B., Slade, M. A., Campbell, D. B., Dyce, R. B., Jurgens, R., and Thompson, T. W., "The Case for the Radar Radius of Venus", *J. Atmos. Sci.* 25:560-563, 1968.

Barath, F. T., Barrett, A. H., Copeland, J., Jones, D. E., and Lilley, A. E., "Mariner II Microwave Radiometer Experiment and Results", *Astron. J.* 69:49-58, 1964.

Barrett, A. H., "Passive Radio Observations of Mercury, Venus, Mars, Saturn, and Uranus", *Radio Sci.* 69D:1565-1573, 1965.

Carpenter, R. L., "Study of Venus by CW Radar–1964 Results", *Astron. J.* 71:142-152, 1966.

Colombo, G., and Shapiro, I. I., "The Rotation of the Planet Mercury", *Astrophys. J.* 145:246-307, 1966.

Cook, J. J., Cross, L. G., Bair, M. E., and Arnold, C. B., "Radio Detection of the Planet Saturn", *Nature* 188:393-394, 1960.

Dent, W. A., Klein, M. J., and Aller, H. D., "Measurements of Mars at λ 3.75 cm From February to June 1965", *Astrophys. J.* 142:1685-1688, 1965.

Dickel, J. R., "Measurement of the Brightness Temperature of Venus Over a Full Cycle of Planetary Phase Angles", *Icarus* 5:305, 1966.

Dickel, J. R., Warnock, W. W., and Medd, W. J., "Lack of Phase Variation of Venus", *Nature* 220:1183-1185, 1968.

Drake, F. D., "Microwave Observations of Venus, 1962-63", *Astron. J.* 69:62-64, 1964.

Dyce, R. B., Pettengill, G. H., and Sanchez, A. D., "Radar Observations of Mars and Jupiter at 20 cm", *Astron. J.* 72:771-777, 1967.

Dyce, R. B., Pettengill, G. H., and Shapiro, I. I., "Radar Determination of the Rotations of Venus and Mercury", *Astron. J.* 72:351-359, 1967.

Epstein, E. E., "Mars, Jupiter and Saturn 3.4 mm Brightness Temperature", *Astrophys. J.* (*Letters*) 151:L149, 1968.

Epstein, E. E., Oliver, J. P., Soter, S. L., Schorn, R. A., and Wilson, W. J., "Venus: On an Inverse Variation with Phase in the 3.4 mm Emission During 1965 Thru 1967", *Astron. J.* 73:271-274, 1968.

Epstein, E. E., Dworetsky, M. M., Fogarty, W. G., Montgomery, J. W., and Cooley, R. C., "Mercury: Epilith Physical Parameters and Hermocentric Longitude Dependence of Its 3.3 mm. Radiation", *Radio Sci.* 5:401-409, 1970.

Evans, J. V., and Hagfors, T., *Radar Astronomy*, McGraw-Hill Book Co., New York, 1968.

Evans, J. V., "Radar Surveys of the Solar System", *Proc. Amer. Phil. Soc.* 113:203-223, 1969a.

Evans, J. V., "Radar Studies of Planetary Surfaces", *Ann. Rev. Astron. Astrophys.* 7:201-248, 1969b.

Fjeldbo, G., "Bi-Static Radar Methods for Studying Planetary Atmospheres and Surfaces", Stanford University Radio Science Laboratory Report SU-SEL 64-025, Stanford, California, 1964.

Fjeldbo, G., and Eshleman, V. R., "The Atmosphere of Mars Analyzed by Integral Inversion of the Mariner IV Occultation Data", *Planet. Space Sci.* 16:1035-1059, 1968.

Fjeldbo, G., and Eshleman, V. R., "Atmosphere of Venus as Studied with the Mariner V Dual Radio Frequency Occultation Experiment", *Radio Sci.* 4:879-897, 1969.

Fjeldbo, G., Kliore, A. J., and Seidel, B. L., "The Mariner '69 Occultation Measurements of the Upper Atmosphere of Mars", *Radio Sci.* 5:381-386, 1970.

Goldreich, P., and Peale, S. J., "Is the Rotation of Venus Locked with the Earth?", *Nature* 209:1117, 1966.

Goldstein, R. M., and Carpenter, R. L., "Rotation of Venus: Period Estimated from Radar Measurement", *Science* 139:910-991, 1963.

Goldstein, R. M., and Gillmore, W. F., "Radar Observations of Mars", *Science* 141:1171-1172, 1963.

Goldstein, R. M., "Radar Observations of Jupiter", *Science* 144:842-843, 1964.

Goldstein, R. M., "Preliminary Venus Radar Results", *J. Res. Nat. Bur. Stand.* 69D:1623-1625, 1965.

Goldstein, R. M., "Radar Observations of Icarus", *Science* 162:903-904, 1968.

Goldstein, R. M., "Mercury: Surface Features Observed During Radar Studies", *Science* 168:467-468, 1970.

Goldstein, R. M., Melbourne, W. G., Morris, G. A., Downs, G. S., and O'Handley, D. A., "Preliminary Radar Results of Mars", *Radio Sci.* 5:475-478, 1970.

Goldstein, R. M., and Rumsey, H., Jr., "A Radar Snap Shot of Venus", *Science*, 1970.

Gulkis, S., MacDonald, T. R., and Craft, H., "The Microwave Spectrum of Saturn", *Icarus* 10:421-427, 1969.

Howard, W. E., III, Barrett, A. H., and Haddock, F. T., "Measurement of Microwave Radiation From the Planet Mercury", *Astrophys. J.* 136:995, 1962.

Ingalls, R. P., and Evans, J. V., "Scattering Properties of Venus at 3.8 cm", *Astron. J.* 74:258-272, 1969.

Kellerman, K. I., "The Thermal Radio Emission from Mercury, Venus, Mars, Saturn and Uranus", *Icarus* 5:478-490, 1966.

Kellerman, K. I., "Thermal Radio Emission From the Major Planets", *Radio Sci.* 5:487-493, 1970.

Klein, M. J., "Mercury: Recent Observations at 3.75 cm Wavelength Summary", *Radio Sci.* 5:397-400, 1970.

Kliore, A., Cain, D. L., and Hamilton, T. W., "Determination of Some Physical Properties of the Atmosphere of Mars from Changes in the Doppler Signal of a Spacecraft in an Earth Occultation Trajectory", California Institute of Technology Jet Propulsion Laboratory Technical Report 32-674, Pasadena, California, 1964.

Kliore, A., Cain, D. L., Levy, G. S., Eshleman, V. R., Fjeldbo, G., and Drake, F. D., "Occultation Experiment: Results of the First Direct Measurement of Mars' Atmosphere and Ionosphere", *Science* 149:1243-1248, 1965.

Kliore, A., Cain, D. L., Levy, G. S., Eshleman, V. R., Fjeldbo, G., and Drake, F. D., "Preliminary Results of the Mariner IV Occultation Measurement of the Atmosphere of Mars", *Proceedings of the Cal-Tech/JPL Lunar and Planetary Conference*, 1966, pp. 257-266.

Kliore, A., Cain, D. L., and Levy, G. S., "Radio Occultation Measurement of the Martian Atmosphere over Two Regions with the Mariner IV Space Probe", in *Space Research VII–Moon and Planets*, North Holland Publishing Company, Amsterdam, 1967.

Kliore, A., Cain, D. L., Levy, G. S., Fjeldbo, G., and Rasool, S. I., "Atmosphere and Ionosphere of Venus from the Mariner V S-Band Radio Occultation Measurement", *Science* 158:1683-1688, 1967.

Kliore, A., Cain, D. L., Levy, G. S., Fjeldbo, G., and Rasool, S. I., "Structure of the Atmosphere of Venus Derived from Mariner V S-Band Measurements", in *Space Research IX–Moon and Planets*, North Holland Publishing Company, Amsterdam, 1969.

Kliore, A., and Seidel, B. L., "S-Band Occultation Experiments", in *Advanced Space Experiments–Advances in the Astronautical Sciences*, O. L. Tiffany and E. Zaitseff, eds., American Astronautical Society, 1969, Vol. 25, pp. 79-101.

Kliore, A., Fjeldbo, G., and Seidel, B. L., "Summary of Mariner 6 and 7 Radio Occultation Results on the Atmosphere of Mars", Paper No. M. 25, 13th Plenary Meeting of COSPAR, Leningrad, U.S.S.R., May 20 to 29, 1970.

Kliore, A., "Current Methods of Radio Occultation Data Inversion", *Proceedings of the Workshop on the Mathematics of Data Inversion*, NASA Ames Research Center, Moffett Field, California, July 12, 1971.

Kotelnikov, V. A., Dubrovin. V. M., Marozov, V. A., Petrov, G. M., Rzhiga, O. N., Trundva, V. G., and Shakovskoi, A. M., "Results of Radar Observations of Venus in 1961", *Radiotekh. Elektron.* 7:1860-1872, 1962.

Kotelnikov, V. A., et al., "Radar Detection of the Planet Jupiter", *Dokl. Akad. Nauk.* 155:1037-1038, 1964.

Kraus, J. D., *Radio Astronomy*, McGraw-Hill Book Co., New York, 1966.

Levy, G. S., Otoshi, T. Y., and Seidel, B. L., "Ground Instrumentation for the Mariner IV Occultation Experiment", California Institute of Technology Jet Propulsion Laboratory Technical Report 32-984, Pasadena, California, 1966.

Lincoln Laboratory, *Radar Studies of Mars* (final report), Massachusetts Institute of Technology, Lexington, Massachusetts, January 15, 1970.

Low, F. J., "Observations of Venus, Jupiter and Saturn at λ 20μ", *Astron. J.* 71:391, 1966.

McGovern, W. E., Gross, S. H., and Rasool, S. I., "Rotation Period of the Planet Mercury", *Nature* 208:375, 1965.

Mayer, C. H., McCullough, T. P., and Sloanaker, R. M., "Observation of Mars and Jupiter at a Wavelength of 3.15 cm", *Astrophys. J.* 127:11-16, 1958.

Mayer, C. H., McCullough, T. P., and Sloanaker, R. M., "3.15 cm Observations of Venus in 1961", *Mem. Soc. Roy. Sci. Liège* 7:357, 1963.

Morrison, D., and Klein, M. J., "The Microwave Spectrum of Mercury", *Astrophys. J.* 160:325, 1970.

Muhleman, D. O., "Planetary Characteristics from Radar Observations", *Space Sci. Rev.* 6:341-364, 1966.

Pettengill, G. H., Briscoe, H. W., Evans, J. V., Gehrels, E., Hyde, G. M., Kraft, L. G., Price, R., and Smith, W. B., "A Radar Investigation of Venus", *Astron. J.* 67:181-190, 1962.

Pettengill, G. H., "A Review of Radar Studies of Planetary Surfaces", *J. Res. Nat. Bur. Stand.* 69D:1617-1623, 1965.

Pettengill, G. H., and Dyce, R. B., "A Radar Determination of the Rotation of the Planet Mercury", *Nature* 206:1240, 1965.

Pettengill, G. H., Counselman, C. C., Rainville, L. P., and Shapiro, I. I., "Radar Measurements of Martian Topography", *Astron. J.* 74:461, 1969.

Phinney, R. A., and Anderson, J. D., "On the Radio Occultation Method for Studying Planetary Atmospheres", *J. Geophys. Res.* 73:1819, 1968.

Rogers, A. E. E., and Ingalls, R. P., "Venus: Mapping the Surface Reflectivity by Radar Interferometry", *Science* 165:797-799, 1969.

Rogers, A. E. E., Ash, M. E., Counselman, C. C., Shapiro, I. I., and Pettengill, G. H., "Radar Measurements of the Surface Topography and Roughness of Mars", *Radio Sci.* 5:465-473, 1970.

Shapiro, I. I., "Theory of the Radar Determination of Planetary Rotations", *Astron. J.* 72:1309, 1967.

Shapiro, I. I., et al., "Fourth Test of General Relativity: Preliminary Results", *Phys. Rev. Lett.* 20:1265-1269, 1968.

Sinclair, A. C. E., Basart, J. P., Buhl, D., Gale, W. A., and Liwshitz, M., "Preliminary Results of Interferometric Observations of Venus at 11.1 cm Wavelength", *Radio Sci.* 5:347-354, 1970.

Slee, O. B., "A Search for Radio Emission From Uranus", *Astrophys. J.* 140:823-824, 1964.
Smith, W. B., Ingalls, R. P., Shapiro, I. I., and Ash, M. E., "Surface Height Variations on Venus and Mercury", *Radio Sci.* 5:411-423, 1970.
Stanford Mariner Group, "Venus: Ionosphere and Atmosphere as Measured by Dual Frequency Radio Occultation of Mariner V", *Science* 158:1678-1683, 1967.
Victor, W. K., and Stevens, R., "Exploration of Venus by Radar", *Science* 134:46-48, 1961.
Zohar, S., and Goldstein, R. M., "Venus Map: A Detailed Look at the Feature", *Nature* 219:357-358, 1968.

BIBLIOGRAPHY

Passive Radio Astronomy

Barrett, A. H., and Staelin, D. H., "Radio Observations of Venus and Their Interpretations", *Space Sci. Rev.* 3:109-135, 1964.
Barrett, A. H., "Passive Radio Observations of Mercury, Venus, Mars, Saturn and Uranus", *Radio Sci.* 69D:1565-1573, 1965.
Basharinov, A. E., Vetukhnovskaya, Y. N., Kuzmin, A. D., Kutuza, B. G., and Salomonovich, A. E., "Measurements of the Brightness Temperature of Venus at 8 mm", *Astron. Zh.* 41:707-710, 1964.
Berge, G. L., and Grisson, E. W., "High Resolution Interferometry of Venus at 3.12 cm Wavelength", *Astrophys. J.* 156:1125-1134, 1969.
Clark, B. G., and Spencer, C. L., "Some Decimeter Observations of Venus During the 1962 Conjunction", *Astron. J.* 69:59, 1964.
Clark, B. G., and Kuzmin, A. D., "The Measurement of the Polarization and Brightness Distribution of Venus at 10.6 cm Wavelength", *Astrophys. J.* 142:23, 1965.
Copeland, J., and Tyler, W. C., "Preliminary Results from Measurements of 8.6 mm Radiation From Venus", *Astrophys. J.* 139:409-412, 1964.
Davies, R. D., and Williams, D., "Observations of the Continuum Emission From Venus, Mars, Jupiter and Saturn at 21.2 cm Wavelength", *Planet. Space Sci.* 14:15, 1966.
Dickel, J. R., "6 cm Observations and the Microwave Spectrum of Venus", *Icarus* 6:417-426, 1967.
Drake, F. D., "10 cm Observations of Venus Near Superior Conjunction", *Nature* 195:894, 1962.
Drake, F. D., "A Search for the 1.36 cm Water Vapor Line on Venus", *J. Res. Nat. Bur. Stand.* 69D:1577, 1965.
Epstein, E. E., "3.3 mm Observations of Venus in 1964", *Astron. J.* 70:721-725, 1965.
Gibson, J. E., "The Brightness Temperature of Venus at 8.6 mm", *Astrophys. J.* 137:611-619, 1963.
Gibson, J. E., and Corbett, H. H., "Radiation of Venus at the 13.5 mm Water Vapor Line," *J. Res. Nat. Bur. Stand.* 69D:1577, 1965.
Grant, C. R., Corbett, H. H., and Gibson, J. E., "Measurements of the 4.3 mm Radiation of Venus", *Astrophys. J.* 137:620-627, 1963.

Griffith, T. H., Thornton, D. D., and Welch, W. J., "The Microwave Spectrum of Venus in the Frequency Range 18-36 gc/s", *Icarus* 6:175-188, 1967.

Hardebeck, H. E., "Mars and Venus at 70 cm Wavelengths", *Radio Sci.* 69D:1573, 1965.

Hughes, M. P., "Planetary Observations at the Wavelength of 6 cm", *Planet. Space Sci.* 14:1017, 1966.

Jones, D. E., "The Microwave Temperature of Venus", *Planet. Space Sci.* 5:166-167, 1961.

Kisliakov, A. G., Kuzmin, A. D., and Salomonovich, A. E., "4 mm Radio Emission From Venus", *Sov. Astron. A. J.* 6:328-332 (1962).

Kuzmin, A. D., and Salomonovich, A. E., "Observations of Radio Emission From Venus and Jupiter at 8 mm Wavelength", *Sov. Astron. A. J.* 6:518-524, 1963.

Lilley, A. E., "The Temperature of Venus" (Abstract), *Astron. J.* 66:290, 1961.

Mayer, C. H., McCullough, T. P., and Sloanaker, R. M., "Observations of Venus at 3.15 cm Wavelength", *Astrophys. J.* 127:1-10, 1958.

Mayer, C. H., McCullough, T. P., and Sloanaker, R. M., "Observations of Venus at 10.2 cm Wavelength", *Astron. J.* 65:349-350, 1960.

McCullough, D. P., and Boland, J. W., "Observations of Venus at 2.70 cm", *Astron. J.* 69:68, 1964.

Muhleman, D. O., "Interferometric Investigations of the Atmosphere Venus", *Radio Sci.* 5:355-362, 1970.

Roberts, J. A., "Radio Emission From the Planets", *Planet. Space Sci.* 11:221-259, 1963.

Staelin, D. H., and Barrett, A. H., "Radio Measurements of Venus Near 1 cm Wavelength", *Astron. J.* 70:330, 1965.

Staelin, D. H., and Barrett, A. H., "Spectral Observations of Venus Near the 1 cm Wavelength", *Astrophys. J.* 144:352-363, 1966.

Thornton, D. D., and Welch, W. J., "Radio Emission from Venus at 8.35 mm", *Astron. J.* 69:71, 1964.

Tolbert, C. W., and Straiton, A. W., "35-gc/s, 70-gc/s, and 90-gc/s Cytherean Radiation", *Nature* 204:1242, 1964.

Mars

Giordmaine, J. A., Alsop, L. E., Mayer, C. H., and Townes, C. H., "Observations of Jupiter and Mars at 3 cm Wavelength", *Astron. J.* 64:332, 1959.

Hardebeck, H. E., "Mars and Venus at 70 cm Wavelength", *Radio Sci.* 69D:1573, 1965.

Hobbs, R. W., McCullough, T. P., and Waak, J. A., "Measurements of Mars at 1.55 cm and 0.95 cm Wavelengths", *Icarus* 9:360, 1968.

Kellerman, K. I., "Radio Observations of Mars", *Nature* 206:1034-1035, 1965.

Kellerman, K. I., "The Thermal Radio Emission From Mercury, Venus, Mars, Saturn and Uranus", *Icarus* 5:478-490, 1966.

Mercury

Barrett, A. H., "Passive Radio Observations of Mercury, Venus, Mars, Saturn and Uranus", *Radio Sci.* 69D:1565, 1965.

Epstein, E. E., "Mercury: Anomalous Absence from the 3.4 Millimeter Radio Emission of Variation with Phase", *Science* 151:445, 1966.

Epstein, E. E., Soter, S. L., Oliver, J. P., Schorn, R. A., and Wilson, W. J., "Mercury: Observations of the 3.4 Millimeter Radio Emission", *Science* 157:1550-1552, 1967.

Gary, B., "Mercury's Microwave Phase Effect", *Astrophys. J. (Letters)* 148:L141, 1967.

Golovkov, V. K., and Losovskii, D. Y., "Measurements of the Phase Dependence of the 0.8 cm Radio Emission of Mercury and Some Properties of Its Surface Layer", *Sov. Astron. A. J.* 12:299, 1968.

Kaftan-Kassim, M. A., and Kellerman, K. I., "Measurements of the 1.9 cm Thermal Radio Emission From Mercury", *Nature* 213:272, 1967.

Kellerman, K. I., "11 cm Observations of the Temperature of Mercury", *Nature* 205:1091, 1965.

Kellerman, K. I., "The Thermal Radio Emission From Mercury, Venus, Mars, Saturn and Uranus", *Icarus* 5:478-490, 1966.

Klein, M. J., "The Planet Mercury: Measurements of Variations in the Microwave Disk Temperature", PhD thesis, University of Michigan, Ann Arbor, Michigan, 1968a (submitted to *Astrophys. J.*).

Klein, M. J., "Measurements of the 8 GHz Phase Effect of Mercury During Seven Synodic Periods" (Abstract), *Astron. J.* 73:S102, 1968b.

Kutuza, B. G., Losovski, B. J., and Salomonovich, A. E., "Measurements of the Radio Emission of Mercury at 8 mm Wavelength", *Astron. Tsirk.* 5(327).

Morrison, D., and Sagan, C., "The Microwave Phase Effect of Mercury," *Astrophys. J.* 150:1105-1110, 1967.

Morrison, D., "On the Interpretation of Mercury Observations at Wavelengths of 3.4 and 19 mm", *Astrophys. J.* 152:661, 1968.

Morrison, D., "Thermal Models and Microwave Temperatures of the Planet Mercury", Smithsonian Astrophysical Observatory Special Report No. 292, Cambridge, Massachusetts, 1969.

Salomonovich, A. E., "Measuring Mercurian Radio Emission at 8 mm", *Astron. Circ.* 327, 1965.

Vetukhnovskaya, Ya. N., and Kuzmin, A. D., "A Theory for the Radio Emission for Mercury", translated in *Solar System Res.* 2:55, 1968.

Jupiter and the Major Planets (Not Including Nonthermal Emission)

Barrett, A. H., "Passive Radio Observations of Mercury, Venus, Mars, Saturn and Uranus", *Radio Sci.* 69D:1565, 1965.

Berge, G. L., "The Brightness Distribution of Jupiter's 10 and 21 Centimeter Radio Emission" (Abstract), *Astron. J.* 70:132, 1965.

Davies, R. D., Beard, M., and Cooper, D. F. C., "Observations of Saturn at 11.3 cm", *Phys. Rev. Lett.* 13:325-327, 1964.

Davies, R. D., and Williams, D., "Observations of the Continuum Emission From Venus, Mars, Jupiter and Saturn at 21.2 cm Wavelength", *Planet. Space Sci.* 14:15, 1966.

Drake, F. D., "Microwave Spectrum of Saturn", *Nature* 195:893-894, 1962.

Giordmaine, J. A., Alsop, L. E., Mayer, C. H., and Townes, C. H., "Observations of Jupiter and Mars at 3 cm Wavelengths", *Astron. J.* 64:332, 1959.

Gulkis, S., and Carr, T. D., "Radio Rotation Period of Jupiter," *Science* 154:257-259, 1966.

Kellerman, K. I., and Pauliny-Toth, I. I. K., "Observations of the Radio Emission of Uranus, Neptune and Other Planets at 1.9 cm", *Astrophys. J.* 145:954, 1966.

Mayer, C. H., McCullough, P. P., and Sloanaker, R. M., "Observations of Mars and Jupiter at the Wavelength of 3.15 cm.", *Astrophys. J.* 127:11-16, 1958.

Roberts, J. A., "Radio Emission From the Planets", *Planet. Space Sci.* 11:221-259, 1963.

Smith, A. G., Lebo, G. R., Six, N. F., Jr., and Carr, T. D., "Decameter Wavelength Observations of Jupiter: The Apparitions of 1961 and 1962", *Astrophys. J.* 141:457-477, 1965.

Thornton, D. D., and Welch, W. J., "8.35 mm Radiation From Jupiter", *Icarus* 2:228-232, 1963.

Tolbert, C. W., "Observed Millimeter Wavelength Brightness Temperature of Mars, Jupiter, and Saturn," *Astron. J.* 71:30, 1966.

Warwick, J. W., "Radio Physics of Jupiter", *Space Sci. Rev.* 6:184-191, 1967.

Welch, W. J., Thornton, D. D., and Lohman, R., "Observations of Jupiter, Saturn and Mercury at 1.53 Centimeters", *Astrophys. J.* 146:799, 1966.

Radar Astronomical Observations

Venus

Anderson, J. D., Efron, L., Goldstein, R. M., Melbourne, W. G., O'Handley, D. A., Pease, G. E., and Tausworthe, R. C., "The Radius of Venus as Determined by Planetary Radar and Mariner 5 Radio Tracking Data", *J. Atmos. Sci.* 25:1171-1173, 1968.

Campbell, D. B., and Muhleman, D. O., "Measurements of the Electron Content of the Interplanetary Medium Between Earth and Venus", *J. Geophys. Res., Space Phys.* 74:1138-1143, 1969.

Carpenter, R. L., "Study of Venus by CW Radar", *Astron. J.* 69:2-11, 1964.

Evans, J. V., Brockelman, R. A., Henry, J. C., Hyde, G. M., Kraft, L. G., Reid, W. A., and Smith, W. B., "Radio Echo Observations of Venus and Mercury at 23 cm Wavelength", *Astron. J.* 70:486-501, 1965.

Evans, J. V., Ingalls, R. P., Rainville, L. P., and Silva, R. R., "Radar Observations of Venus at 3.8 cm Wavelengths", *Astron. J.* 71:902-915, 1965.

Evans, J. V., and Ingalls, R. P., "Absorption of Radar Signals by the Atmosphere of Venus", *J. Atmos. Sci.* 25:555-559, 1968.

Evans, J. V., and Cramer, G. N., "Radio Echo Observations of Venus", *Nature* 184:1358-1359, 1969.

Goldstein, R. M., "Venus Characteristics by Earth Based Radar", *Astron. J.* 69:12-18, 1964.

Goldstein, R. M., "Radar Time-of-Flight Measurements to Venus", *Astron. J.* 73:829, 1968.

Goldstein, R. M., "A Radar View of the Surface of Venus", *Proc. Amer. Phil. Soc.* 113:223-228, 1969.

Goldstein, R. M., "Radio and Radar Studies of Venus and Mercury", *Radio Sci.* 5:391-395, 1970.

James, J. C., and Ingalls, R. P., "Radar Observations of Venus at 38 mc/s", *Astron. J.* 69:19-22, 1964.

Karp, D., Morrow, W. E., Jr., and Smith, W. B., "Radar Observations of Venus at 3.6 cm", *Icarus* 3:473-475, 1964.

Klemperer, W. K., Ochs, G. R., and Bowles, K. L., "Radar Echoes From Venus at 50 mc/s", *Astron. J.* 69:22-28, 1964.

Kotelnikov, V. A., "Radar Observations of Venus in the Soviet Union in 1962", *Dokl. Akad. Nauk.* 145:1035, 1962.

Kotelnikov, V. A., "Radar Observations of Venus in the Soviet Union in 1964", *J. Res. Nat. Bur. Stand.* 69D:1634, 1965.

Levy, G. S., and Schuster, D., "Further Venus Radar Depolarization Experiments", *Astron. J.* 69:29-33, 1964.

Maron, I. G., Luchak, G., and Blitzstein, W., "Radar Observations of Venus", *Science* 138:1419-1421, 1961.

Melbourne, W. G., Muhleman, D. O., and O'Handley, D. A., "Radar Determination of the Radius of Venus", *Science* 160:987-989, 1968.

Muhleman, D. O., "Early Results of the 1961 JPL Venus Radar Experiment", *Astron. J.* 66:292, 1961.

Muhleman, D. O., Holdridge, D. B., and Block, N., "The Astronomical Unit Determined by Radar Reflections from Venus", *Astron. J.* 67:191-203, 1962.

Muhleman, D. O., "Radar Results as Constraints on Models of Venus", *Astron. J.* 67:277, 1962.

Muhleman, D. O., "The Electrical Characteristics of the Atmosphere and Surface of Venus From Radar Observations", *Icarus* 1:401-411, 1963.

Muhleman, D. O., "Radar Scattering from Venus and the Moon", *Astron. J.* 69:34-41, 1964.

Muhleman, D. O., "Radar Scattering from Venus and Mercury at 12.5 cm", *J. Res. Nat. Bur. Stand.* 69D:1630-1631, 1965.

Muhleman, D. O., "Microwave Opacity of the Venus Atmosphere", *Astron. J.* 74:57-69, 1969.

Pettengill, G. H., and Price, R., "Radar Echoes from Venus and a New Determination of the Solar Parallax", *Planet. Space Sci.* 5:71-74, 1964.

Pettengill, G. H., and Shapiro, I. I., "Radar Astronomy", *Ann. Rev. Astron. Astrophys.* 3:377-410, 1965.

Pettengill, G. H., Dyce, R. B., and Campbell, D. B., "Radar Measurements at 70 cm of Venus and Mercury", *Astron. J.* 72:330-337, 1967.

Ponsonby, J. E. B., Thomson, J. H., and Imrie, K. S., "Radar Observations of Venus and a Determination of the Astronomical Unit", *Mon. Notic. Roy. Astron. Soc.* 128:1-17, 1964.

Price, R., Green, P. E., Goblick, T. J., Kingston, R. H., Kraft, L. G., Pettengill, G. H., Silver, R., and Smith, W. B., "Radar Echoes from Venus," *Science* 129:751-753, 1959.

Rzhiga, O. N., "Results of a Radar Study of the Planets", *Cosmic Res.* 7:76-82, 1969.

Schuster, D., and Levy, G. S., "Faraday Rotation of Venus Radar Echoes", *Astron. J.* 69:42-48, 1964.

Shapiro, I. I., "Resonance Rotation of Venus", *Science* 157:423-425, 1967.

Shapiro, I. I., "Theory of the Radar Determination of Planetary Rotations", *Astron. J.* 72:1309-1323, 1967.

Slade, M. A., and Shapiro, I. I., "Interpretation of Radar and Radio Observations of Venus", *J. Geophys. Res.* 75:3301-3317, 1970.

Smith, W. B., "Radar Observations of Venus, 1961 and 1959", *Astron. J.* 69:42-48, 1964.

Wood, A. T., Watson, R. B., and Pollack, J. B., "Venus: Estimates of the Surface Temperature and Pressure from Radio and Radar Measurements", *Science* 162:114-116, 1968.

Mars

Dyce, R. B., "Recent Arecibo Observations of Mars and Jupiter", *J. Res. Nat. Bur. Stand.* 69D:1628-1629, 1965.

Kotelnikov, V. A., et al., "Radar Detection of the Planet Mars in the Soviet Union", *Dokl. Akad. Nauk.* 151:811-814, 1963.

Muhleman, D. O., Goldstein, R., and Carpenter, R., "A Review of Radar Astronomy", *IEEE Spectrum*, 278:83-89, 1965.

Mercury

Carpenter, R. L., and Goldstein, R. M., "Radar Observations of Mercury", *Science* 142:381-382, 1963.

Dyce, R. B., Pettengill, G. H., and Shapiro, I. I., "Radar Determination of the Rotations of Venus and Mercury", *Astron. J.* 72:351-359, 1967.

Evans, J. V., Brockelman, R. A., Henry, J. C., Hyde, G. M., Kraft, L. G., Reid, W. A., and Smith, W. B., "Radio Echo Observations of Venus and Mercury at 23 cm Wavelength", *Astron. J.* 70:486-501, 1965.

Kotelnikov, V. A., et al., "Radar Detection of the Planet Mercury", *Dokl. Akad. Nauk.* 147:1320-1323, 1962.

Muhleman, D. O., "Radar Scattering From Venus and Mercury at 12.5 cm", *J. Res. Nat. Bur. Stand.* 69D:1630-1631, 1965.

Spacecraft and Occultation Measurements (Not Including Celestial Mechanics)

Cain, D. L., Kliore, A. J., and Levy, G. S., "The Mariner IV Occultation Experiment–Summary of Data and Reduction Methods", *AIAA Preprint 66-148*, 1966.

Cain, D. L., Drake, F. D., Eshleman, V. R., Fjeldbo, G., Kliore, A., and Levy, G. S., "Radio Propagation Measurements of the Atmosphere and Ionosphere of Mars", *Proceedings of the AGARD/IRC Meeting on Propagation Factors in Space*, Rome, Italy, September 25, 1967.

Eshleman, V. R., Fjeldbo, G., Hudson, J. D., Kliore, A., and Dyce, R. B., "Venus: Lower Atmosphere Not Measured", *Science* 162:661-662, 1968.

Fjeldbo, G., Eshleman, V. R., Garriott, O. K., and Smith, F. L., III, "The Two Frequency Bi-Static Radar Occultation Method for the Study of Planetary Ionospheres", *J. Geophys. Res.* 70:3701-3710, 1965.

Fjeldbo, G., Eshleman, V. R., Kliore, A., Cain, D. L., Levy, G. S., and Drake, F. D., "Preliminary Results of the Mariner IV Radio Occultation Measurements of the Upper Atmosphere of Mars", *Proceedings of the Cal-Tech/JPL Lunar and Planetary Conference,* 1966, pp. 267-272.

Kliore, A., and Tito, D. A., "Radio Occultation Investigations of the Atmosphere of Mars", *J. Spacecr. Rockets* 4:578-582, 1967.

Kliore, A., Tito, D. A., and Cain, D. L., "A Radio Occultation Experiment to Probe the Atmosphere of Venus", *J. Spacecr. Rockets* 4:1339-1346, 1967.

Kliore, A., and Cain, D. L., "Mariner V and the Radius of Venus", *J. Atmos. Sci.* 25:549-554, 1968.

Kliore, A., Fjeldbo, G., Seidel, B. L., and Rasool, S. I., "Mariner VI and VII: Radio Occultation Measurements of the Atmosphere of Mars", *Science* 166:1393-1397, 1969.

Kliore, A., Fjeldbo, G., and Seidel, B. L., "First Results of the Mariner VI Radio Occultation Measurements of the Lower Atmosphere of Mars", *Radio Sci.* 5:373-379, 1970.

Page intentionally left blank

CHAPTER 10
NATURE AND INTERPRETATION OF THE APOLLO 11 LUNAR SAMPLES

J. A. Wood
Smithsonian Astrophysical Observatory
Cambridge, Massachusetts

New information has recently become available from the analyses of Apollo 11 samples. It is the most important data on the Moon that we have ever acquired during the long history of observations. We will first discuss the available information about the Moon, then present the properties of the lunar samples acquired by Apollo 11, and, finally, attempt to interpret these data in terms of the processes that could have shaped the interior of the Moon.

I. CRATERING

Cratering has been overwhelmingly the most important mechanism that has shaped the lunar surface; therefore, an examination of this process is essential to an understanding of the lunar problem.

This section will be organized as follows:

(1) Mechanics of individual cratering events.
(2) Systems of interacting craters, and crater statistics.
(3) Development of the regolith.
(4) Physical effects of cratering events on lunar material.

A. Mechanics of Individual Cratering Events

Most or all of the craters on the Moon were generated by impacts of bodies (meteorites and comets) hitting the surface at high velocities. Formerly, the craters were thought by some to be of volcanic origin, but this point of view has not been defended recently, particularly since the return of the Apollo 11 samples. It is not clear why this has been so, as there was nothing in the Apollo 11 samples that

decisively contradicted volcanic activity. However, the author does not consider it likely that many of the lunar craters are related to volcanism; hence, only impact cratering will be discussed here.

The escape velocity of the Moon is 2.4 km/s; thus, all projectiles impinging on it strike the surface with at least this velocity. Usually, the velocity is a great deal higher. The Moon revolves with the Earth around the Sun in a nearly circular orbit, and the material striking it comes from much farther out in the solar system, either from the asteroid belt (between Mars and Jupiter) or from comets, which have orbits extending even farther out. These meteoroids travel in elliptical (eccentric) orbits, so there is a substantial difference between their velocity vectors and that of the Moon about the Sun. This inherent velocity difference is additive to the 2.4 km/s that the Moon's gravitational field imparts to the projectiles. As a result, the latter strike the lunar surface at some 10 to 15 km/s. This corresponds to a great deal of kinetic energy: more kinetic energy per gram than the chemical energy per gram of nitroglycerine. (This analogy between chemical explosions and meteorite impact is somewhat overworked. People have tended to say that a meteorite buries itself in the surface of the material and then, like nitroglycerine, explodes, but this is inaccurate. There are parallels between chemical or nuclear explosions and hypervelocity impact, but there are also important differences: In a chemical explosion, the ejection of material is caused largely by the vaporization of the explosive and perhaps some target material and a buildup of pressure which blasts the material out. In the case of impact, waves of compression and relaxation in the solid target material pitch the material out.)

Donald Gault and his co-workers at the NASA Ames Research Center have done a great deal of experimental work on the cratering process, though necessarily on a small scale. Gault has a gas gun capable of firing projectiles in the kilometer-per-second velocity range, simulating impact velocities on the lunar surface. The gun can be elevated to fire projectiles at various angles into the target material. The target material is placed in an evacuated chamber (pressure about 500 μm), and the projectile penetrates a thin membrane as it enters the chamber, but the impact occurs before air can rush in and affect the event.

The experimental projectiles are made of a variety of materials: glass, steel, plastic, rock, and so forth. The targets also consist of a variety of materials, but often of loose sand. Each cratering event is photographed with high frame-rate cameras (10 000 frames/s). After a crater is made in sand, it is soaked with thermosetting plastic and preserved for sectioning and study. The experimenters sometimes put layers of different-colored sand in the target to see after the collision

how each layer is sheared and deformed by the impact. An exhaustive series of such experiments has been carried out, testing all variables. Solid rock thought to be a realistic imitation of the lunar surface material has been included among the targets.

Gault, Quaide, and Oberbeck (1968) perceived the following stages in the hypervelocity cratering process:

(1) Compression:
  (a) initial contact
  (b) jetting
  (c) terminal engulfment
(2) Excavation:
  (a) radial expansion
  (b) lateral flow
  (c) ejection
(3) Subsequent modification.

1. Compression Stage

When the projectile makes contact with the target, a shock wave starts to propagate down through the target, and another starts to propagate up through the projectile (Figure 1). There is a bounded area (shaded in the diagram) between the two shock waves that is subjected to extreme pressures, of the order of megabars. In this region, the strength of the material is completely overriden, and the material behaves as a fluid. Before the projectile has proceeded very far, this lens-shaped region gives way and starts spurting material out, a process that Gault refers to as "jetting". It is a completely hydrodynamic process due to the rapid buildup of pressure and is not yet governed by shock-wave behavior. Material is initially jetted out at a low angle; as the projectile is buried deeper, the angle increases. The material comes out in the vapor state or the liquid state; it is very hot and at least in the initial stage has an extremely high velocity, possibly even three or four times the projectile velocity. So, if the projectile hit the lunar surface at greater than the escape velocity, some jetted material is almost certain to escape from the Moon. The jetted material is a combination of target and projectile, melted and mixed together. Jetting continues until terminal engulfment, but the velocity of jetted material decreases.

The compression stage is formally ended when the shock wave that was propagated back through the projectile reaches its rear face. At this point, all of the kinetic energy of the particle has been transformed into shock-wave velocity, and the next stage begins.

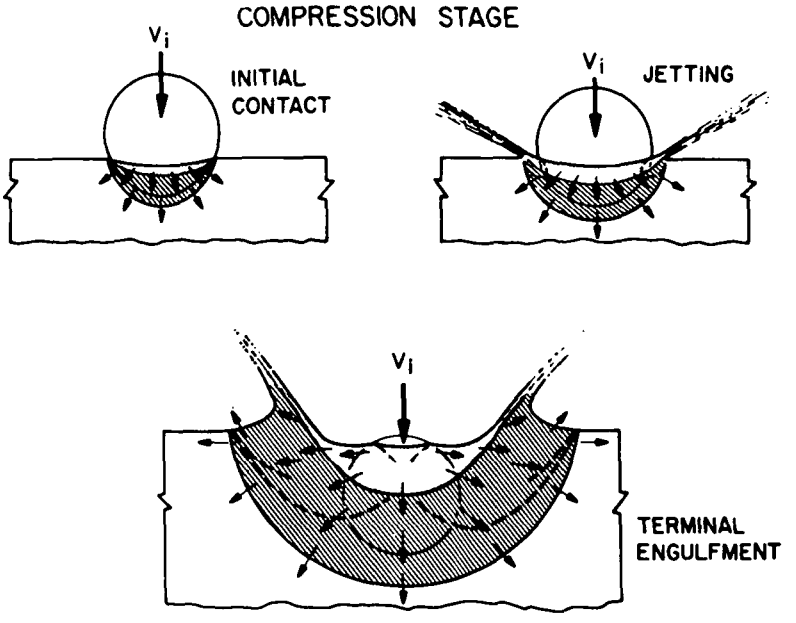


Figure 1.—Events in the compression stage of the formation of a crater by hypervelocity impact (from Gault, Quaide, and Oberbeck, 1968).

2. Excavation Stage

A compression wave propagates from the point of impact downward into the target material; it is followed by a wave of rarefaction. Figure 2(a) shows a series of dashed lines as though there were a series of rarefaction waves, but actually there is only one continuous and gradual decompression in the target material behind the compression wave. The dashed lines should be thought of as isobars shown through the decompressing system. At the target surface, the decompression wave converges onto the compression wave.

The shock wave moves with a certain velocity U_s, and the impact has also imparted a velocity to the material itself; i.e., the wave is moving, and each element of target material is also moving at a different velocity U_p. However, the shock wave is moving faster than the material ($U_s > U_p$). The rarefaction isobars therefore overtake and wash over any particular element of target material that we might decide to keep track of. Now, the rarefaction isobars are not parallel to the compression front, since the two have to converge at the target surface. The element in question was originally moving away perpendicular to the compression front;

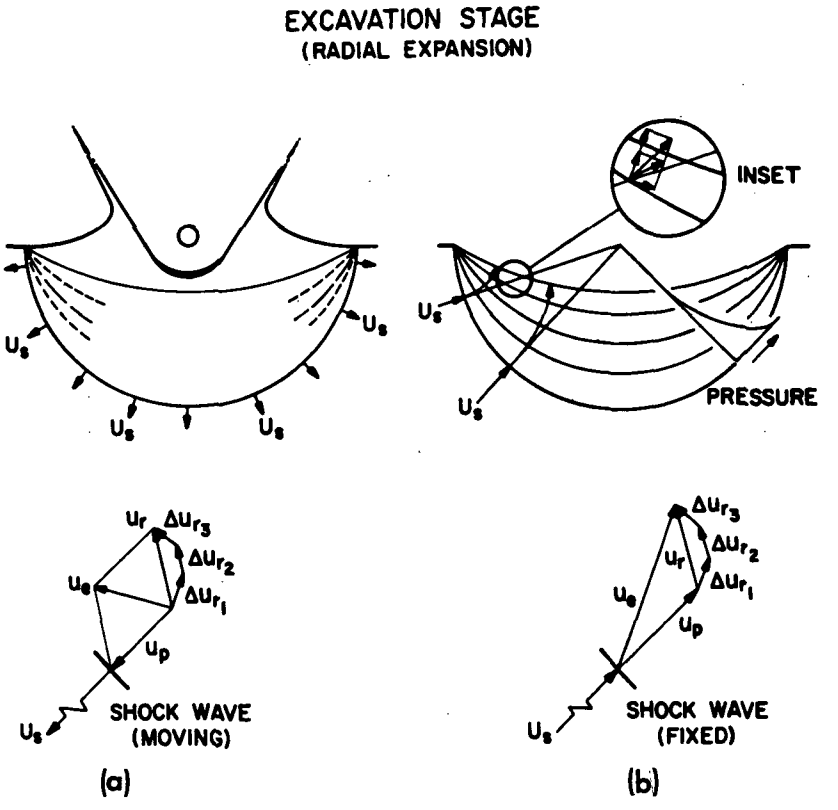


Figure 2.–Radial expansion of material behind a shock wave produced by hypervelocity impact (from Gault, Quaide, and Oberbeck, 1968).

therefore, when these rarefaction isobars overtake it, they wash over it in such a way that the isobars are not perpendicular to the motion of the element. Because of this, the rarefaction has the tendency to deflect the direction of motion of the element. In Figure 2(b), the rarefaction isobars are considered to be a fixed reference frame; thus, the volume-element motion is in the opposite direction (upward) to that indicated in Figure 2(a), where the shock wave is moving. The system of rarefaction isobars tends to accelerate the element backward, which is equivalent to decelerating its motion forward. In such a system, the acceleration can only take place in a direction perpendicular to the isobars. The net effect of the passage of a series of isobars is to deflect the motion of the element in the direction of the

curved arrows in Figure 2(b), i.e., upward and outward from the original point of impact.

Figure 3 develops the idea further and shows the paths that two volume elements would be expected to follow as they are ejected from a crater. As the shock waves expand radially outward, the elements are gradually deflected upward and outward. The volume elements originally close to the target surface are driven in a predominantly horizontal direction, whereas the deeper elements are first accelerated downward, and then deflected toward the surface.

The net effect, as Gault's slow motion films show very graphically, is that during the impact a conical sheet of material is continuously shooting up out of the sides of the hole. As time passes, the crater gets deeper and wider, but the relative dimensions (the shape of the cavity) remain the same. The slope of the crater wall is smooth and continuous, with no ridges or joints.

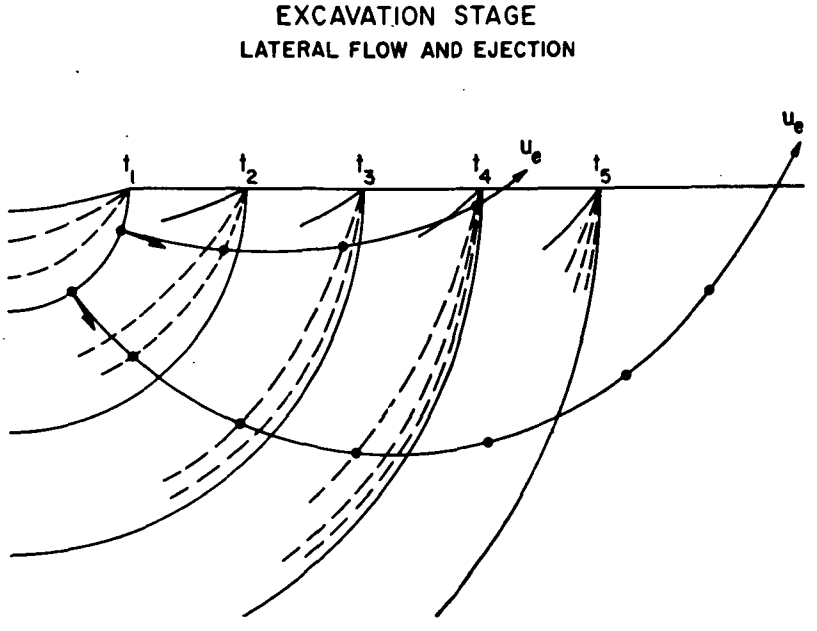


Figure 3.—Expansion of shock compressional wave (concentric solid line segments) and rarefaction isobars (families of dashed lines) away from hypervelocity crater, at successive times t_1, t_2, etc. Passage of rarefaction wave accelerates elements of target material along curved lines U_e, outward and upward from the impact point (from Gault, Quaide, and Oberbeck, 1968).

The velocity at which material is ejected during this stage is relatively small compared with that during the jetting stage. Moreover, the ejection velocity decreases as the process proceeds. Thus, the first material to be ejected is thrown the farthest. The next material is thrown out at a lower velocity; hence, it does not go as far, but there is more of it. The last material to be ejected goes the least distance, but it is most abundant. Therefore, crater ejecta are distributed in such a way that material near the original target surface is thrown the farthest, whereas the material of deepest origin is deposited close to the crater. In fact, the great bulk of the ejecta is deposited within 1 crater diameter of its source.

The choice of Fra Mauro as a landing site for Apollo 13 (and, subsequently, Apollo 14) was based on the hope that the material at this site, obviously ejecta from the vast crater that is now Mare Imbrium, might have been derived from deep in this crater and therefore would provide samples of the lunar interior. But as has been seen, this will not be the case: The deepest material from the Imbrium crater will be found on the rim of Mare Imbrium, and not as far away as Fra Mauro.

The last material to be ejected comes down in such abundance that it behaves like a fluid for a short period of time. This phase of movement leaves a characteristic pattern on the ground, which is said to be due to the fluid flow of material when it comes down and is fanning out. Some writers refer to this as the base-surge mechanism, after the expanding cloud that one sees in films of deep-seated nuclear weapon tests. H. Masursky of the U.S. Geological Survey, in particular, has studied this phenomenon as it was manifested during the Sedan underground nuclear explosion. The ground around the crater was found to have a hummocky appearance. Masursky finds similar patterns on the Moon, and he attributes them to the base-surge mechanism. Gault, however, objects to analogies drawn to the Sedan test which, because of its depth, involved ejection mechanisms quite dissimilar to those of the impact cratering process. Figure 4 shows the fan of hummocky base-surge ejecta surrounding the lunar crater Copernicus.

Figure 5 is a cutaway diagram of a crater that Gault and his co-workers made in sand, with markers embedded for recording the movement of each point. Gault has drawn vectors to indicate the direction of movement of each particle as a consequence of the cratering event. Note that there is a general tendency for the material to fan radially outward, as discussed earlier in this section.

3. Modification Stage

Figure 6 suggests the ways in which a crater can be modified after it has been formed. One thing that can happen is that the walls of the crater can slump into it,

as a result of the weakened state of the wall rock. Very often lunar photographs show crater walls that have slumped; e.g., Copernicus (Figure 4). Another process is isostatic adjustment. A large crater is out of hydrostatic equilibrium with its surroundings: The weight of a column outside the crater is greater than the weight of a column inside it. If the material in the Moon has any tendency to creep and flow over a period of time (and rock does behave this way), then to some extent it will flow and readjust by squeezing up to fill the center of the crater. A final method of modification is by erosion and filling in: Material flows in some way down into

Figure 4.–The lunar crater Copernicus (lower left), successively surrounded by hummocky base-surge deposits and then arrays of secondary craters. Diameter of Copernicus, 90 km (Lunar Orbiter photograph LOIV-121H2).

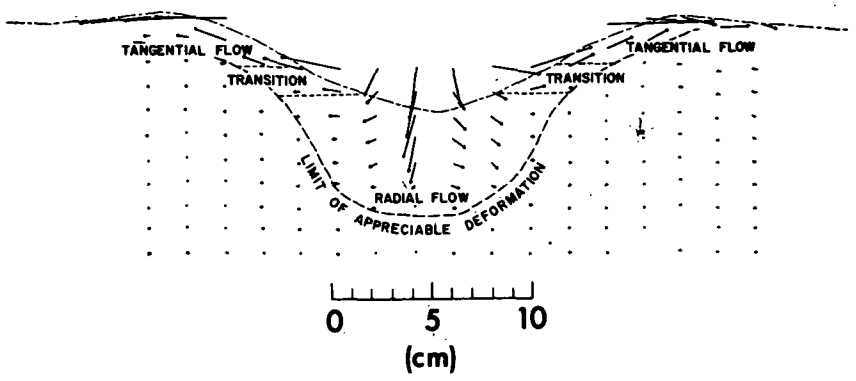


Figure 5.–Movement of individual point masses in a loose sand target during experimental cratering impact (from Gault, Quaide, and Oberbeck, 1968).

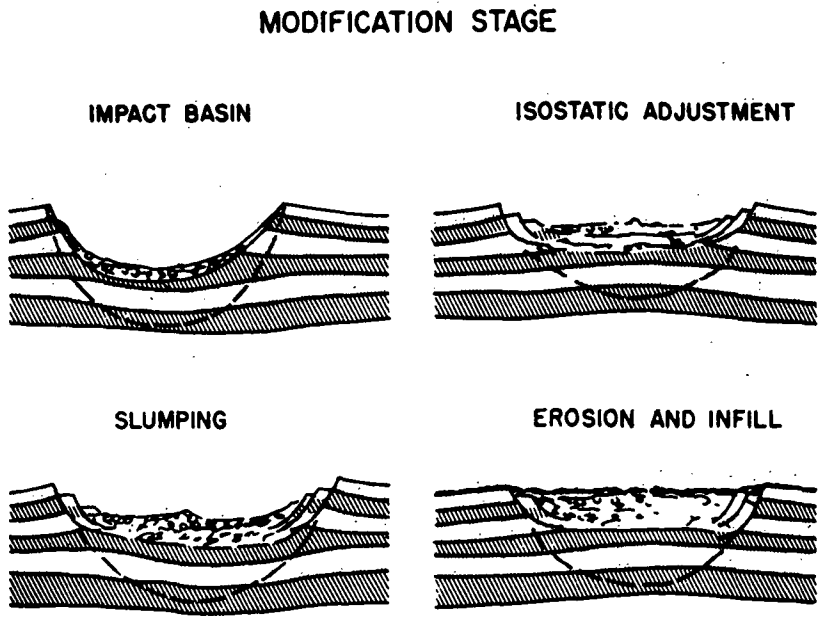


Figure 6.–Processes that would tend to fill and otherwise modify impact craters once they are formed (from Gault, Quaide, and Oberbeck, 1968).

the crater and fills it up. Formerly, it was fashionable to suggest that activity of water on the moon could be such a filling-in mechanism. Or, as Tom Gold has suggested, dust could behave as a fluid if the dust particles became electrostatically charged to such an extent as to repel one another strongly. Such fluidized dust is supposed to flow downward and fill the low spots on the Moon, i.e., the craters. One more possible mechanism of modification, not shown in Figure 6, is igneous activity: Lava could erupt onto the surface and fill craters. Some craters do appear to be filled with lava (e.g., Copernicus).

Craters on the Moon have been studied for a number of years, and a fairly universal relation between their depths and diameters seems to have emerged (Figure 7). The slope of the relation in Figure 7 is almost 45 deg, so that the ratio of depth to radius is almost constant, but not quite. However, above a certain point (about 10-km radius), the curve departs from linearity. Craters larger than this are not as deep as they ought to be according to their diameters. It seems very likely that these craters were originally deeper, but that secondary processes have filled them. Apparently, craters larger than a certain size cannot maintain a deep floor.

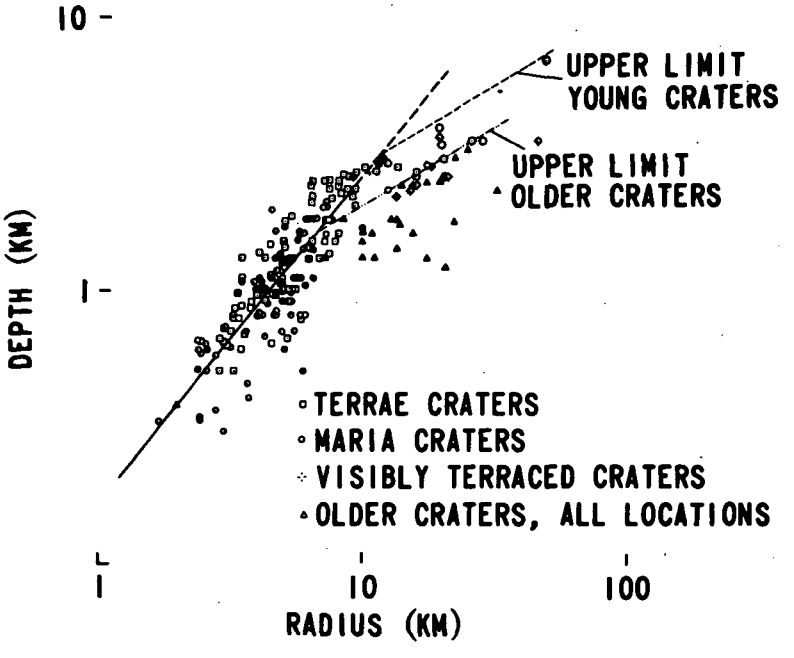


Figure 7.—Depth versus radius for a number of lunar craters (from Quaide, Gault, and Schmidt, 1965).

The effect of low impact angle on crater morphology has been studied. Gault's experiments show that the crater remains circular in form even when angles of incidence are as low as 15 deg, but at low impact angles the ejecta blanket becomes asymmetric.

If Gault's experimental results are to be extrapolated to the case of the large craters on the lunar surface, we need to know the relationship between the dimensions of the crater, the energy of the impact, and its scaling law. Gault (1964) cites experiments by other investigators on craters in water. The formation energy E_f of a crater in water is approximately

$$E_f \approx \frac{\pi}{4} (8Tp^2 + \rho g p^4),$$

where T is the surface tension of the water and p is the penetration depth of the crater. The first term in brackets describes the physical strength of the water; the second is a gravitational term referring to the energy necessary to take the water out of the crater and move it elsewhere. For very small craters, the surface tension, or strength, term dominates. For very large craters, the gravitational term dominates. Therefore, if the gravitational term predominates, the depth of penetration should be proportional to the fourth root of E, i.e., $p \propto E^{1/4}$. If the strength term dominates, $p \propto E^{1/2}$. Thus, there are two different scaling laws for two different ranges of crater dimension. The depth of penetration is related to the mass ejected, so we should find a relation as sketched in Figure 8. That is, the curve relating these two parameters should start out with square-root scaling, because the mass is small, and the surface tension term predominates. At some critical dimension, it would change to fourth-root scaling. Experimental work tends to confirm this, as is shown in Figure 9. The dashed curve refers to data for water. One could wish there were more of a spread in the data. There seems to be a changeover from square-root to fourth-root scaling, but there are no points for larger crater dimensions to confirm this. Under the curves for water, there is a large body of data for impacts on solid materials, including Gault's data and weapons tests. Again, there is some indication that the curve changes over from square-root to fourth-root scaling. Of course, more than surface tension is involved in bonding a solid together, and cratering in solids is a more complex process than cratering in water; yet the data do fit a pattern similar to square-root and fourth-root scaling. This correspondence in behavior is used by Gault to justify extrapolating his laboratory studies on craters of centimeter dimensions up to lunar craters that are kilometers in size.

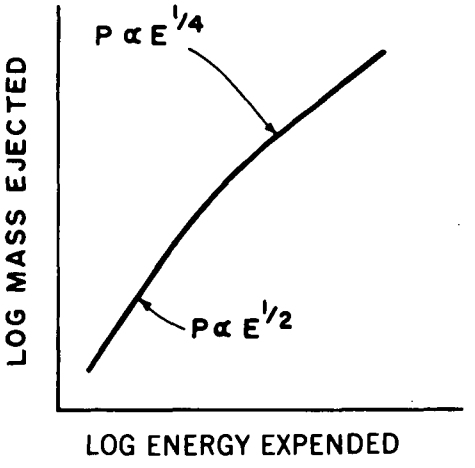


Figure 8.—Relationship of energy expended in crater formation and dimension of crater (mass ejected). Two different scaling laws obtain for different ranges of impact energy, depending upon whether gravitational or material-strength forces dominate.

B. Systems of Interacting Craters, and Crater Statistics

We have to consider not only the formation of individual craters on the Moon, but also the interaction of one crater with another, and the creation of complex patterns of overlapping craters. If a smooth surface is pelted with objects of a given size, it at first records the number of impacts it suffers. However, a point is reached when the virgin surface is largely used up, and old craters are being obliterated as fast as new ones are formed. We define, somewhat artificially, the term "saturation" as that state of affairs when a surface cannot possibly record any more impacts, because craters are lying next to one another in a closely packed hexagonal array (Figure 10). In this case, the ratio of the cratered area to the total area is 0.905.

In dealing with crater statistics, it is customary to use a cumulative plot, as in Figure 11. The logarithm of the cumulative number of craters per unit area is plotted against the logarithm of the crater diameter. For example, Figure 11 tells us that the surface described contains n craters per unit area, of diameter equal to or greater than d.

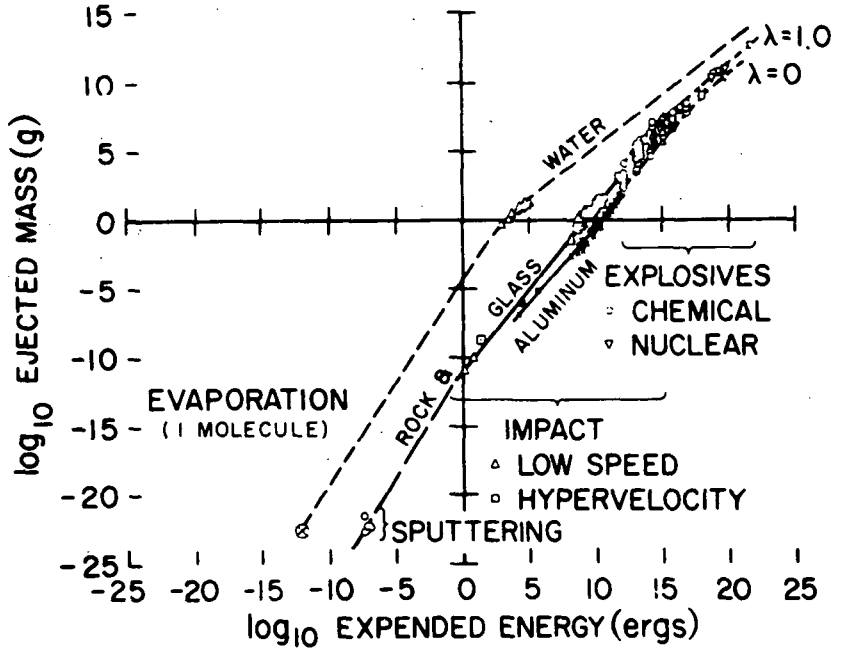


Figure 9.–Relationship between energy expended during a cratering event and amount of mass excavated, for a wide variety of target materials and impact velocities (from Gault, 1964).

For any given crater diameter, there is an easily calculable number of craters per unit area corresponding to saturation, or 10-percent saturation, etc. We can plot families of curves for these various degrees of saturation, as is shown in Figure 12. Starting with a smooth surface, bombardment builds up the number of craters, and at any point in time a cumulative-distribution curve can be drawn that describes the crater population. The curve representing the crater distribution migrates upward (see heavy lines in Figure 12). The crater-distribution curve cannot, of course, cross the curve corresponding to full saturation of the surface. Actually, it does not even approach the saturation curve. Bombarded surfaces reach a state of equilibrium long before saturation (at a value of about 5 to 7 percent of saturation, in fact). As soon as the curve representing cumulative crater distribution (heavy line) reaches the equilibrium curve, it flattens off as in Figure 12. This situation can actually be found on the lunar surface. When craters in a uniform surface area are measured and

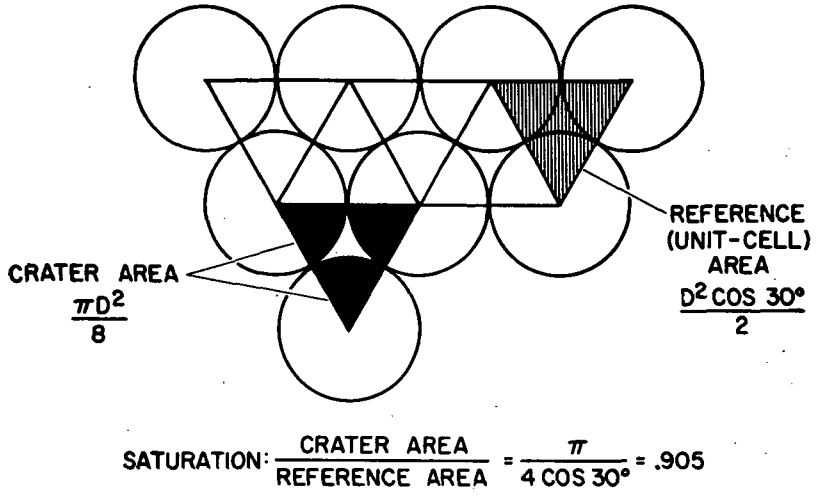


Figure 10.–"Saturation" defined for a crater field: Craters are pressed together side by side in a hexagonal close-packed array. The craters constitute 0.905 of the target area (from Gault, 1970).

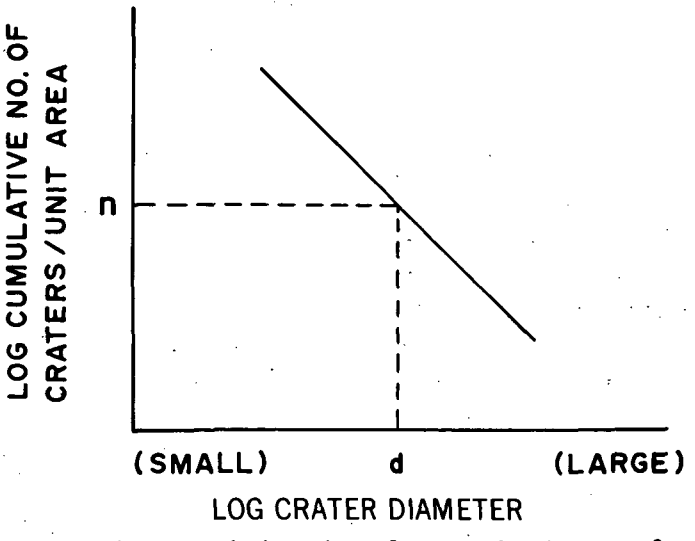


Figure 11.–Example of a cumulative plot of crater density as a function of crater size for an impacted surface.

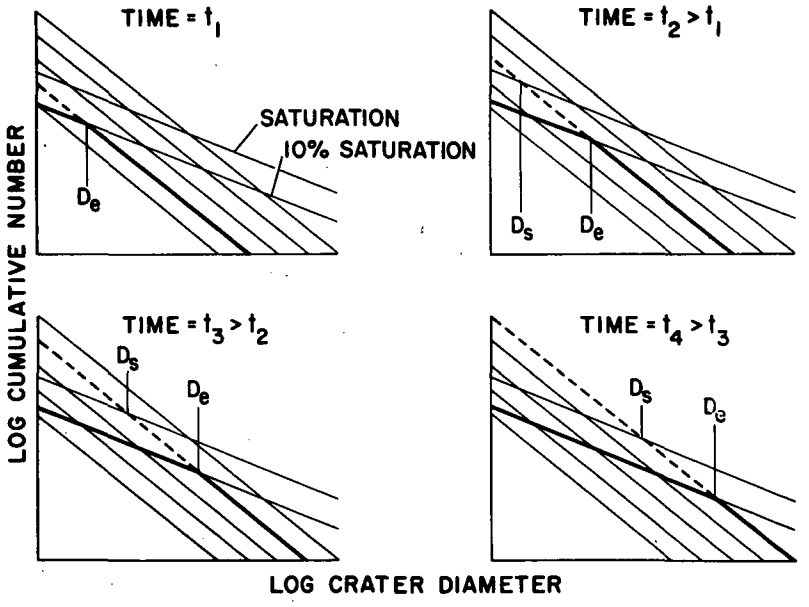


Figure 12.—Cumulative crater density plots at four successive times showing evolution of a cratered surface with time. Crater population curve (heavy line) migrates to the right and up with time but cannot cross shallower curve representing approximately 10-percent crater saturation (from Gault, 1970).

counted, their distribution plots just like the example in Figure 12. If we knew the flux of objects striking the Moon, we could calculate how long it took to reach a given point, and we could then date different areas of surface on the Moon. It is an observational fact that some areas of the Moon are less densely cratered than others; hence, they must be younger.

So that actual ages may be calculated, the flux of impacting objects must be known. Measurements of the fluxes of various-sized objects on the Earth's surface have been made (Figure 13); obviously, a correction must be made before this flux pattern is applied to the Moon, to allow for the smaller lunar gravitational-capture cross section.

When these fluxes are used to date surface areas of the Moon, results like those in Table 1 are obtained. The two values shown for each area represent the extreme

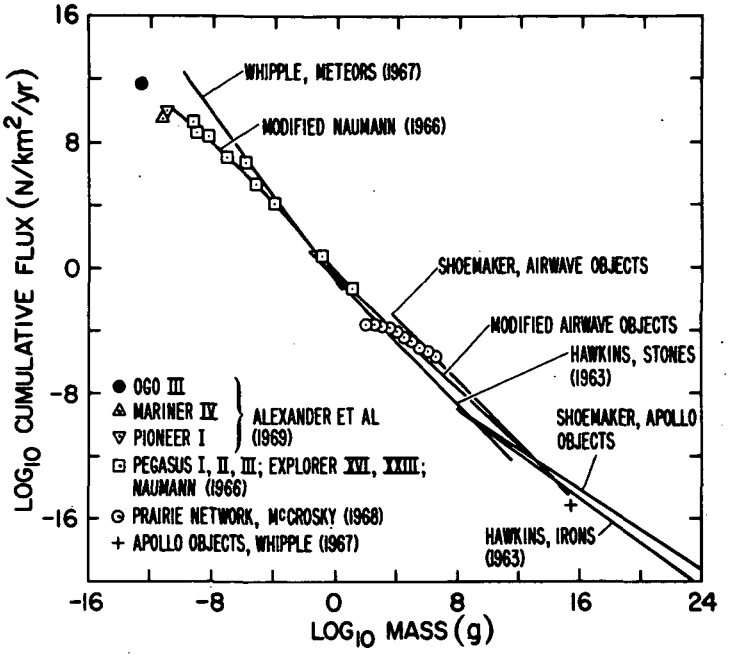


Figure 13.—Influx rate of meteoroids at the Earth as a function of meteoroid mass; compilation of data from a number of different sources (from Gault, 1970).

possible range of particle fluxes. Yet, radiometric dating of Apollo 11 rocks showed that Mare Tranquillitatis is about 3600 million years old, not 20 to 120 million years old as predicted by Gault. Apparently, the flux rate used in calculating these surface ages is much too large. Gault has tried to recalibrate the system by using the now-known age of Mare Tranquillitatis. However, when the Mare Tranquillitatis age is increased to 3600 million years, the age of the Southern Highlands must be increased to some 3×10^{12} years. This indicates that the flux rate cannot have been constant through time, something that some writers have maintained previously for other reasons altogether. These results suggest that early in the history of the Moon, the influx of objects was colossally higher than it is now, and that the flux rate fell off exponentially with time. This is a very crucial aspect of the history of the Moon, affecting more than just the appearance of the lunar surface; this concept will be returned to when the interpretation of the lunar samples is discussed.

Table 1.—Ages of various lunar surfaces, based on crater density and two estimates of the meteoroid flux rate at the Moon (from Gault, 1970).

Site	Age (millions of years)	
	Naumann-Hawkins Flux	Whipple-Shoemaker Flux
Southern Highlands	$>10^5$	$>10^4$
Mare Orientale		
Ejecta blanket	2000	250
Mare-material fill	300 to 400	30 to 60
Highland Plains		
Hipparchus	1700	170
South of Rima Ariadacus	1400	160
Sinus Medii		
Orbiter 2 P7	300	40
Orbiter 3 P7	100	20
Oceanus Procellarum		
Orbiter 3 P11	100	15 to 20
Orbiter 2 P13	200	30
Orbiter 1 P8	50	3 to 6
Mare Tranquillitatis		
Orbiter 2 P6 (Apollo 11 landing site)	120	20 to 30
Copernicus		
Ejecta blanket	30	8
Tycho		
Ejecta blanket	10	2

C. Development of the Regolith

Meteorite impact on the surface of the Moon generates a layer of rocky rubble, termed the regolith. As the bombardment continues, the layer of rubble gets deeper,

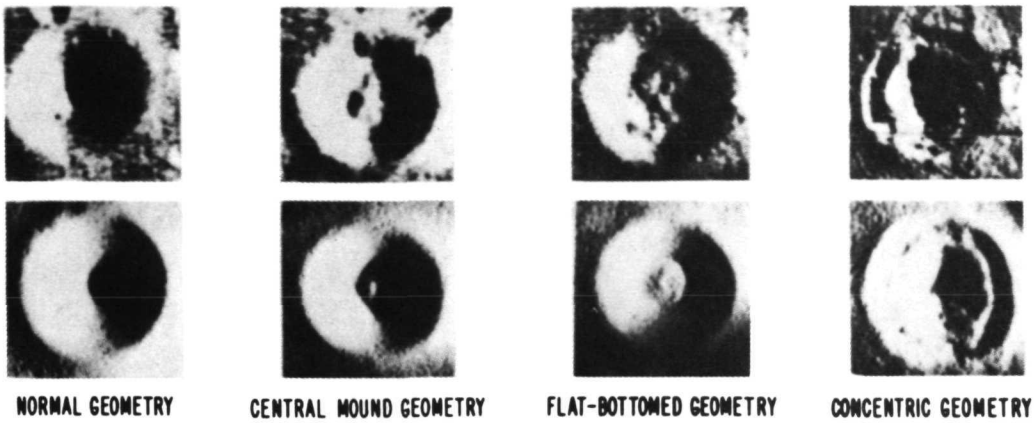


Figure 14.–Four types of impact crater with differing morphology, the difference depending upon the degree to which the crater penetrates an unconsolidated surface layer and impinges upon solid target material beneath. In each pair, a genuine lunar crater appears at the top and a laboratory-produced crater at the bottom (from Quaide and Oberbeck, 1968).

but at a decreasing rate, because once the regolith becomes reasonably thick, new craters made by small meteoroids will be confined to the regolith; they will not penetrate to the bedrock below and furnish additional debris to the regolith. Only the larger impacts can grind more bedrock and increase the mean thickness of the regolith. The more frequent small impacts only churn the upper few centimeters, reducing the material to a fine powder. If one could view a cross section through the regolith, he might find there were more coarse fragments and boulders near the bedrock than near the surface zone.

In order to approximate a lunar surface consisting of hard bedrock overlaid by a loose regolith in their cratering experiments, Gault and his co-workers used cohesive sand (bonded together with plastic) overlaid by a layer of loose sand. Surfaces of this type were bombarded at different velocities and angles of incidence; the layer thickness was varied also. Common sense tells us that a crater made by a small meteorite that was confined to the regolith should look different from one made by a meteorite that was sufficiently large to penetrate to the bedrock. This is exactly what the Ames group found, as can be seen in Figure 14. This figure shows several different crater morphologies obtained by varying the relationship between the penetration depth and the thickness of the layer of loose sand. As Figure 14

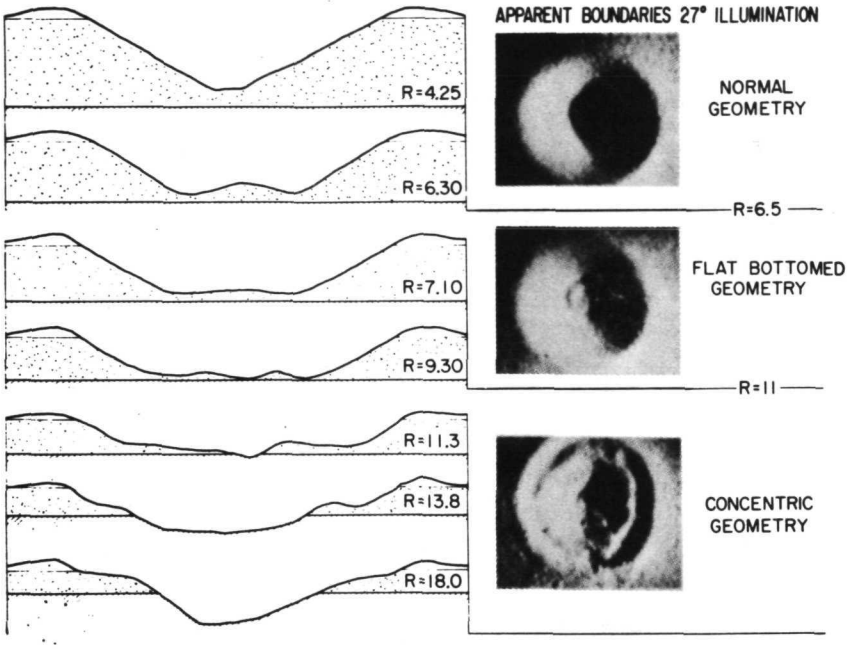


Figure 15.—Experimentally produced crater profiles in situations (from top to bottom) where the "regolith" was successively thinner and crater impingement on underlying solid material was more profound (R is the ratio of crater diameter to unconsolidated layer thickness). Photographs are of experimentally produced craters in three different ranges of R (from Oberbeck and Quaide, 1967).

shows, this same range of crater morphologies can be found on the lunar surface. Figure 15 shows profiles through a series of experimentally produced craters and how these correspond to the geometry as viewed from the top (R is the ratio of the crater diameter to the unconsolidated layer thickness). Craters confined to the regolith are conical in shape, but those which penetrate to the bedrock tend to be flat bottomed. Evidently, craters are conical in shape if $R \leqslant 4$. One might wonder, however, if this criterion is affected by the angle of incidence, or by the strength of the projectile material and target layers. Figure 16 summarizes the results of experiments designed to test the effect of these variables. In the figure, D_F is the diameter of the floor of the crater, D_A is the diameter of the lip of the crater, and t is the loose layer thickness. Bars at the top of the figure denote the gross

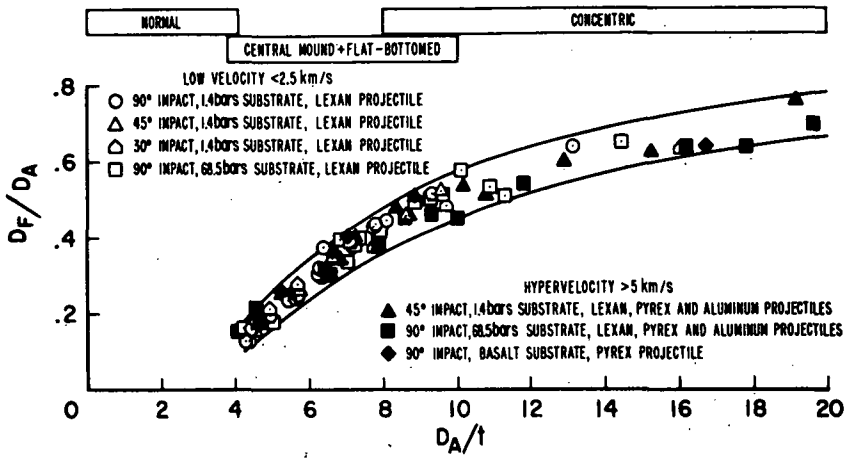


Figure 16.–Relationship of crater morphology (D_F/D_A is the ratio of diameter of floor of crater to diameter of lip of crater) to thickness of unconsolidated or "regolith" layer (t). The wide variety of angles of impact, projectile materials, and strengths of substrate shown do not seem to affect this relationship (from Quaide and Oberbeck, 1968).

morphologies observed for different ranges of values of D_A/t. For almost any conditions, the changeover from craters with normal (conical) geometry to craters having central mounds or flat bottoms comes at $R \approx 4$. Therefore, to determine the mean thickness of the regolith in an area of the Moon, one has merely to survey a number of craters, note at which crater diameter there is a changeover from simple, conical geometry to more complex geometry, and then divide that diameter by 4.

Oberbeck and Quaide (1968) surveyed 12 different sites on the lunar surface (mainly tentative Apollo landing sites) in this way, examining a large number of craters in order to obtain good statistics. They found that the regolith thickness was variable in each region but that four characteristic distributions of thickness could be distinguished, as is shown in Figure 17. This figure shows cumulative distributions and tells us, for example, that for regolith Type I, only 5 percent of the area has a regolith thickness greater than about 7 m. Higher-numbered regolith types are progressively thicker.

The crater density also increases from Type I through Type IV, as we would expect. The age should also increase from Type I to IV, based on the number of craters present. Thus, the highlands, the most cratered lunar terrain, should be the oldest, as they have the thickest regolith (Type IV).

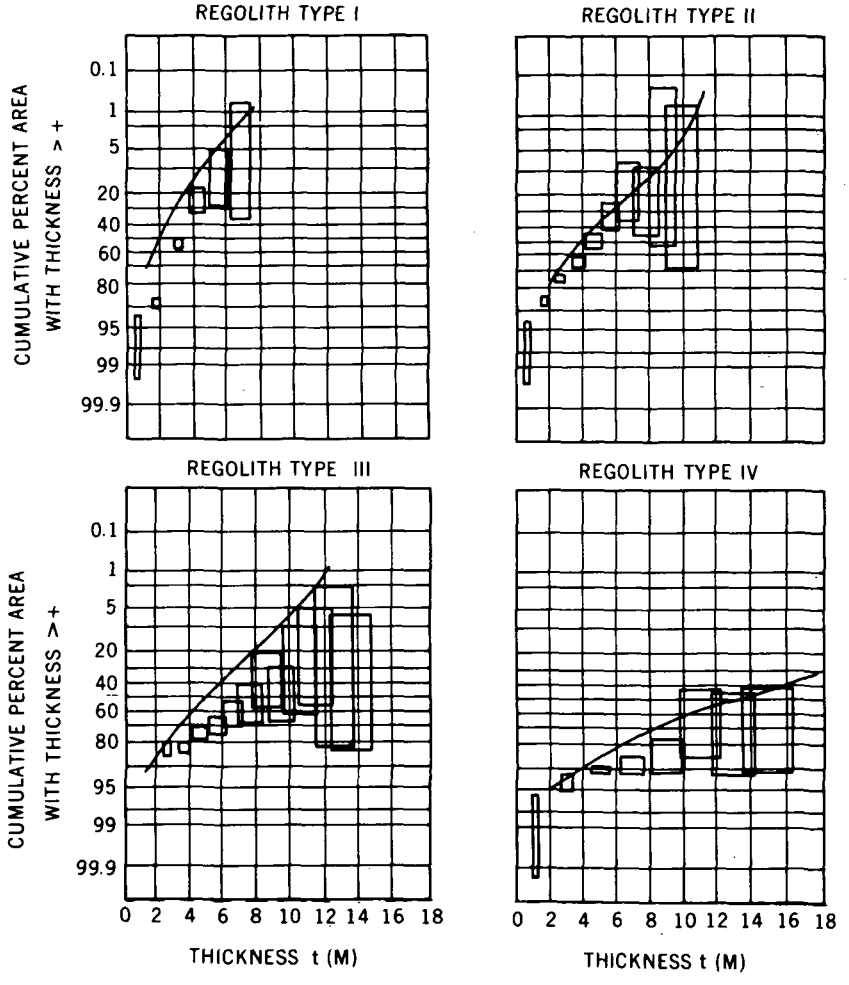


Figure 17.–Cumulative graphs of distribution of regolith thickness for four different regions or regolith types on the Moon, as differentiated by Oberbeck and Quaide (1968).

D. Physical Effect of Cratering Event on Lunar Material

The Ames research group has tried to understand how energy is partitioned during a cratering event. For the case of an aluminum projectile striking hard rock

(basalt) at 6.25 km/s, they found that the energy is partitioned as shown in Table 2. It appears that most of the kinetic energy originally possessed by the projectile is converted into kinetic energy of ejected material.

Figure 18 shows the physical effect of shock waves as they pass through a target away from the point of impact. At the point of impact, the shock affects the target profoundly, but as the shock moves radially outward, it is attenuated and does less and less damage. In this figure, r/D is the ratio of the distance the shock wave has traveled through the target to the diameter of the final crater; m/m_p is the ratio of the target mass that has been affected to the mass of the projectile; and M_e is the total target mass that will ultimately be excavated. For a velocity of 10 km/s, which is realistic for lunar impacts, several projectile masses of target material are actually vaporized; approximately five projectile masses of target material are partly vaporized and partly melted, and so forth.

On the basis of Figure 18, we can sketch semiquantitatively the effect of an impact of a mass M at 15 km/s on the regolith, if the impact is assumed to be confined to the regolith (Figure 19). It is seen that immediately at the impact point, melting and vaporization of the material would occur, giving rise to blobs and droplets of glass. A larger region involving 10 to 1000 projectile masses would be shock lithified, the loose dust being pressed into a coherent mass. In this zone, shock pressure is not great enough to melt the material completely, but it is sufficient to cause incipient melting at the points of contact between mineral grains and rock

Table 2.—Partition of energy on impact of aluminum projectile against basalt at 6.25 km/s (from Gault and Heitowit, 1963).

Form in Which Energy Is Expended	Percentage of Projectile Kinetic Energy
Irreversible heat	
In projectile	4 to 12
In target	19 to 23
Comminution (i.e., breaking up of material)	10 to 24
Kinetic energy of ejecta	43 to 45
Miscellaneous	
Residual elastic wave	>1
Radiant energy	negligible

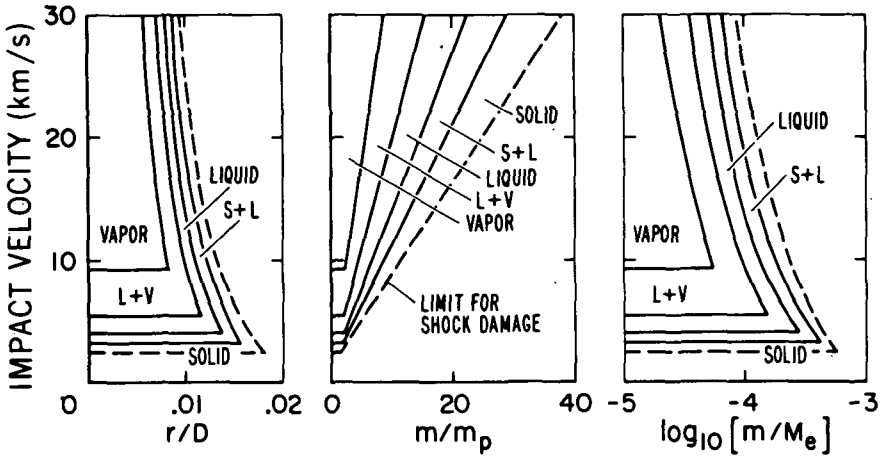


Figure 18.–Nature of damage done to an unconsolidated sand or tuff target by basalt or granite projectiles in the velocity range 0 to 30 km/s, where r is the distance outward from point of impact; D is the diameter of crater that is ultimately produced (1 m, in this case); m is the mass of target material that has been affected by shock wave; m_p is the mass of projectile; and M_e is the mass of target material that will ultimately be excavated from crater (courtesy of D. E. Gault, personal communication).

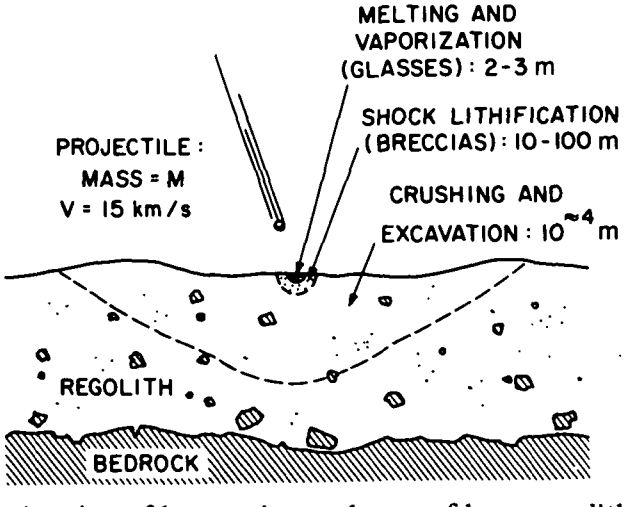


Figure 19.–Approximation of how various volumes of lunar regolith will be affected by a meteorite impact.

fragments in the regolith. This degree of melting is enough to bind the rock together. Beyond this zone, there is a large conical region of some 10^4 M where the regolith is heaved out onto the rim and to some extent is broken finer than it had been before.

II. LUNAR GEOLOGY

About 20 years ago, lunar geologists realized that they could apply well-known aerial geologic mapping techniques to the Moon. A program of lunar mapping has been carried out by the Astrogeology Branch of the United States Geological Survey (USGS) using earth-based telescopic photos and, more recently, the high-resolution pictures obtained by Lunar Orbiter spacecraft. Much of the output of this program has been published as colored geologic maps, but publication tends to be slow, and maps of some crucial areas are still available only as open-file reports.

To distinguish different rock units on the Moon, the following criteria are used:

(1) Albedo (lightness and darkness of the rocks). This varies strongly with Sun angle; strict comparisons can be drawn only at full Moon (zero phase angle).

(2) Structure (landform). In the simplest terms, some lunar terrains are flat and others are hilly; a number of more subtle variations of landform are also distinguishable. These are best differentiated at relatively low Sun angle, when shadows impart a three-dimensional appearance to the lunar terrain.

There are also ways of determining the relative ages of different rock units. For example, it is sometimes clear that a particular body of rock (presumably lava) has flowed into and filled valleys and craters in the preexisting terrain. Obviously, the lava is younger. Again, a fault plane may be seen that offsets topography in one rock unit and not in another, in which case the unfaulted unit would be the younger one. Crater rays are also useful; these may be seen to cross one rock unit but not another, which we conclude was deposited after the ray system. Finally, crater densities can be used to determine relative ages, as was discussed earlier.

Most USGS maps are confusing to novices because the overlapping ejecta blankets from various craters are mapped as different rock units, depending upon how early or late they were excavated; yet many of these ejecta deposits are indistinguishable from one another in chemical or physical properties, having been derived from the same type of bedrock. Figure 20 is a much-simplified geologic map of the Moon, in which an attempt is made to indicate only the fundamentally different types of bedrock present on the Moon. Ejecta from craters of less than mare-basin size are not shown. Three main types of rock units can be distinguished here:

(1) Mare filling—flat and usually dark in color,

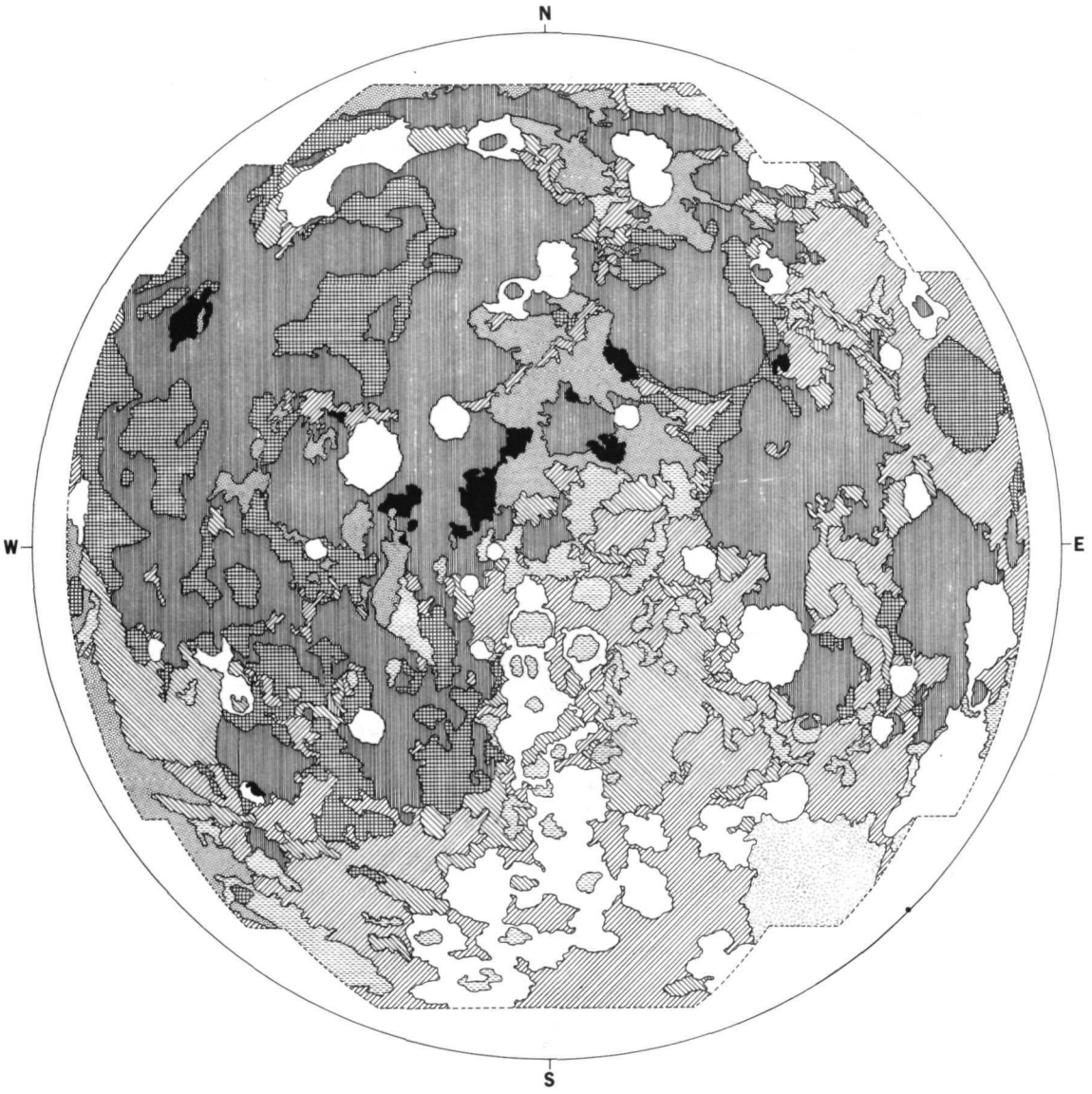
(2) Terra—high, rugged, highly cratered, light-colored highlands or mountains,

(3) Mare ejecta—also high and rugged terrains, but these have obviously been pitched out of the large mare craters.

The USGS has developed a system of relative ages built around the Imbrian event on the Moon; that is, the colossal impact that excavated the basin that is now Mare Imbrium. Ejecta from Mare Imbrium was deposited over much of the visible face of the Moon; it furnishes a useful time marker. The Imbrian period (Table 3)

Table 3.—Summary of the lunar geological column.

Time Period		Nature of Principal Geologic Materials
Older ←	Copernican	Ejecta surrounding rayed craters of relatively young age (e.g., Kepler, Copernicus, Tycho, Aristarchus, and Theophilus).
	Eratosthenian	Crater ejecta of intermediate age (e.g., surrounding craters Eratosthenes, Plinius, and Bullialdus). Last of the dark lavas that were deposited in maria and Oceanus Procellarum (Apollo 12 basalts are a sample).
	Imbrian	Most dark lavas that filled mare basins and Oceanus Procellarum (Apollo 11 basalts are a sample). Debris impact-ejected from mare basins and deposited concentrically about them. Lighter-colored plains-forming material (including Cayley formation), older than dark lavas.
	Pre-Imbrian	Ancient crust of the Moon; possibly very complex geologically, but differences in ancient rock types largely unresolved by USGS mappers.


N
W
E
S

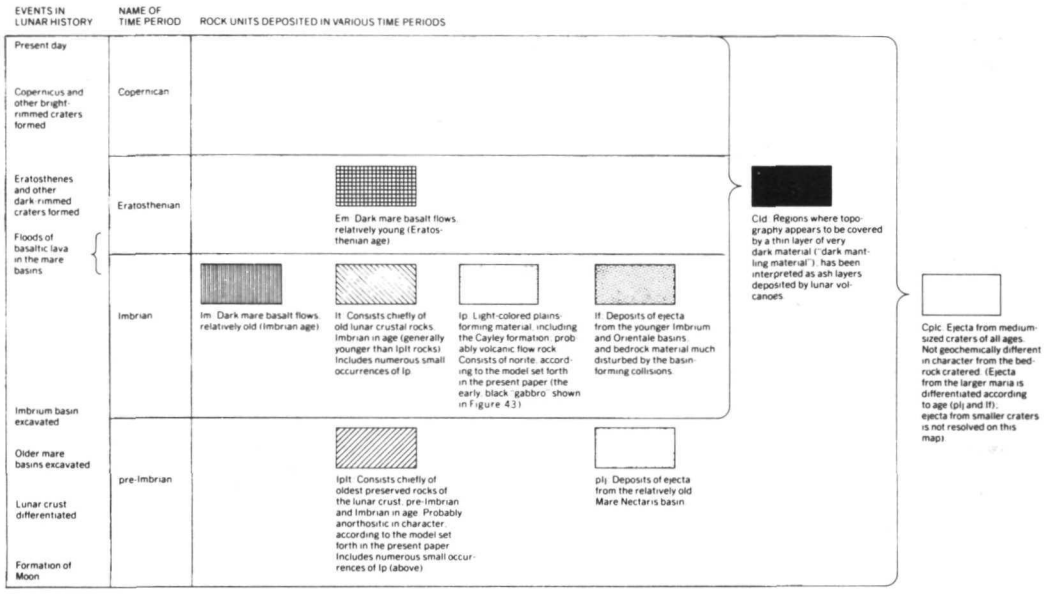


Figure 20.–Geologic map (facing page) of the near side of the Moon, much simplified. Within areas marked by a particular type of pattern, the lunar surface consists largely of one distinctive type of terrain, which is set forth in the key above. Eratosthenian and Copernican crater ejecta blankets, which actually cover a substantial proportion of the Moon, are omitted from this map so that more fundamental differences in lunar geologic units can be observed. This map is a simplified version of the 1:5,000,000 compilation of lunar geology by the United States Geological Survey (Wilhelms, D. E., and McCauley, J. F., USGS Misc. Geol. Inv. Map I 703, 1971, in press).

began with the impact that formed the crater and ended when the crater was filled with lava. Everything before the Imbrian period is lumped together as "pre-Imbrian". The pre-Imbrian terrains are very complex in structure, but are also much obscured by later geology; pre-Imbrian lunar geology has not been worked out in much detail. If anything can be said to represent the original crust of the Moon, it would be found in the region mapped as pre-Imbrian.

There are extensive deposits of material that was obviously thrown out of Mare Imbrium; this unit is named the Fra Mauro formation after the region around the crater of that name, where Imbrian ejecta occur in abundance. Another important rock unit of Imbrian age is a light-colored, plains-forming material. Its flatness suggests that it originated as a lava flow; but it is lighter in color than the dark rock filling the maria, which implies a different chemical composition. This unit, locally called the Cayley formation, has a greater crater density than the darker-colored mare-plains areas surrounding it and is therefore older.

The youngest major unit of lunar rocks is the Procellarum group of lavas. These are the dark rocks that fill the mare basins. They were not erupted simultaneously but over a substantial span of time. The Eratosthenian rock units shown in Figure 20 are lavas that flowed into the maria last of all; they are slightly darker than the main body of Procellarum lavas.

III. PROPERTIES OF THE APOLLO SAMPLES

The material on the lunar surface is highly variable in particle size, ranging from rocks down to very small particles. The author's research group has been most interested in the fine material, rather than in specimens of the larger rocks. Table 4 is an inventory of the material returned from the Apollo 11 mission. The internal grain size of the rocks is quite variable but is generally of the order of 100 μm. Therefore, any fragment larger than approximately 100 μm is likely to be a rock, i.e., an assemblage of coexisting minerals, whereas particles finer than that are likely to be single-mineral fragments. In general, rocks are more interesting than single minerals because they contain information on a whole chemical system and not just on a fragment of that system. Most investigators requested large-rock samples; our group asked for soils, and we were delighted to find that most of the tiny soil particles were themselves rock fragments. We concentrated our studies on samples in the 1-mm to 1-cm range, and examined some 1676 of them.

There are a variety of types of rock material in the samples in this size range (Figure 21). First, the most abundant type of lunar rock, which is basalt, will be described (Figure 22).

The average chemical composition of the Apollo 11 basalts (major elements) is given in Table 5. The column on the right in Table 5 shows the average composition of terrestrial basalts from the deep ocean bottom. Differences are not great; the principal difference between the two types of basalt is in the titanium content: The Apollo 11 lunar samples contain approximately 6.0 percent by weight as compared with less than 1 percent by weight for terrestrial oceanic basalts. The sodium content is also quite different; it is approximately 2 percent in terrestrial basalts and approximately 0.3 percent in lunar basalts.

The various elements in the lunar basalts do not fit themselves into one type of crystal but into a variety of crystalline compounds. To study these compounds (i.e., minerals), the petrologist makes thin sections of rocks and studies them microscopically (Figure 23). There are three main minerals in the lunar basalt: pyroxene, $(Mg,Fe,Ca)(Si,Al)O_3$; feldspar, $(NaSi,CaAl)AlSi_2O_8$; and ilmenite, $(Mg,Fe)TiO_3$. Elements tend to substitute for one another in naturally occurring compounds like these, as indicated by the parentheses; e.g., in pyroxene, the (Mg,Fe,Ca) means that these three elements can substitute for one another in any proportion.

Another important rock type in the lunar soil, although much less abundant than basalt, is a light-colored one (Figure 24). This rock contrasted with the dark-colored basalt and was totally unexpected in the lunar sample. The predominant mineral here is high-calcium feldspar, or anorthite, with minor amounts of pyroxene. This rock is called anorthosite. Some anorthosite particles have been heavily shocked, as shown by their distorted crystalline state. Such shock damage is an expected result of the high-energy cratering events that occur on the lunar surface.

These are the two principal crystalline rock types found in the Apollo 11 lunar sample, and a first-order interpretation can be made of their significance. The explanation for the abundant basalt in the soil is straightforward enough: It is clearly a sample of the lunar bedrock underfoot at Tranquility Base and is what lunar maria are made of. Basalt is a volcanic rock of widespread occurrence on the Earth. It seems clear that Mare Tranquillitatis (and now also Oceanus Procellarum, from Apollo 12) consists of giant lava flows, or lava lakes; this was anticipated by many lunar investigators before the Apollo missions.

The anorthosites, however, were a great surprise. Anorthosite is practically the only rock type that has never been suggested as a possible lunar material. Our group,

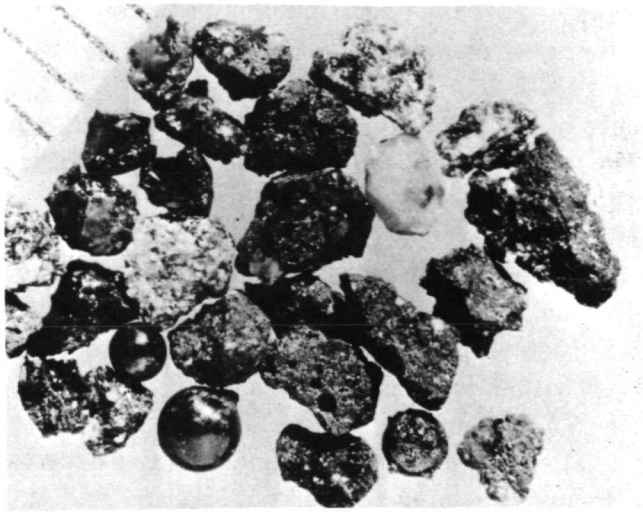


Figure 21.–Assortment of washed soil fragments in the 1- to 3-mm size range, from Apollo 11. Visible are basalt fragments (mottled white-and-black), basaltic glasses (glossy black, present as irregular forms and spherules), soil breccias (medium-gray tone), and a single fragment of anorthositic glass (white). Millimeter scale (from Wood et al., 1970a).

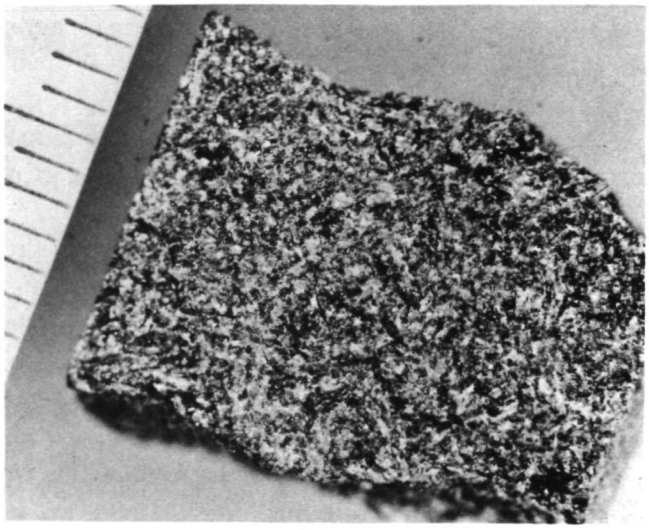


Figure 22.–Sample of basaltic rock from Oceanus Procellarum (Apollo 12 mission). Millimeter scale.

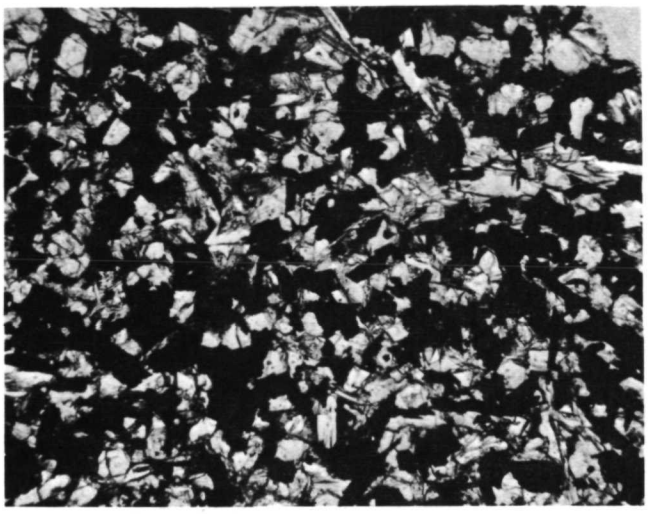


Figure 23.–Thin section of a lunar basalt, illuminated by transmitted light. The field of view is approximately 3 mm wide. Several minerals can be seen: ilmenite, which is opaque and thus appears black in the figure; and pyroxene and feldspar, transparent minerals which form a complex interlocking fabric of light-colored crystals (from Wood, 1970).

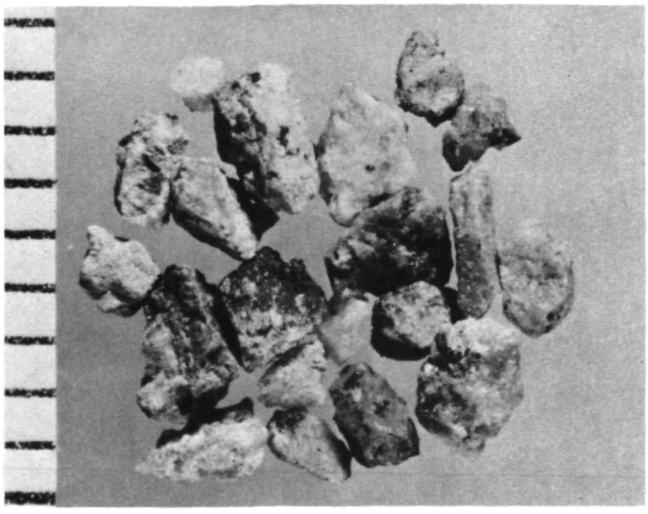


Figure 24.–Light-colored particles of anorthosite, hand picked from the Apollo 11 soil. Millimeter scale (from Wood et al., 1970a).

Table 4.–Inventory of material returned by Apollo 11.

Fragment Diameter Range	Approximate No. of Fragments Returned	Nature of Fragments
3 cm to 10 cm	36	Rocks
1 cm to 3 cm	125	
1 mm to 1 cm	10^6	
100 μm to 1 mm	10^9	
10 μm to 100 μm	10^{13}	Minerals
1 μm to 10 μm	10^{15}	

Table 5.–Comparison of basalt compositions on the Moon and on Earth.

Elements	Percentage by Weight	
	Moon*	Earth**
O	38.9	39.0
Si	22.9	28.0
Fe	14.8	6.8
Ca	7.1	8.5
Ti	6.0	0.9
Al	4.8	9.1
Mg	4.6	4.4
Na	0.3	2.0
Cr	0.2	–
Mn	0.2	0.1
K	0.15	0.13
P	0.05	0.12
	100.0	100.0

*Average of analyses of Apollo 11 type A and B rocks.
**Average of basalts from deep ocean bottoms.

and others as well, concluded that the anorthosites are not samples of the maria at all, but come from the lunar highlands. The basalts are dark and the anorthosites are light, after all, and the maria are dark and the highlands are light, so it is tempting to suggest that the anorthosites are highlands material. Moreover, the Apollo 11 site is only about 50 km from the highlands (Figure 25). In the discussion of cratering, it was noted that this process can throw some material long distances. It is therefore reasonable to suppose that some material from the highlands would have gotten mixed in with the mare material at Tranquility Base. Shoemaker et al. (1970) have

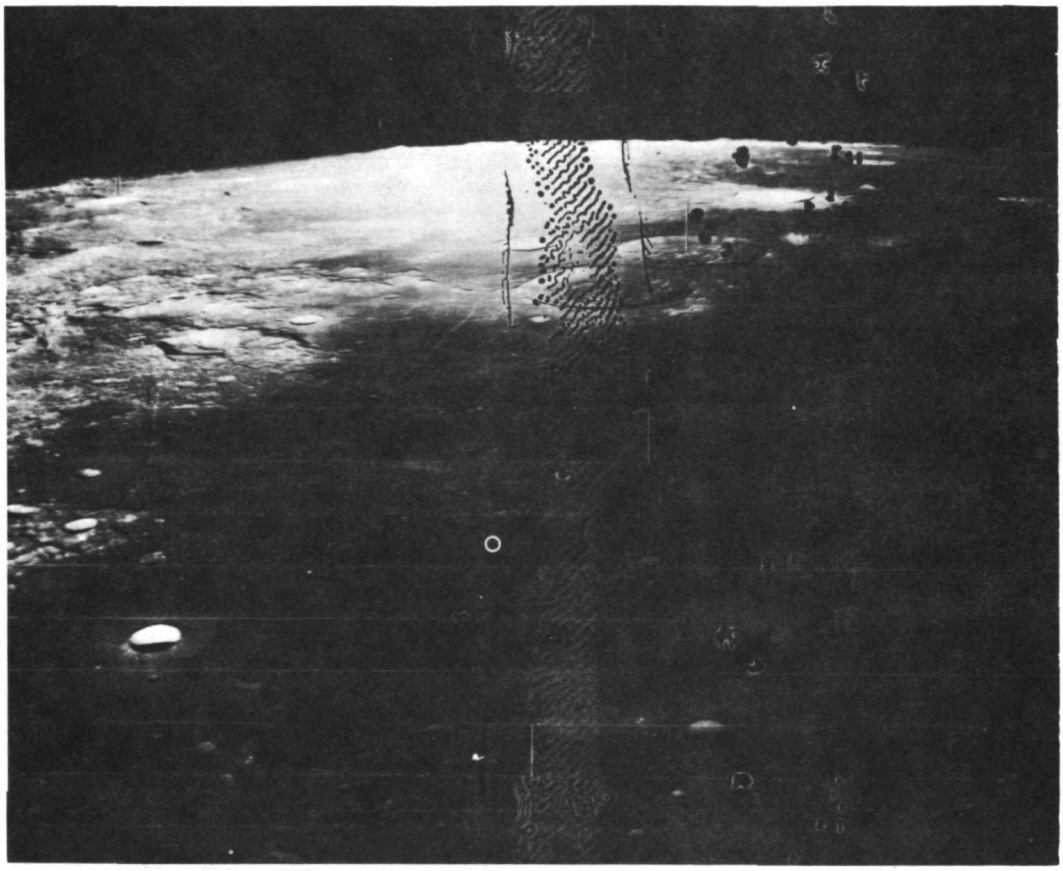


Figure 25.—Lunar Orbiter photograph of the Apollo 11 landing site (circle) on the vast dark plain that is Mare Tranquillitatis. The view is west along the lunar equator; looming to the south (left in figure) are the lunar highlands, composed of a different and lighter-colored type of rock (from Wood et al., 1970a).

recently calculated, from considerations of cratering dynamics, that the mare material should have about 5-percent highlands material mixed in. This turns out to be very nearly the amount of anorthosite found in the lunar soil.

This argument is greatly strengthened by the comparison (Figure 26) that can be made between the composition of the anorthosite, as measured by electron microprobe in our laboratory, and that of the lunar highlands, as measured by the alpha-backscattering experiment carried by the Surveyor 7 spacecraft, which landed in the highlands (on the ejecta blanket of the great crater Tycho). The agreement in composition between the two materials is excellent, except in the case of fluorine, which has not been measured in lunar anorthosites. It can be expected that such high levels of fluorine would not be present in the anorthosites, since no fluorine

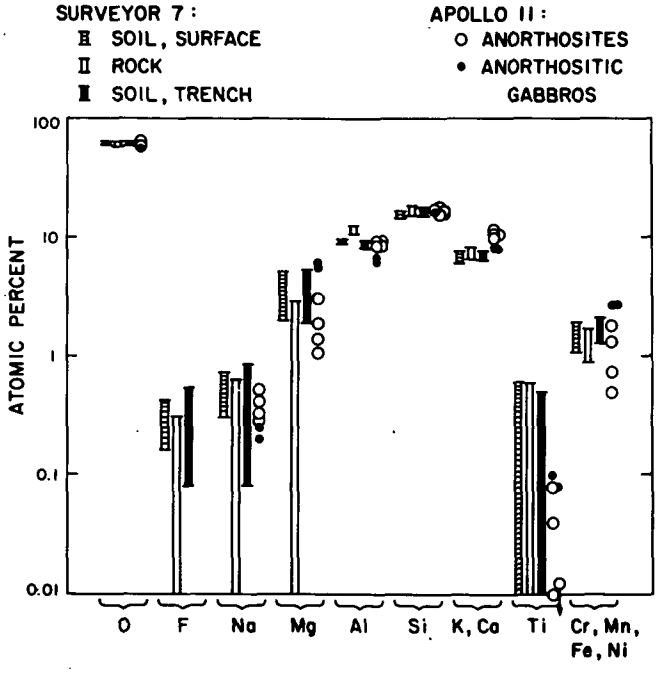


Figure 26.—Comparison of compositions of lunar highlands materials in the vicinity of the crater Tycho, determined by the Surveyor 7 alpha-backscattering experiment (data shown as error bars, which reflect 2-sigma counting statistics), with compositions of six anorthositic particles from the Apollo 11 soil, determined by microprobe analysis (from Wood, 1970).

minerals were observed. In spite of this discrepancy, however, the similarity of compositions is very convincing.

The following is a list of things the lunar soils might be expected to contain:

(1) Fragments of meteoritic projectiles.
(2) Fragments of lunar bedrock, essentially intact.
(3) Fragments of lunar rock in a degraded form.

The basalts and anorthosites fall into category 2 and are the most important components of the lunar samples because it is possible to learn the most about the evolution of the Moon from them. However, degraded materials are also present, and these deserve some discussion.

Recall from the discussion of cratering that some material is melted, immediately at the point of impact. The molten rock may harden in flight into spherules, or, after splashing on the surface, into irregular forms (Figure 27). The glass ranges in hue from colorless to dark brown, the differences being related to variability of chemical composition.

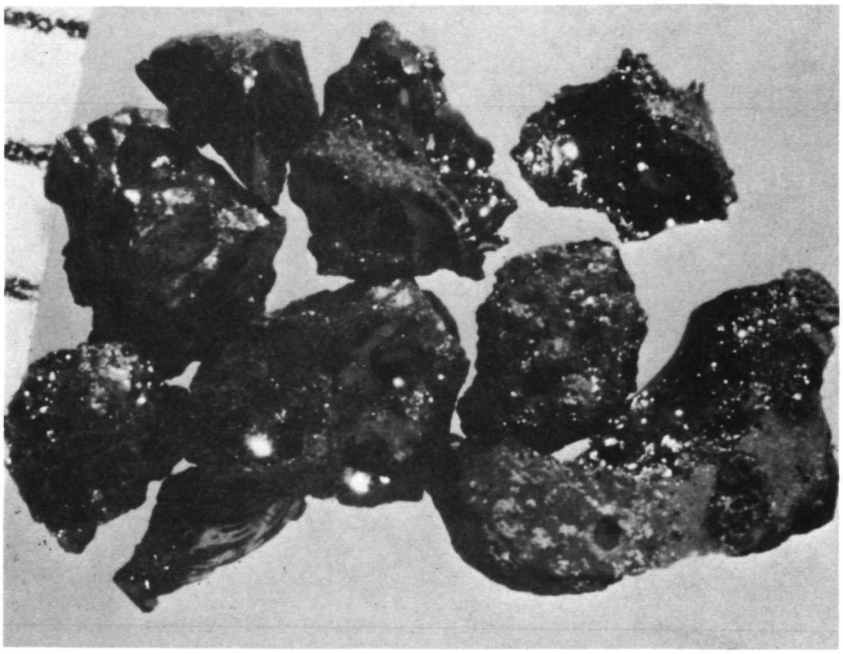


Figure 27.–Irregular fragments of glossy black glass from the Apollo 11 soil sample. Note vesicles (bubbles) (from Wood et al., 1970a).

A large proportion of the Apollo 11 sample consisted of a rock type called breccia. This is an aggregate of angular fragments of minerals, glasses, and rocks (Figure 28). Breccias are formed from loose soil via shock lithification, in a region somewhat farther away from the impact point than the small zone in which total melting occurred (Figure 19).

Figure 29 expresses the chemical composition of soils and breccias, anorthosites, and basalts on a triangular, three-component phase diagram. This type of diagram overlooks some of the possible compositional differences, but it is impossible to draw diagrams in all the dimensions required to represent rock

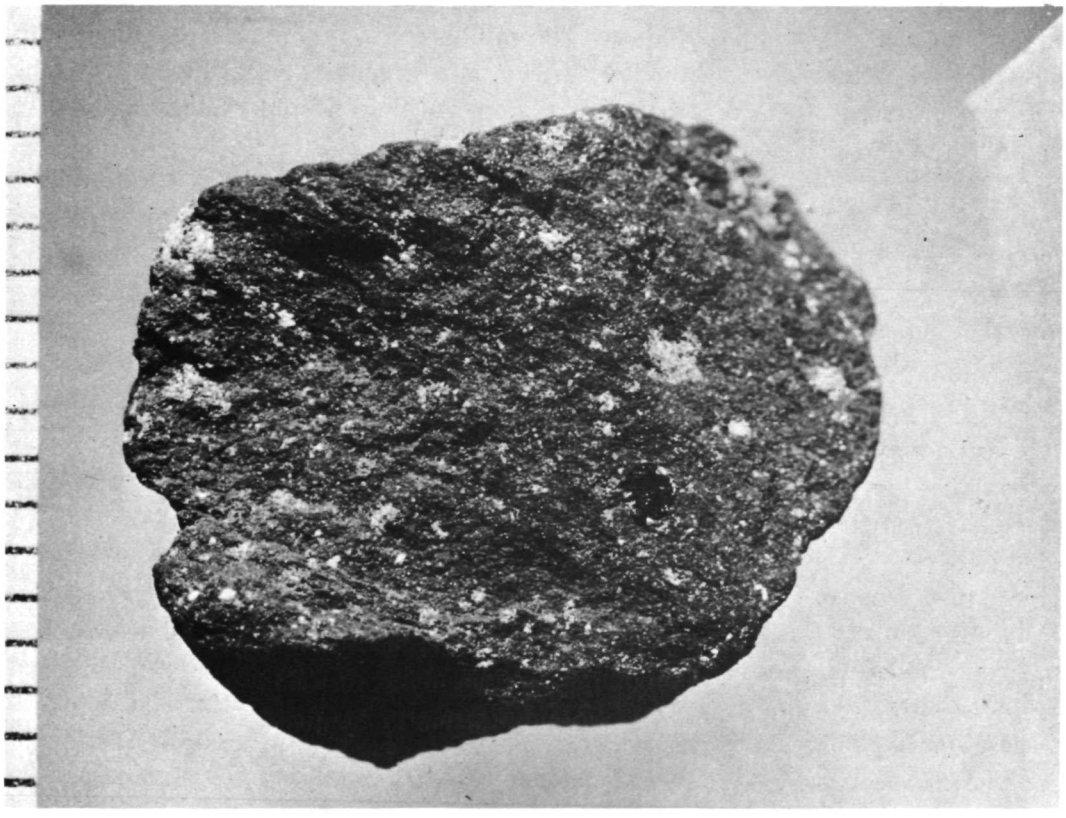


Figure 28.—Lump of soil breccia from Tranquility Base; an aggregate of mineral, rock, and glass particles embedded in a matrix of fine soil that has been shock lithified into a solid mass. Round black structure is a split-open glass spherule. Millimeter scale (from Wood et al., 1970a).

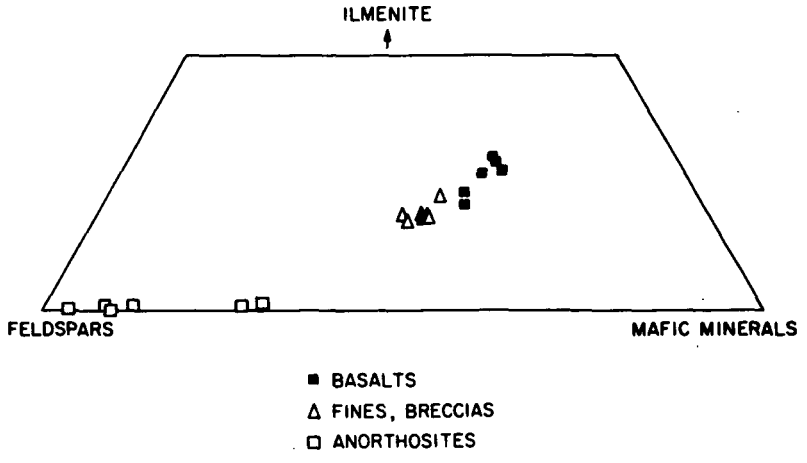


Figure 29.–Compositions of components of the Apollo 11 sample plotted in terms of their relative contents of three principal minerals. The data actually come from bulk chemical analyses of lunar substances but had to be recalculated in terms of three components in order to be plotted in two dimensions (from Wood et al., 1970b). Pyroxene is the dominant mafic mineral.

compositions containing eight or more chemical constituents. Therefore, the bulk chemical composition of each rock has been recalculated in terms of the three major mineral components shown. The diagram represents the trapezoidal base of a triangular diagram.

The three groups of rocks plot rather compactly. The soils and breccias consist largely of crushed basalt; the bedrock beneath Tranquility Base is basalt which has been ground up into a regolith during the course of geologic time, and the regolith has been subsequently lithified (locally) into breccia. So it would seem the breccias ought to have the same composition as the basalt; but they do not, and a reason suggests itself. Some alien material, apparently the anorthosite, has been mixed into the basaltic regolith. Any mixture of anorthosite and basalt should lie between the two end members in Figure 29, and this is what is observed.

A more detailed plot of the chemical compositions (Figure 30) shows more convincingly that the breccias (and the soil) are a mixture of basalt and anorthosite; the breccia composition corresponds satisfactorily to a mixture of 20-percent anorthosite and 80-percent basalt. Only about 5 percent of the visible fragments in the

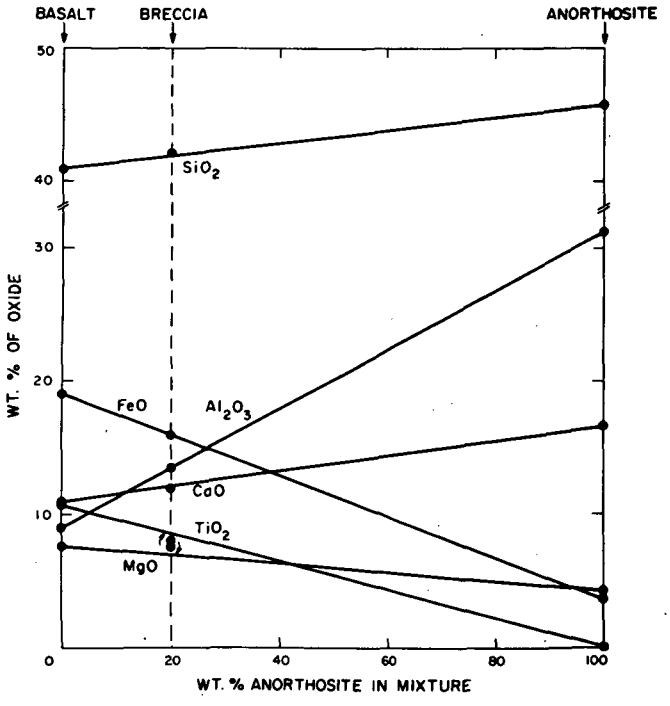


Figure 30.–Plot of major element contents of average Apollo 11 basalt, breccia, and anorthosite, showing that breccia composition corresponds closely to a mixture of 20-percent anorthosite and 80-percent basalt (from Wood et al., 1970b).

soil and the breccias are anorthosite, but it is impossible to determine compositions of the more abundant very small grains or glassy beads in the breccias, and these may contain larger proportions of anorthosite.

Figure 31 shows the compositions of the Apollo 11 glasses; the dashed lines reproduce the groupings of rock compositions from Figure 29. The color of the glasses noted refers to the colors in the thin section. The glass compositions cluster in two groups, one (colorless) corresponds to anorthosite compositions and the other corresponds to basaltic soils and breccias (darker colored glasses). Clearly, one family of glass is produced by impacts in the maria and the other by impacts in highlands anorthositic terrains.

Table 6 is a breakdown of the proportions of the various rock types surveyed in our sample of about 1676 fragments.

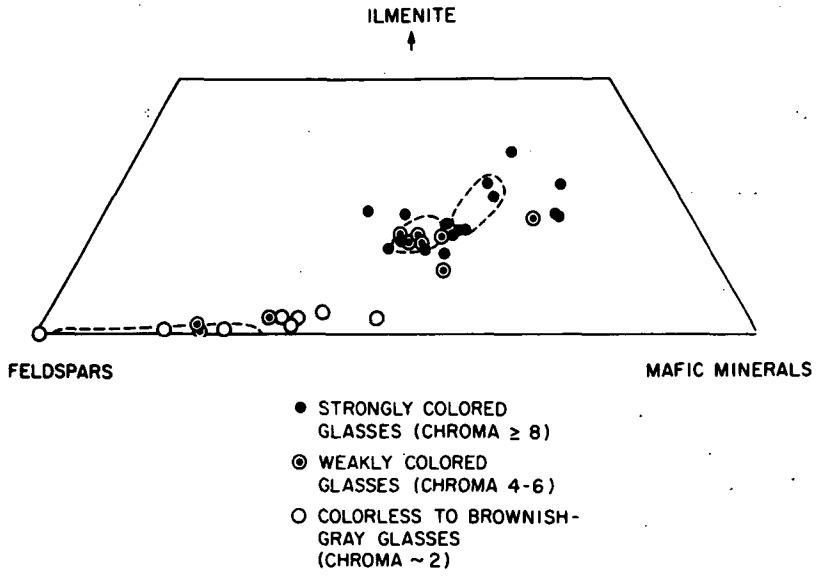


Figure 31.–Compositions of lunar Apollo 11 glasses compared with rock compositions (fields surrounded by dashed lines) from Figure 29 (from Wood et al., 1970b).

Table 6.–Percentages of rock types found among 1676 coarse (> 1-mm) particles surveyed (from the Apollo 11 soil sample).

Constituent	Percentage by Weight	
Crystalline basalt	37.4	
Basaltic glasses	4.3	
Basaltic breccias	52.4	
Basalts		94.1
Crystalline anorthosites	2.0	
Anorthositic glasses	1.5	
Anorthositic breccias	1.5	
Anorthosites		5.0
Others (including meteorite fragments)		0.9
Total		100.0

Among the large rock samples returned by Apollo 12, there are considerably fewer breccias than in the Apollo 11 samples. Some have suggested that this means the regolith is much less evolved at the Apollo 12 site than at Tranquility Base, so that too little fine material was present to make a substantial amount of breccia. This seems to be erroneous; we found abundant breccia in our Apollo 12 soil samples. The ages of the rocks from the two sites are not very different. Moreover, the regolith thicknesses in the two areas are not much different, according to the estimates made by Quaide and Oberbeck using their crater morphology techniques. Possibly, the difference is an operational one; for Apollo 11, the flight planners were cautious and made a point of planning a landing in the smoothest and safest place they could find, which meant being as far away from craters as possible. The result was a site that had not been disturbed by a major cratering event for a very long time. Recall from the discussion of cratering that the regolith occasionally receives a major impact that turns things over and mixes everything up, but normally receives only small impacts that affect nothing but the upper few centimeters. This means that over a period of time a profile is built up such that coarse boulders are at the bottom of the regolith, and progressively finer material is nearer the top. The material in the topmost few centimeters will have been bombarded by so many small impacts that it will have been ground into a very fine powder. When material of this sort is impacted, it is readily turned into breccia. This was presumably the kind of area in which Apollo 11 landed. On the other hand, Apollo 12 came down on the edge of the Surveyor crater, and several other large craters were nearby. Large craters such as these have dredged up coarse material from the bottom of the regolith, as well as from fresh bedrock, and distributed it on the surface. Therefore, much of the material on the surface was derived recently from the bottom of the regolith, not from shock lithification of fine material from the top, as was the case at the Apollo 11 site.

IV. EARLY INTERPRETATIONS OF PROPERTIES OF APOLLO SAMPLES

In the previous description of lunar soil, it was noted that approximately 5 percent of the sample consisted of a type of rock called anorthosite and that this was unexpected. What this peculiar rock implies in terms of the structure and history of the Moon will be discussed next. Anorthosite is very unusual in that it is composed almost entirely of a single mineral, calcic feldspar (anorthite), of

approximate composition $CaAl_2Si_2O_8$. It is always an unusual thing when nature conspires to bring a lot of one particular mineral together. Before going any further in understanding how this has happened, we should try to see how much of this anorthosite is present on the Moon. This problem can be approached if some of the gravitational properties of the Moon are considered. Figure 32 shows idealized block diagrams across a portion of the Moon; the slightly higher region represents the highlands, and the lower region represents the mare. In Figure 32(a), the Moon is assumed to be made of a single rock type having a uniform mass density. If this were the case, then, when a satellite such as one of the lunar orbiters flew over this terrain (from right to left), it ought to have been deflected downward slightly when it passed over the highlands region. The highlands region would exert a greater gravitational pull than the mare, simply because it has an extra amount of mass in it. It would display a positive gravity anomaly. One can calculate that for the fairly well-known height of the highlands, the anomaly relative to maria would be about 200 mgal.* This proposition has been tested by the lunar orbiters, since their acceleration in the direction of the Earth can be measured very precisely by the doppler shifts in the radio signals they transmit. The results of such measurements are shown in Figure 33, which is a compilation of the distribution of gravity on the near side of the Moon. The shaded regions are maria, the white regions highlands. Over the highlands, there are few variations in the gravitational attraction of greater than 50 mgal, although several positive gravity anomalies (mascons) were found over maria regions. However, nowhere is there a systematic increase in gravity as one goes from mare to highland. This means that the landscape is isostatically compensated, and must look in cross section approximately like Figure 32(b); the highlands must be underlaid by some thickness of a lower density material. This lower density material is effectively floating in denser material, i.e., buoyant equilibrium prevails.

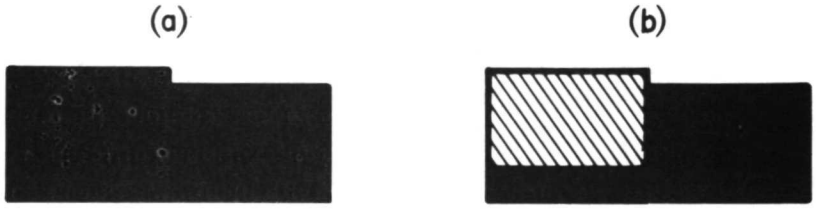


Figure 32.–Block diagrams suggesting possible structure under the lunar highlands and maria. Hatched material has lower mass density than black material.

*1 gal = 1 cm/s^2.

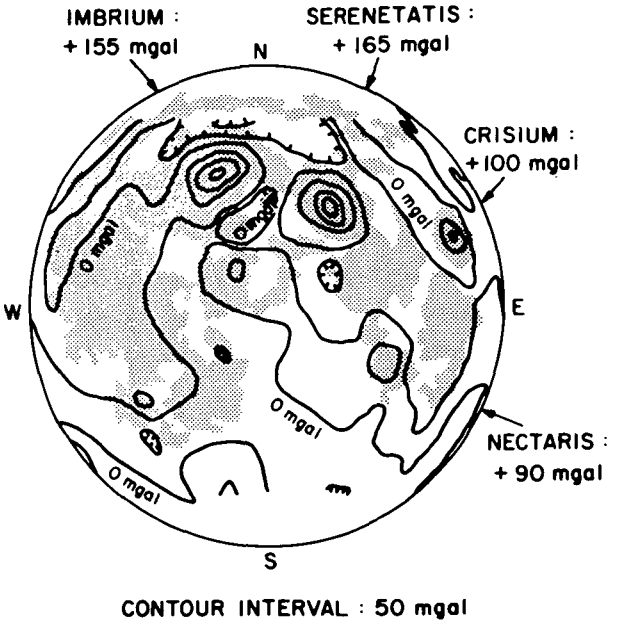


Figure 33.—Gravity map of the near side of the Moon, replotted from data of Muller and Sjogren (1968). Major positive gravity anomalies (mascons) are centered on the circular maria (from Wood, 1970).

Under these circumstances, there would be no significant differences between the gravitational attraction over the highlands and that over the maria.

Earlier, the highlands material was tentatively identified as anorthosite, and the mare material was identified as basalt. The thickness of the anorthosite plate is simple to calculate if one knows the difference in mass density between the two materials (which we can measure from the samples we have) and the mean elevation of the highlands above the maria. The latter is only a moderately well-known quantity. The best Earth-based data have been obtained by observing certain topographic features on the Moon at different libration angles; from the stereo effect produced, one calculates the position in space of these features and derives from that their distances from the center of the Moon. Figure 34 shows the whole body of these data as prepared by the United States Air Force Aeronautical Chart and Information Center (ACIC). The histograms for maria and highlands are offset from one another by a distance of 1.4 km, which represents a mean elevation difference between maria and highlands. However, it is perhaps unsafe to take this

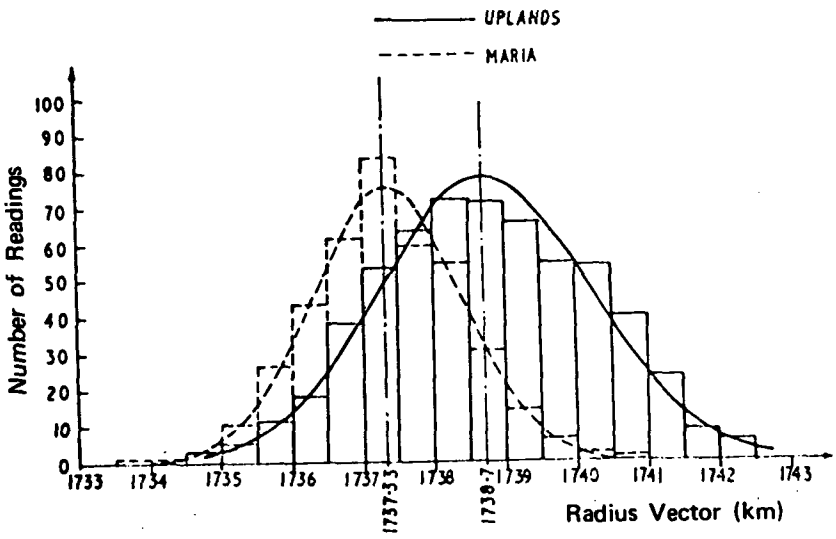


Figure 34.—Histograms of elevations of control points on the lunar surface, made by the USAF ACIC. Two histograms are shown: one for control points in the uplands, the other for points in the maria. The lunar uplands are higher than the maria by 1.4 km, on the average (from Runcorn and Shrubsall, 1968).

number, which applies to the Moon as a whole, and apply it to the Mare Tranquillitatis region in particular. Figure 35 shows a plot of relative elevations of some topographic features in the neighborhood of Tranquility Base. The black bars represent heights in Mare Tranquillitatis, and the light ones represent points in the highlands to the south and southwest of Mare Tranquillitatis. The uncertainties are those stated by the ACIC; these are probably rather optimistic. Nevertheless, one can see at least a qualitative difference between the heights of the highlands and of the mare, and it appears to amount to 1 or 2 km.

If 1.5 km is taken as the real difference between Mare Tranquillitatis and the adjacent highlands, it turns out that, with densities of 3.3 g/cm^3 for basalt and 2.9 g/cm^3 for anorthosite, the total thickness of the anorthosite plate must be about 10 km. One could argue that these densities are not representative of the material at depth, because only material from near the surface has been sampled; if the densities did change at depth, however, it is unlikely that the density contrast would become even greater than the one used here, which is an extreme case. If the rock beneath the anorthosite is of greater density than the 2.9 g/cm^3 used, the floating layer

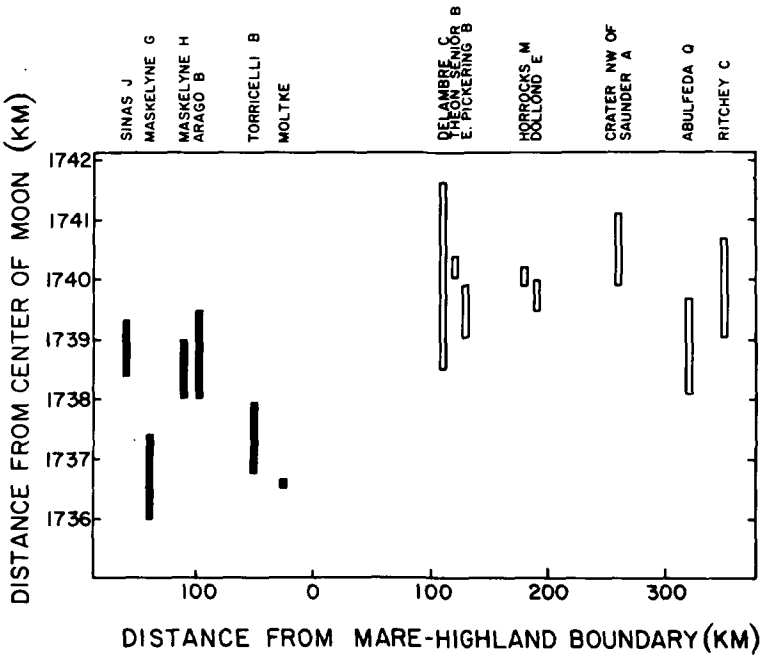


Figure 35.—Plot of elevations of ACIC control points in the vicinity of Tranquility Base. Black control points are in Mare Tranquillitatis; white are in central highlands.

would have to be more than 10 km thick in order to project 1.5 km above the mare. Thus, 10 km can be taken as a lower limit for the thickness of this crust that underlies the highlands.

But why an anorthositic crust on the Moon, a layer consisting of little other than the mineral plagioclase? An explanation for such a concentration of one mineral can be gained from an examination of the behavior of rocks at high temperature. A rock is a very complex system having a number of different elements in it. When heated, it does not simply melt at a given temperature the way a piece of pure ice or iron would. Instead, it melts over a range of temperatures and in a rather complex way. For example, the composition of a typical basaltic rock system may be represented by two components: anorthite (feldspar), $CaAl_2Si_2O_8$; and diopside, $CaMg(SiO_3)_2$. The phase diagram for such a system is shown in Figure 36, with the proportions of these components plotted against temperature. At low temperatures, these phases do not form a continuous solid-solution series, but instead they form a physical mixture of the two minerals in a proportion determined by the bulk

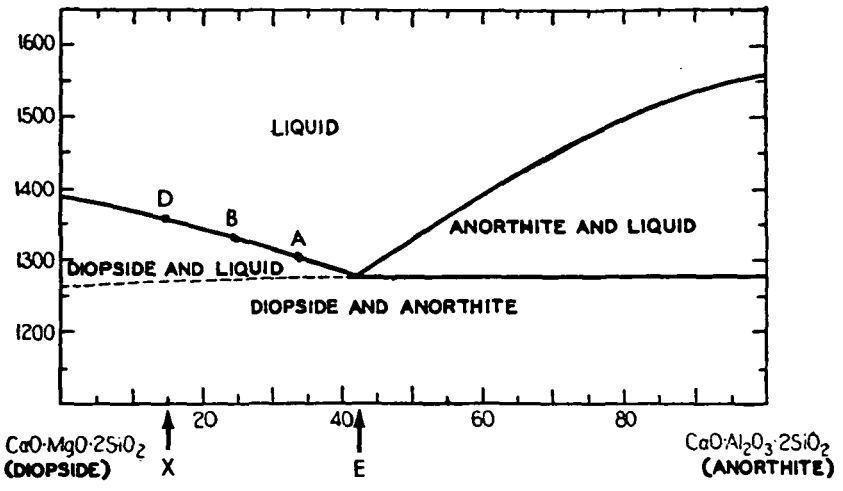


Figure 36.–Phase diagram for the system diopside-anorthite (from Osborn, 1942).

chemical composition of the rock. From the phase diagram, it is seen that each pure phase melts at a relatively high temperature, but that mixtures of the two components begin to melt at lower temperatures along the curved lines (melting curves or liquidus). Rocks of bulk composition *E* melt at the lowest temperature (eutectic point). On either side of the eutectic point, regions are found where both a liquid phase and a solid phase are present. Consider what happens when a system of bulk composition *X* is heated. As soon as the temperature reaches the eutectic temperature, some melting begins to occur. The composition of the initial melt is not *X*, but rather *E* (the eutectic composition). As the temperature continues to rise, more and more of the solid residue melts, and the composition of the liquid changes from *E* to *A* to *B* and so on, until the whole system is molten at point *D*; the liquid then has a composition equal to the bulk composition (*X*) of the system. The process works exactly symmetrically in reverse: On cooling, when point *D* is reached, crystals of diopside precipitate, and the amount of diopside increases as the remaining melt changes in composition from *D* to *E*. Finally, plagioclase precipitates, and the final, completely solid system is a mixture of these two phases, in the proportions determined by *X*.

A complication arises because the two solid phases and any coexisting liquid all have different mass densities. If this situation exists in a gravitational field, and if the

system is not turbulent and does not melt or freeze completely too rapidly, there will be a tendency for the diopside to sink and separate from the liquid. Suppose such a system were partially melted inside a planet. The heavier residual crystals would tend to sink, and the liquid would tend to become concentrated near the top of the system. The liquid might find fissures and be extruded onto the surface of the planet. Note that the liquid extruded would have a composition substantially different from the overall material of the planet: It would be richer in plagioclase (i.e., partial melting of a system of composition X may cause a liquid of, for example, composition A to form; gravitational settling would remove the solid phase, thus leaving a liquid which is enriched in plagioclase relative to the initial composition X).

Now suppose that this liquid containing approximately equal amounts of diopside and plagioclase were removed and concentrated either directly on the surface or close beneath it. The liquid would now begin to cool and crystallize. Diopside and plagioclase crystals would appear. The diopside crystals would certainly be denser than the residual liquid and, hence, would sink. The plagioclase crystals might or might not be denser; there is a chance they would float rather than sink. The final result of such a process would be a layer of diopside at the bottom of the system, a layer of plagioclase (i.e., anorthosite) at the top, and a layer of liquid of some intermediate composition in between. If such a differentiation occurred at the surface of the Moon, the plagioclase layer, after solidification was complete, would comprise an anorthosite crust.

It seems that this is the only mechanism adequate to explain the existence of large masses of anorthosite on the Moon. This requires that the Moon, early in its history, had a partially molten layer at its surface. How thick must this layer of melted material have been? The anorthosite is known to be approximately 10 km thick, and these 10 km are only a fraction of the total thickness of the igneous system that is required to produce this layer. At the very least, such a system would deposit an equal thickness of diopside and/or other mafic minerals. So, a layer of melted rock that was at least 20 km thick must be contemplated. If the Moon is assumed to have a bulk composition similar to the silicate phase of chondritic meteorites (thought to be the most primitive type of meteorite and containing elements in proportions very similar to the cosmic abundance), then, from simple mass-balance considerations, it turns out that the upper 200 km of the Moon would have to melt for enough of the needed elements, especially aluminum, to be available to make 10 km of anorthosite. It is not necessary to melt these 200 km completely; it is only necessary to partially melt the material because, as is noted in

Figure 36, the first liquid to be produced carries with it most of the aluminum in the system. Figure 37 summarizes the model under discussion.

Another question one may ask is whether this early anorthositic crust extended all around the Moon in a symmetric shell (which was subsequently punctured by large impacts to form the maria) or whether it was localized. One cannot say for certain, although the highlands are of uniform brightness on both sides of the Moon. J. Salisbury et al. (1970) of the United States Air Force Cambridge Research Laboratories have made measurements of infrared emissivities of various regions on the Moon; these tell something about the mineralogy of the rocks. The results suggest that the material in the bottoms of two large lunar-highland craters has the same spectral properties as that in the vicinity of the crater Tycho, which is known to be anorthositic. This evidence suggests that the highlands are rather uniform in their properties.

It appears, then, that the outermost 200 km of the Moon was heated at least to the point at which it began to melt. This outer 200 km embraces about one-third of the volume of the Moon. How did the Moon get as hot as this, and what does this mean in terms of the evolution of the Moon? It is not feasible to appeal to the decay of radioactive elements such as U, Th, and K to explain this melting. In the near-surface regions, heat is lost easily, and if generated on the time scale of long-life

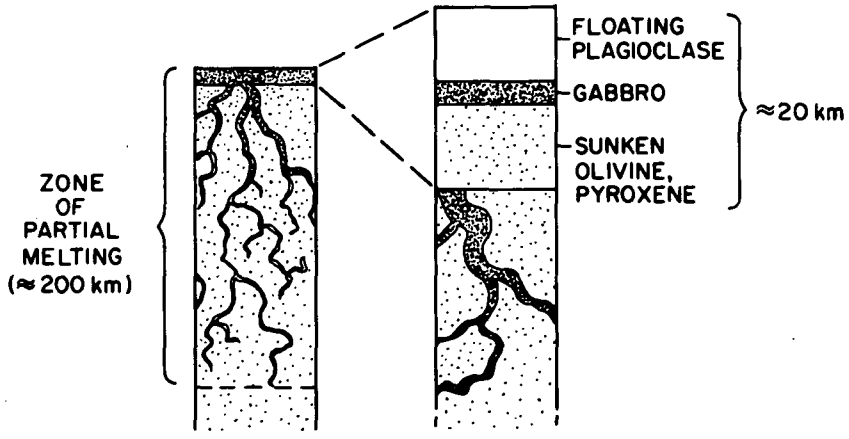


Figure 37.—Postulated early lunar-surface magma system during stage of partial melting and migration of basaltic magma toward surface (left), and enlarged sketch showing layers of floating plagioclase (anorthosite) and sunken mafic minerals produced by density fractionation of crystals (right).

radioactivity decay, it cannot accumulate to the extent needed to raise temperatures to the melting range. If such melting occurred in the interior, the melt would tend to come up to the surface in small increments and be squeezed out, and each increment would have a chance to solidify before the next one came along. This is not an adequate mechanism of producing anorthosites. The magma on the surface of the Moon should be 20 to 30 km thick and all liquid at one time in order for crystal flotation to produce a 10-km layer of anorthosite. Hence, some mechanism of heat generation is needed that is much more drastic and violent than the decay of long-lived radionuclides.

The best possibility is that the Moon retained a fairly large fraction of its accretional energy in the form of heat. If accretion proceeded at a slow rate, such that the impact of each particle could radiate away the heat it produced, the Moon would not heat up significantly. But if the accreting material came in very rapidly and in the form of large masses (i.e., preaccreted moonlets) rather than in small bits, there is a good chance that much of the energy would remain trapped in the system as heat. The resulting distribution of temperature would mean that the lunar interior would be cooler than the outer layers because the moonlets would be added to the accreting Moon at the escape velocity, which of course would increase as the Moon grew larger. The last to accrete would have the highest velocity, hence, the most kinetic energy to be converted to heat. If all particles accreted at the escape velocity, and if all the energy of accretion were conserved as heat, the final temperature distribution would be approximately that shown in Figure 38. The lunar magmatic history discussed above appears to require that the Moon accreted in such a way that much of the accretional energy was retained, so as to melt approximately the outer 200 km.

This is a novel initial condition for the Moon, and it inspires one to perform some heat-flow calculations for the Moon. There has been a long history of lunar heat-flow calculations, but nobody has used this starting condition. One either started with a cold Moon and let it heat up by radioactivity, or if one thought that initial temperature played an important role, one gave the Moon some uniform initial temperature. In the following heat-flow runs under the initial condition previously stated, the various parameters that are important to planetary heat flow (but are poorly known) have been juggled in an attempt to understand the relationship between this hypothetical early epoch of high temperature and the eruption of lavas into the mare basins substantially later.

Figure 39 shows some simple heat-flow calculations. This figure (and subsequent ones) does not give the temperature profiles as a function of time for the

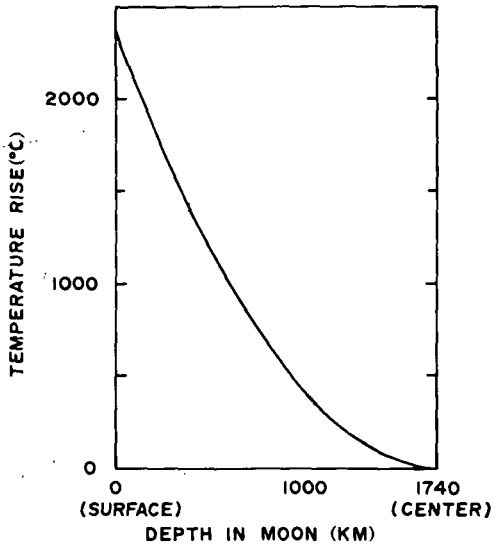


Figure 38.–Temperature rise in an accreting Moon, as a function of radius, on the assumption that accreting objects strike at escape velocity and all kinetic energy is conserved in the form of heat.

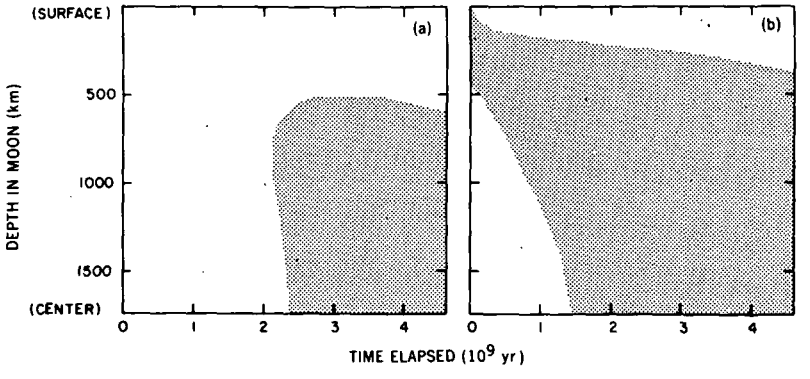


Figure 39.–Melting diagrams for lunar models in which magma migration does not redistribute radioactivity. Shaded regions are above the melting temperature of basalt. Chondritic levels of radioactivity are assumed. (a) Initially cold Moon (uniform starting temperature of 0°C); (b) Moon in which 100 percent of accretional energy is retained as heat, causing melting in the outermost 400 km of the Moon (from Wood, in preparation).

planet, because these tend to make rather complicated diagrams. What is really important is whether a portion of the Moon is melted or not, and that is all the information given on these plots. The shaded areas represent the times and depths in the Moon at which basalt would melt. The program takes into account the way in which the melting temperature of basalt varies with pressure. Figure 39(a) assumes that the Moon was originally cold and that the amount of radioactivity in the Moon is equal to that found in chondritic meteorites. (This last assumption is probably not valid; the proportions of radionuclides appear to be different in the Moon and in chondrites, but the calculations are not affected drastically by relative proportions, as long as the total amount of radioactivity is equivalent.) Melting occurs after about 2 billion years, but it is confined to the lower two-thirds of the Moon's radius. This does not mean that the lunar interior is now a liquid, because a eutectic temperature was used, i.e., the melting temperature of basalt. Thus, there would be some liquid and much solid material, as can be seen from the phase diagram presented earlier (basalt melts between 1100° and 1200°C).

Figure 39(b) shows what would happen if, instead of choosing 0°C as an initial temperature, we had assumed that the Moon retained all of its original accretional energy in the form of heat. Initially, the outer layers, down to 500 km below the surface, are above the melting point of basalt. As the surface loses heat by radiation, it begins to solidify, and the depth beneath which melting is possible increases, rather swiftly at first, then after a solid crust about 100-km thick is formed, more slowly. The important point is that there is a monotonic increase with time of the depth of the zone at which basalt is molten. The frequency of surface igneous activity (volcanism) on the Moon might be expected to decrease with time, as the depth to the melted zone (i.e., to the source of the lava) increased. However, the evidence is that the volcanism on the Moon has not decreased in such a fashion, although the evidence is very fragmentary. The basalts that have been dated from Apollo 11 range in age from 3.5 to 4.0 billion years, with an average of 3.6 billion years. Apollo 12 basalts average 3.3 billion years in age. From admittedly totally inadequate statistics, it is tempting to conclude that the igneous activity on the surface of the Moon has been episodic and not a monotonically decaying process as described earlier. It appears that there was a burst of activity about 1 billion years after the Moon was formed, but it must be emphasized again how fragmentary the evidence is.

Clearly, the heat-flow calculations do not fit such a history. However, something important has been left out of the calculations. The melting in the interior of a planet tends to be self-defeating because when melting begins, the first

liquid that appears carries the lion's share of radioactive elements (U, Th, and K). If this initial low-density liquid is squeezed up toward the surface of the planet soon after it melts and carries most of the radioactivity with it, the source region in the Moon is no longer able to heat itself as well. As soon as melting begins, the melted zone becomes depleted in the material necessary to continue the melting. The heat-flow program was changed to allow for this phenomenon. One problem is that one does not know how efficient the process is, so the efficiency of radioactive transfer has been treated as an unknown parameter that can be varied in each run. Figure 40 shows a matrix of possibilities, with two parameters varied. In all cases, chondritic radioactivities were assumed. Different transfer efficiencies have been tried, and different amounts of accretional energy retained as heat. The top row of panels looks peculiar because, if one tells the program that the Moon is 100 percent efficient in cleaning out its radioactivity when it reaches the melting point, one has set up a self-governing system. At any given point along the radius of the Moon the temperature will rise until it gets to the melting point; as soon as this happens, all of the radioactivity is removed, the temperature cannot rise any more, and that point can only cool. Once cooled below the melting point, it becomes a trap for any radioactivities rising from below, and then it begins to heat up again. Thus, a given point oscillates between temperatures above and below the melting point. What is plotted are those instances when an area was momentarily above the melting point.

The interesting thing about most of the plots shown is that no matter how the parameters are varied, the melting diagrams show little humps on top. These are times fairly early in the history of the Moon when the increase in depth to the melted zone reverses itself, i.e., the melted region rises closer to the surface. It is very tempting to identify this event in the thermal history of the Moon with the hypothetical resurgence of igneous activity on the surface of the Moon that seems to be indicated by the ages of lunar basalts. The trouble is that it does not come at quite the right time: In order to satisfy the ages of Apollo 11 and 12 Moon rocks, it should come at 1 billion years after the formation of the Moon, but the humps in Figure 40 come somewhat later.

It is possible to adjust some of the free parameters so that the peak will come just 1 billion years after the Moon formed. The reason for the occurrence of the peak is clear: As melting occurs at increasingly greater depths in the Moon, and as radionuclides are removed and concentrated in the upper layers, so much radioactivity is finally piled up in this rather restricted region that it temporarily reverses the trend of heat loss and begins to increase in temperature, ultimately causing melting to occur nearer to the surface than previously. To make this event

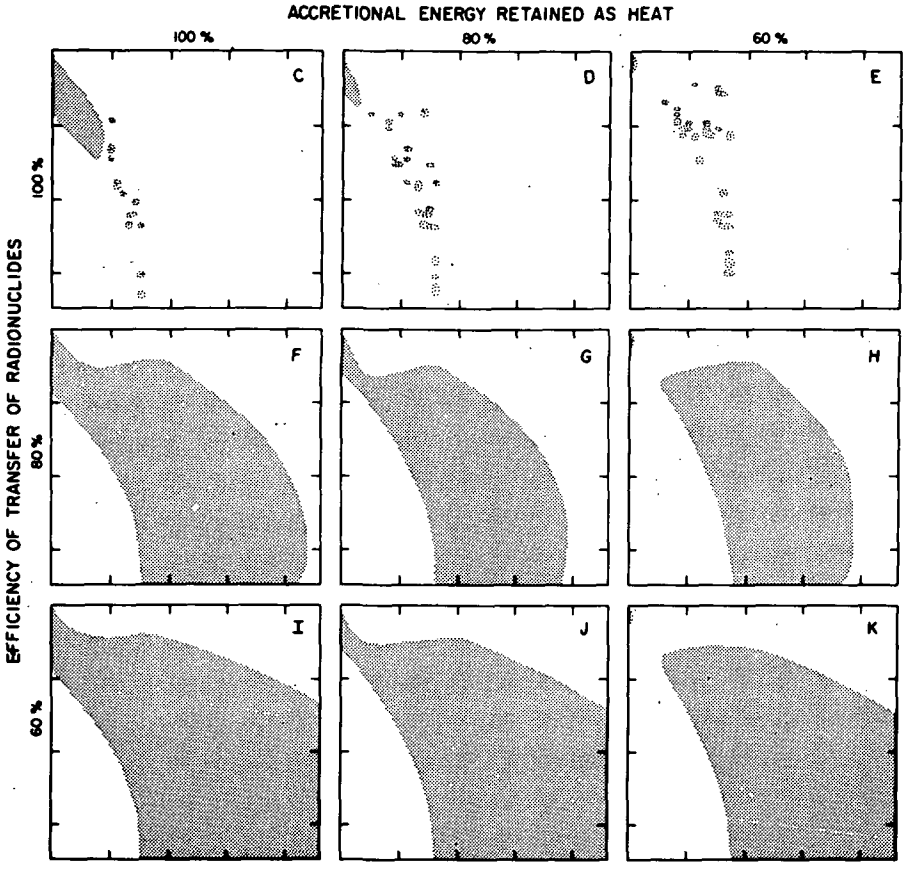


Figure 40.—Melting diagrams for lunar models in which magma migration redistributes radioactivity. Several different transfer efficiencies are tried, and several values for completeness of retention of accretional energy. These models display "humps" 1.5 X 10^9 to 2 X 10^9 years after the lunar origin when the zone of melting rises in the Moon. Chondritic radioactivity, *B* (base initial temperature) = 0°C (from Wood, in preparation).

occur earlier, one must cause melting at depth to occur earlier. Therefore, the early Moon must be made hotter, either by assuming a greater content of radioactivity or a higher initial temperature. It turns out that either one of these mechanisms or a combination of them will work. Figure 41 shows what happens when the radioactivity level is increased. It can be seen that the hump is moved over toward 1 billion years if the abundance is increased to 1.5 times the chondritic abundances.

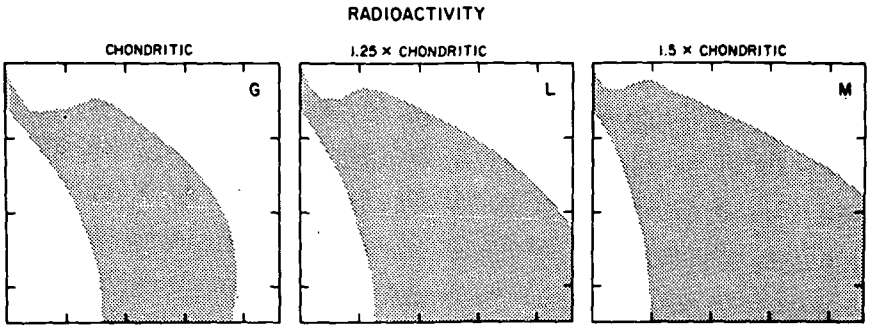


Figure 41.—Comparison of model G (Figure 40) with models that contain higher levels of chondritic radioactivity but which are otherwise similar to G. Additional radioactive heat generation causes the "hump" to occur earlier (from Wood, in preparation).

Figure 42 shows the effect of varying the initial temperature. The previous calculations had the accretional energy superimposed on 0°C; i.e., the material was at 0°C before accreting on the Moon. If a higher temperature were supposed instead, the peaks will be shifted forward as the temperature increases; it takes 600°C, plus accretional energy, for the peak to come close to the desired point.

This seems to say that if the idea of episodic magmatism is to be taken seriously a very hot origin for the Moon must be assumed, or a great deal of radioactivity, or both. In any case, there is no escaping the idea that the Moon must have melted to its center very early in its history.

In closing, the related problem of mascons on the Moon should be considered. Figure 33 shows the locations of the lunar mascons. Something has happened in these areas to generate additional gravity; i.e., extra mass has been concentrated in these areas of the Moon. The author feels mascons can be understood quite simply in terms of the deposition of lavas on the surface of the Moon. There are two requirements for these lunar mascons. First, an incentive is needed for the lava to rise to the surface. Second, a means of localization of the lava is needed, because it will not form a mascon if it merely floods the whole surface of the Moon. The author feels both of these requirements can be filled. In the first case, it is largely overlooked that the density of igneous rock decreases by about 10 percent when it melts; it is lighter than overlying solid rock, so it tends to rise above it. However, if the Moon were homogeneous and everywhere perfectly compensated (i.e., if it were in gravitational equilibrium), it would have a perfectly spherical surface; any lava

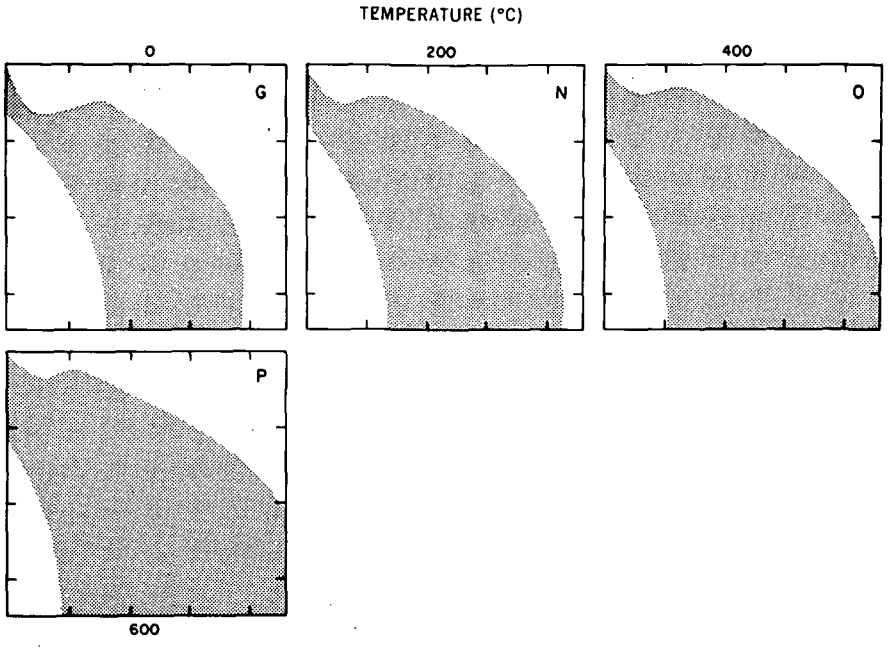


Figure 42.—Comparison of model G (Figure 40) with models that start at higher base initial temperatures but which are otherwise similar to G. Higher base initial temperatures cause the "hump" to occur earlier (from Wood, in preparation).

erupted would tend to fan out everywhere and form a uniform layer on the surface, which would not give rise to gravity anomalies. One needs, therefore, a means for localization of the lavas. In spite of the Moon's being in gravitational equilibrium everywhere, some parts must be lower than others to serve as traps for the lava. But we know this to be the case with the maria: The highlands, being of less dense material, stand up higher than do the maria. Thus, even without gravity anomalies, we can have topographic irregularities in that the maria are substantially lower than the highlands. Lava that flowed onto the surface would localize itself in the low places (the maria), and if the region was strong enough to support the additional mass, a mascon would be produced.

Figure 43 shows how different processes on the lunar surface would produce a terrain with many different levels, which could then be overfilled by lava to greater or lesser thicknesses and thus could cause greater or lesser gravity anomalies. In Stage I of Figure 43, a pristine lunar surface is shown which has just differentiated according to the model discussed earlier: Plagioclase has floated up to form the

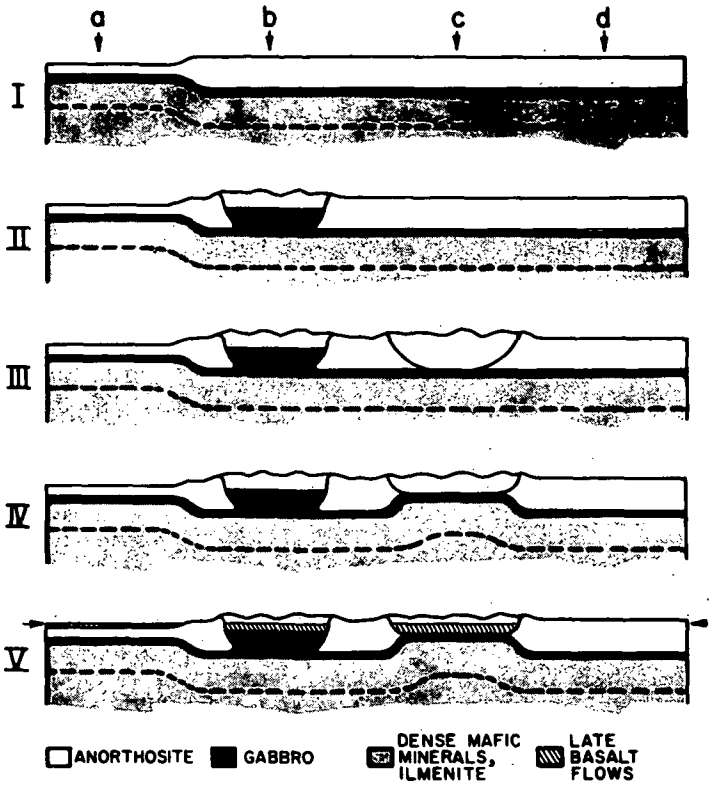


Figure 43.–Crustal model of the Moon showing stages of formation of irregular but compensated topographic features: I: Isostatically compensated crust, product of crystal fractionation in surface magma system, in early times. Anorthositic layer is of nonuniform thickness. Gabbroic layer is still liquid. II: Major crater (mare basin) is excavated, fills to equilibrium level with gabbroic magma. All gabbroic liquid then solidifies; hence, mare basin excavated later (III) is not filled with magma, but (IV) by a plug of solid mantle material. Topography at stage IV is isostatically adjusted; floor of mare *c* stands lower than *b* because mantle material underlying *c* is denser than basalt beneath *b*. V: Episode of lava generation extrudes basalt to level shown by arrows. In some terrains this does not bring lava to the surface (*d*, highlands), but elsewhere lava flows are deposited. These may be very thin (*a*, Procellarum?), of moderate thickness (*b*, Tranquillitatis?), or greater thickness (*c*, Imbrium?). Areas receiving the thickest deposits at this stage will display the greatest positive gravitational anomalies.

white layer, diopside has sunk to form the gray layer, and there is still some leftover basaltic liquid in between them (represented by the black band). The process is shown not to have operated uniformly over the entire surface, so the anorthosite layer is thicker in some places than in others.

Stage II shows a large, mare-size crater punching through the anorthosite layer at a time when the underlying basalt is still liquid. The basalt thus fills the crater (i.e., mare) and comes to a level in hydrostatic equilibrium. Stage III shows another mare-forming impact that happened after the basalt had hardened everywhere, so that nothing flowed up into the basin. However, a basin on that scale is so badly out of gravitational equilibrium that the material from depth in the Moon would tend to squeeze up into it via solid-state creep or flow, without actually needing to be melted. This is shown in Stage IV. The author proposes that the surface shown in Stage IV is completely compensated, i.e., is in gravitational equilibrium. The level under *d* is the highest, that under *a* the second highest, that under *b* the third highest, and that under *c* the lowest. Thus, four different topographic levels would exist on the Moon, all in isostatic equilibrium with one another, with no gravity anomalies.

Finally, assume a late resurgence of igneous activity, about 1 billion years after the formation of the Moon, which sends basalt out onto the surface of the Moon and into the low places; suppose it can rise no higher than the arrow shown in Stage V. None gets onto the highlands, although some might get onto an area such as *a*, where the crustal material is very thin. A much thicker layer would fill a mare like that beneath *b*, and still more would fill the deeper maria. Once this filling has occurred, a gravitationally disequilibrated situation would exist. If the highlands region on the right in the above figure were used as the base, a lunar orbiter would sense a strong gravity anomaly over the mare with the thickest accumulation, a weaker anomaly over the other mare, and a very feeble one over the other highland region (i.e., beneath *a*). Two types of mare craters are proposed here in an attempt to explain why some maria have mascons while others do not. The craters that do have mascons tend to be the latest ones; perhaps they were formed at a time when there was no longer any liquid lava to come up and fill them, so they were filled instead by squeezed-up mantle material of higher density. This material stood at a lower topographic level and, therefore, more of the later-generated basalt could flow over it.

REFERENCES

Gault, D. E., and Heitowit, E. D., "The Partition of Energy for Hypervelocity Impact Craters Formed in Rock", paper presented at the 6th Hypervelocity Impact Symposium, Cleveland, Ohio, 1963.

Gault, D. E., "Scaling Relationships for Microscale to Megascale Impact Craters", paper presented at the 7th Hypervelocity Impact Symposium, Tampa, Florida, 1964.

Gault, D. E., Quaide, W. L., and Oberbeck, V. R., "Impact Cratering Mechanics and Structures", in *Shock Metamorphism of Natural Materials*, B. M. French and N. M. Short, eds., Mono Book Corp., Baltimore, Maryland, 1968.

Gault, D. E., *Radio Sci.* 5:273, 1970.

Muller, P. M., and Sjogren, W. L., *Science* 161:680, 1968.

Oberbeck, V. R., and Quaide, W. L., *J. Geophys. Res.* 72:4697, 1967.

Oberbeck, V. R., and Quaide, W. L., *Icarus* 9:446, 1968.

Osborn, E. F., *Amer. J. Sci.* 240:758, 1942.

Quaide, W. L., Gault, D. E., and Schmidt, R. A., *Ann. N.Y. Acad. Sci.* 123:563, 1965.

Quaide, W. L., and Oberbeck, V. R., *J. Geophys. Res.* 73:5247, 1968.

Runcorn, S. K., and Shrubsall, M.H., *Phys. Earth Planet. Interiors* 1:317, 1968.

Salisbury, J. W., Vincent, R. K., Logan, L. M., and Hunt, G. R., *J. Geophys. Res.* 75:2671, 1970.

Shoemaker, E. M., Hait, M. H., Swann, G. A., Schleicher, E. L., Schaber, G. G., Sutton, R. L., Dahlem, D. H., Goddard, E. N., and Waters, A. C., *Proc. Apollo 11 Lunar Sci. Conf., Geochim. Cosmochim. Acta* 34, Suppl. 1, 2399, 1970.

Wood, J. A., *J. Geophys. Res.* 75:6497, 1970.

Wood, J. A., Marvin, U. B., Powell, B. N., and Dickey, J. S., Jr., Smithsonian Astrophysical Observatory Special Report 307, 1970a, 99 pp.

Wood, J. A., Dickey, J. S., Jr., Marvin, U. B., and Powell, B. N., *Proc. Apollo 11 Lunar Sci. Conf., Geochim. Cosmochim. Acta* 34, Suppl. 1, 965, 1970b.

Page intentionally left blank

CHAPTER 11
ORIGIN OF THE SOLAR SYSTEM

E. Schatzman
Institut d'Astrophysique
Paris, France

I. INTRODUCTION

In astrophysics, the problem of the origin of the solar system is a special one in the sense that what we can observe now are only the results of processes that occurred many years ago, not the processes themselves. So, we must be particularly cautious about the methods we use to build a theory of the origin of the solar system. On one hand, we can rely on the observed facts about the solar system, and on the other, we can rely on observed facts concerning stars, groups of stars, clouds of interstellar matter, and so forth. As yet, we have not observed in the sky the actual formation of a planetary system. Until that has happened—and it may, after all—there will always remain in any theory a certain amount of speculation about the origin of the solar system. Speculation, however, should be avoided whenever possible. This means we must use the least possible number of assumptions, and starting from these we must proceed in the most coherent manner to the conclusion. We must also try to establish whether in some cases these assumptions are self-contradictory. Sometimes, we might find that the assumptions do not lead to the expected result; we must then resist the temptation to introduce *ad hoc* new hypotheses that have only the purpose of leading to the desired results but which, in a final analysis, are not justified either by physics itself or by our present knowledge of stars and interstellar matter.

II. HISTORICAL BACKGROUND

We shall begin our historical survey with the era of Descartes and Buffon. Within their lifetimes (the 17th and 18th centuries), two theories that would recur often in later years were formulated: the monist theory and the dualistic theory.

The monist theory of Descartes hypothesized a cloud, rotating around a central body, from which the planets formed. The main problem with this theory was that it had been formulated before Newton postulated his theory of gravitation. Descartes explained the quasi-circular orbital motion of the planets around the Sun in terms of vortices of the ether; but after the acceptance of Newton's theory of planetary motion (which was based on gravitational attraction), Descartes' theory was abandoned and, in fact, forgotten.

About one century later, Buffon proposed his dualistic theory, which stated that the planets might have originated following the collision of a comet with the Sun. Indeed, the idea of a collision between the Sun and something else, as a mechanism of ejecting from the Sun matter that would subsequently condense into planets, was very interesting. The reason the theory was initially based upon a collision between the Sun and a comet is that Buffon had no idea whatsoever as to the mass of a comet; as we now know, this mass is many times too small to affect the Sun.

These, we may say, are the most archaic theories of the Scientific Era. They were followed by the theories of Kant (in 1766) and Laplace (in 1796), which returned to the concept of a primordial nebula, that is, a disk of some material rotating around a central body. Later, because of their similarity, the combination of these theories was referred to as the Kant-Laplace theory. There are, nevertheless, some differences worth mentioning. In Kant's theory, there is a cloud composed essentially of dust particles which collide with each other until the dust particle system becomes a flat disk in circular motion around the central body. In Laplace's theory, there is a lens-shaped central body which contracts while maintaining its shape to form the proto-Sun. The planets are formed from gas expelled along the equator of the contracting central body. Neither Kant nor Laplace made a mathematical analysis of the problem; Laplace, who had done remarkable work in the field of mechanics, offered, without any mathematical development, only a qualitative discussion of his model.

This was the situation at the end of the 18th century. During the great development of the physical sciences in the 19th century, the problem continued to be discussed until, near the middle of that century, Laplace's theory received a very serious blow in the form of a very simple argument concerning a question of angular momentum.

If the Sun was formed by contraction, its speed of rotation should have increased until, at its present size, it should be rotating with a period of about 2 hours, that is, with an equatorial velocity of 350 $km\text{-}s^{-1}$. However, the equatorial

velocity of the Sun is only 1.4 km-s^{-1} (at least on the surface). Given this velocity, we find that only 3 percent of the total angular momentum in the solar system would be due to the Sun, whereas 97 percent would be due to Jupiter and the other planets. If the Sun were to rotate 200 times faster (as Laplace had predicted), this ratio would be almost reversed; i.e., 60 or 70 percent of the angular momentum would be due to the Sun, and the rest to the planets. In Laplace's time, there was no way to explain why the Sun rotated slowly or to understand how the angular momentum could have been exchanged in order to produce such a slow rotation.

The theory of Laplace was abandoned in the last half of the 19th century. About 1900, Jeans revived the theory of Buffon; this time, however, instead of thinking in terms of a comet (a very inefficient mechanism), it was proposed that the passage of a star sufficiently close to the Sun could have produced a gigantic tide, which in turn could have caused a certain amount of matter to separate from the Sun and to remain in space, eventually to condense and form the planets. From 1900 to 1935, the dualistic theory had many developments, until H. N. Russell gave it the final blow. Russell's argument is very simple and is based on two physical properties. The first is, again, the problem of the angular momentum, and the second concerns the internal structure of the Sun. The Sun is a highly concentrated body, and the tidal effects on its surface, due to the passage of a star at any conceivable distance, would be very small. Only if the star almost touched the Sun would the tidal effect be large enough that, possibly, a "tongue" of matter might be ejected from the Sun and left behind; but in such a case, the angular momentum we find in the planets would not have been imparted to the filament, which would have been ejected at a small distance from the Sun and left with a very small angular momentum.

Therefore, we are left with two conflicting notions: either the passage of the star was sufficiently distant to provide the correct angular momentum properties (but then the tidal effects would be only a small perturbation on the surface of the Sun, and no ejection of matter would have taken place) or the star passed close enough to cause the ejection of matter (but then the angular momentum imparted to the ejected matter would be too small).

In order to overcome this contradiction, Lyttleton proposed a more complicated explanation which included a double-star system hit by a third star. The main star, the Sun, and the secondary star would be separated by a distance of the order of the orbit of Jupiter; then, the passage of a third star could cause an ejection of matter which might be distributed at the proper distance with the correct angular momentum.

In 1938, Spitzer supplied a criticism related to the internal structure of the Sun. His argument is very simple. A filament of matter from which the planetary system might have been constructed must have had a mass of about $10^{-2}\ M_{\odot}$. The total mass ejected from the Sun would have to have been several times larger than this to account for several mass losses that could have taken place. For this mass to have been obtained from the Sun, it must have originated below the solar surface in a region where the density would be high enough to provide the required amount of matter. However, when matter is removed from the Sun, the gas pressure within the matter is no longer balanced by the gravitational attraction of the Sun; that is, as soon as the filament is ejected, it is submitted to only its own gravitational potential, which is many times smaller than that of the Sun. The lifetime of such an object is essentially the time it takes an acoustical wave to travel along its dimensions, or about 1 minute, as calculated by Spitzer. Then, the filament explodes and the matter is spread out into space, which returns us to the primitive-nebula theory. Thus, it is not really necessary to go through all these troubles in order to develop a theory of the origin of the solar system.

This is certainly the reason why, since 1945, the only theory that has been developed is that of the primitive nebula, which we shall call the Kant-Laplace theory. We wish to analyze the development of this theory, but first, the facts that have to be interpreted will be stressed.

III. FACTS ON THE SOLAR SYSTEM

A. Mechanical

The mechanical facts of the solar system are fairly well known; they refer mainly to the orbits of the planets. What remains to be explained is why the planets rotate in quasi-circular orbits around the Sun, why these orbits lie in almost the same plane, and, moreover, why there is a law of the planetary distances, that is to say, why the planets do not seem to be randomly distributed around the Sun.

The fact that almost all the planets rotate in the same direction was already mentioned by Laplace as a reason for adopting the planetary-nebula theory because a very simple statistical argument shows that it is very improbable that the planets could have this motion just by chance. The law of planetary distances, discovered in

the 18th century, is usually known as the Bode-Titius law; it expresses the semimajor axis of the elliptical orbits as a geometrical series:

$$\sqrt{r} = a + b^n ,$$

where n is the index number of the planet. At the time when it was formulated, only the planets from Mercury to Saturn were known. When Herschel discovered Uranus, its orbit fitted the Bode-Titius law adequately, and the belief that the planets followed the law guided Leverrier in his search for Neptune. It is impressive that even in the case of Neptune, the departure from the law is not too large. The same idea was tried again for the discovery of Pluto, but this time it did not work very well; Pluto does not obey the Bode-Titius law. Still, the overall situation is remarkable, and a great effort has been devoted to the explanation of the law of planetary distances.

In relation to this, the question of the mechanical stability of the solar system should be raised. This is a very important point because if we are going to try to explain the Bode-Titius law as a cosmogonical fact, we have to assume also that the law of planetary distances has not changed appreciably since the formation of the solar system. This means that we have to assume the mechanical stability of the orbits. They are probably stable for a time of the order of 5×10^9 years; as far as the author knows, there is presently no proof, or disproof, that the orbital motions are stable for a longer period of time. This is the first weak point in any cosmogonical theory. Everyone assumes, in fact, that the Bode-Titius law reflects the conditions at the time of formation of the planets, although it has not been proved yet that the situation remained stable for 5 billion years, which is the age of the solar system.

B. Chemical

Discussion of the chemical makeup of the solar system started quite late. In fact, until a careful analysis of the structure of the planets was made, no thought was given to the question of chemical composition.

The chemical facts fall essentially in two categories. One concerns what we may call the gross chemical properties, which can be simplified by the following

considerations. The telluric planets (Mercury, Venus, Earth, and Mars) are made of rocks; with regard to their gross chemical properties, questions of fractionation in the formation of these planets and the possible differences in composition between each of them do not arise. The giant planets may be divided into two groups. To one belong Jupiter and Saturn, which appear to be almost like the Sun; i.e., the calculation of the radius of a cold sphere composed of hydrogen and helium in relative solar abundances and with the mass of Jupiter or Saturn yields practically the radius of these planets. This means that there is a large difference in chemical composition between the telluric and the giant planets, which is one of the main questions as far as the chemical properties are concerned. To the other group of giant planets belong Uranus and Neptune, which show a hydrogen and helium deficiency compared with Jupiter and Saturn. In other words, given the masses of Uranus and Neptune and a composition of only hydrogen and helium, the calculated radii are definitely greater than those observed.

The second category of chemical facts concerns the detailed chemical analysis of the planets and other objects in the solar system. This raises the question of the composition of the Earth, the Moon, and the meteorites, which show a large variety in chemical properties. There is also, especially in the case of the Earth, possibly in that of the Moon, and certainly in that of the meteorites, the question of deviation from the so-called cosmical composition. This composition is derived from those of the solar atmosphere, some stellar atmospheres that have been studied in great detail, and, for heavier elements, meteorites, which are supposed to be good representatives of the chemical composition of the primitive nebula. There is no evidence that the present composition of the solar atmosphere is entirely the same as that of the primitive nebula, but we believe it is a good first approximation. There are, then, these differences and, moreover, the questions of fractionation and of how it was possible for certain elements to disappear or, at least, considerably decrease in abundance on the surface of the Earth.

C. Mineralogical

The meteorites show a variety of crystalline forms, and the physical conditions under which these crystals have been formed are very restricted. This limits greatly the physical conditions of the primitive nebula, of the bodies in which the meteorites were formed (primary bodies), and of the secondary bodies.

D. Isotopic (Nuclear Composition)

Isotopic data are very recent. They are obtained by extensive mass-spectrographic studies of meteorites, the Earth's crust, and, more recently, Moon rocks. We are faced with a number of remarkable problems: There is a problem of the differences in abundances of D, Li^6, Li^7, Be^9, B^{10}, and B^{11}. Then there is the problem of the heavier isotopes, such as Xe, and in this connection, also that of the melting of a body like the Moon. This in turn raises the question of the formation of Al^{26} (radioactive) that produces Mg^{26}. Finally, there is the problem of all the radioactive elements, from K^{40} to Th.

Once we accept that all these facts must be explained, we find that our freedom is extremely limited. It is within these limits of freedom that we must develop a theory of the origin of the solar system.

IV. RECENT DEVELOPMENTS

The most recent developments are related to a revival of the Kant-Laplace theory and have two different origins. One is the work of von Weizsacker; the other is that of O. Yu Schmidt in the U.S.S.R. Their theories are based, essentially, on the idea of the primitive nebula, that is to say, of a rotating disk revolving around the Sun, but some aspects of this picture, the main one being perhaps the question of the origin of the rotating disk itself, are not really discussed by them. Schmidt suggests that the disk was captured by the Sun, but this is not an essential point in his work. We shall give a brief criticism of these two theories.

Von Weizsacker's theory begins with the Sun and a disk of gas rotating around it. We assume the presence of dust particles in the gas which follow elliptical orbits of the same period as those for circular motion around the Sun (if one neglects friction between dust particles and the gas). Now, if we take a frame of reference that rotates with the Keplerian angular velocity of the circular motion and study orbits of dust particles having, in this frame, a different eccentricity but the same period as the circular motion, we obtain elongated orbits that are closed. Dust particles following these orbits would form vortices along the circular-motion orbit. Thus, with such a rotating system, a finite number of vortices rotating around the Sun can be formed. Consideration of five of these vortices yields a law of planetary distances,

$$r = ca^n ,$$

which reproduces adequately the law of planetary distances from Mercury up to Uranus. It is only a kinematic picture in the sense that the vortices are formed by the orbits of the dust particles and not by some kind of instability or turbulent motion.

The theory of von Weizsacker depends completely on the existence of these vortices, but there is no proof that the vortices do form and are actually stable. This is why in the years that followed, a number of people have discussed the origin of the vortices. In particular, ter Haar and Kuiper tried to abandon the naive kinematical picture of stable vortices in order to introduce a concept borrowed from the theory of turbulence: What actually took place was a turbulent motion in which the vortices formed as eddies.

The problem with the picture of turbulent motion is that it has been proven that a rotating fluid whose angular momentum increases outwards is a system that is stable against turbulent motions. This had been shown already by Rayleigh for the case of a fluid rotating between two cylinders, and the same has been shown to be the case for rotating stars also. If in a rotating star the angular momentum increases outwards, the star is stable against perturbations.

In conclusion, on one hand, it seems improbable that dust particles embedded in a gas would develop the type of motion required by von Weizsacker's theory; on the other hand, the fact that the angular momentum increases outwards in the disk appears to prevent the formation of eddies by turbulent motions.

The theory of O. Yu Schmidt is quite different. According to it, the planets are formed by accretion of matter around an initial center, which can be quite small. Once accretion begins in a certain place within the primitive nebula it continues by taking matter from the nebula. As a center grows, it prevents the next one from growing. Hence, within the nebula a certain "area of accretion" is defined for each center, which is determined by the angular-momentum distribution. This argument leads to a law of planetary distances given by

$$\sqrt{r} = a + bn \ ,$$

n being the index number of the planet. If we look at a plot of planetary number versus distance (Figure 1), it turns out that the law that fits the data better is given by the expression above (if one uses two different sets of parameters a and b, one for the telluric and one for the giant planets).

The main criticisms of Schmidt's theory are that the region of accretion has, in fact, no boundaries and that the method by which boundaries are presently

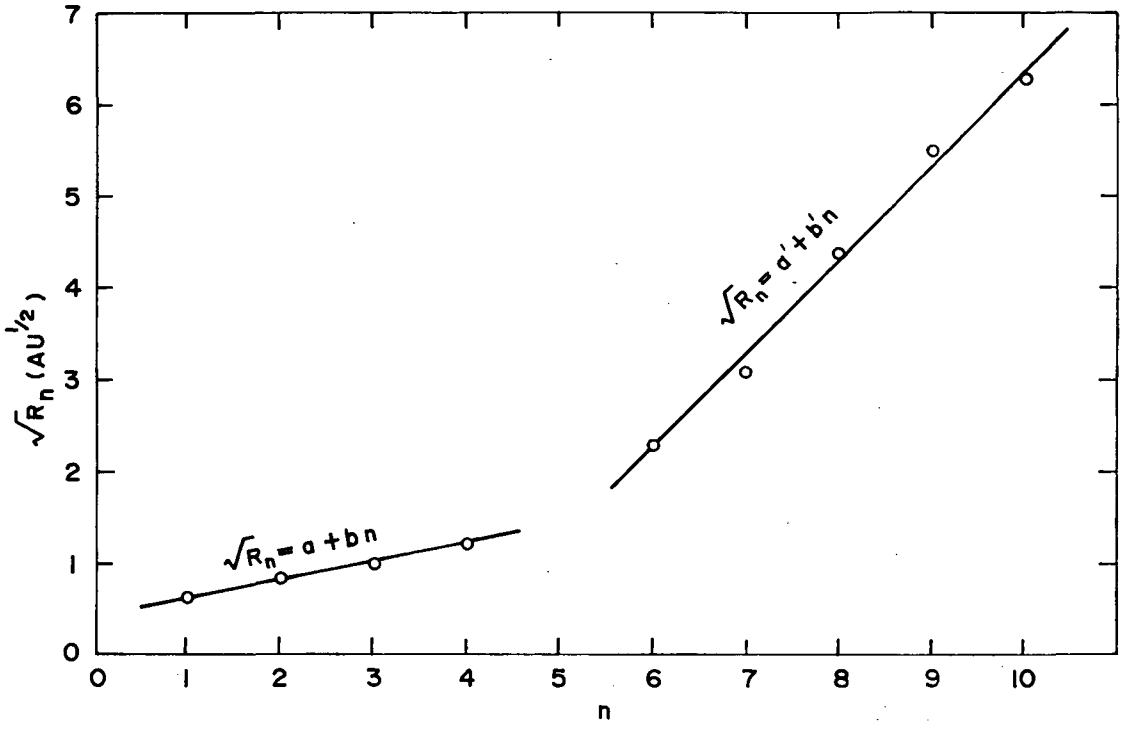


Figure 1.—Planetary distances calculated from the expression $\sqrt{R_n} = a + bn$ using two values for a and b for each of the two groups of planets. The actual distances are also plotted (Schmidt, 1959).

determined gives rise to an overly simplistic model. We will return later to the Schmidt model to discuss it in more detail.

V. THE PROTO-SUN

A. Young Stars and Early Phases of Evolution

As we have stated previously, *ad hoc* assumptions should be avoided as much as possible. Consequently, we must look at the sky to find some further information as to the conditions which might have accompanied the formation of the solar system.

This information includes what we learn from the observation of young stars and what we know about the early stages of stellar evolution. This chapter, of course, does not present a complete review of these subjects; it is only a survey of the questions relevant to the origin of the solar system.

Our information on young stars comes from the observation of extremely young clusters. There are two interesting examples: NGC 2264 (studied by Walker) and the Orion cluster, associated with the very well known Orion nebula, which has been studied extensively by many authors (particularly Haro, Herbig, and Poveda).

What one observes in the Hertzsprung-Russell diagram, of NGC 2264 (Figure 2) for example, is that the blue stars are already leaving the main sequence and the late stars are largely in a region above it.

An estimate of the time scale of evolution by contraction gives an age of the system which is similar in the case of stars that have left the main sequence and of those that have not yet reached it, about 2.4×10^6 years. This is very interesting because the first million years must have been very important in the formation of the planets, as we shall discuss later. Observations of young stars give us an idea of how the Sun might have been at the time of the formation of the solar system.

It is important to notice that the stars that fall in the area above the main sequence (late-type stars) are variable stars, showing either flare activity (visible in the case of the very late spectral types, K or M) or emission spectral lines, especially H_α. In the case of the Orion cluster, whose estimated age is 0.5×10^6 to 1×10^6 years, the stars also show features similar to those of T-Tauri stars (i.e., wide fluctuations in light and broad emission lines due to a circumstellar envelope). In the Orion nebula, a small number of these variable stars lie below the main sequence. The spectroscopic features and the variability characteristics of these stars are similar to those appearing above the main sequence, yet they appear below. A possible explanation for this was given by Poveda, who suggested that they are seen below the main sequence because of absorption in a circumstellar cloud of dust. Poveda claims that the number of stars observed below the main sequence is consistent with the probability of observing a star that is surrounded by a disk of dust, if one takes a random distribution for the position of the absorbing disk. Hence, there is at least a possibility that whenever we look at the stars through the disk, they may appear fainter, thus falling below the main sequence. Certainly, this favors the idea that these stars are perhaps building some kind of planetary system.

We now turn to the question of the type of activity observed, a question raised at the IAU Symposium held in Paris in 1958. At that time, the author suggested that one could observe flare stars by looking for a radio emission of the same type as

that which takes place on the surface of the Sun during a solar flare. (The only data available then were those coming from the observation of the Sun.) The Jodrell Bank group (Lovell) began a systematic survey of about six flare stars; naturally, the nearest and best known were chosen. A most interesting case is that of UV Ceti, which has been studied extensively by Oskanian. This star shows, as a function time,

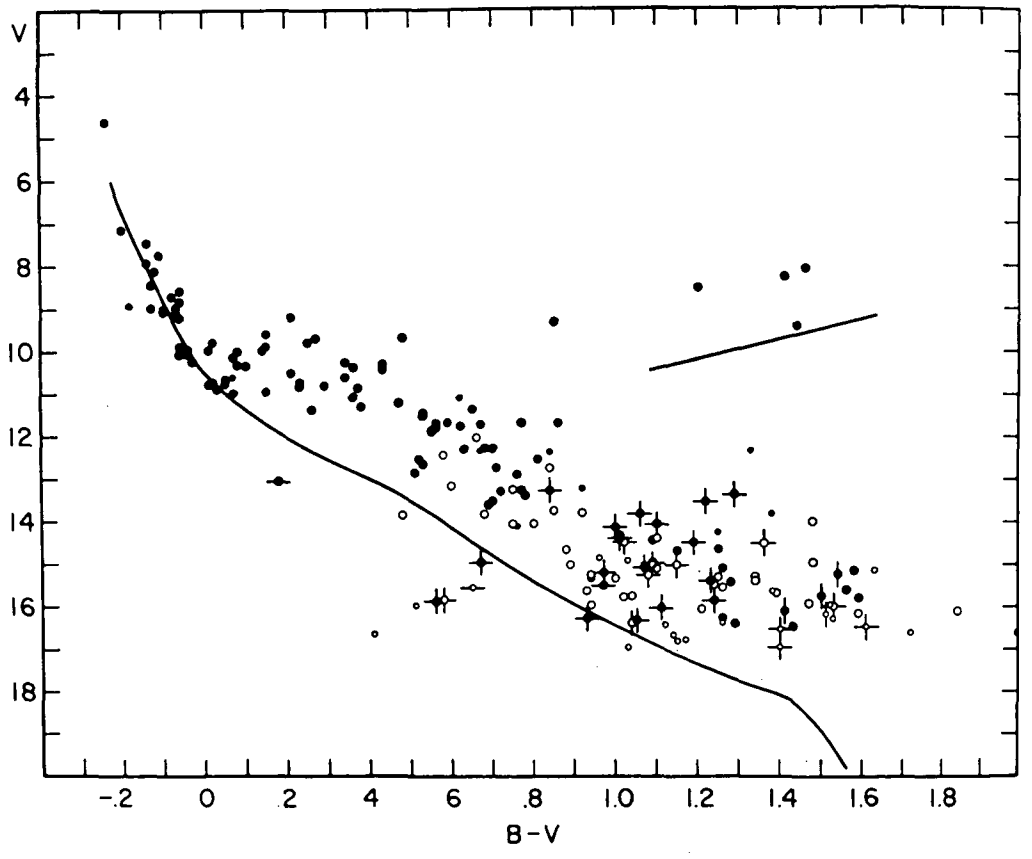


Figure 2.–Color-magnitude diagram of NGC 2264. Dots represent photoelectric, and circles photographic, observations. Vertical lines indicate known light-variables; horizontal lines indicate stars having bright H_α. Smaller symbols indicate observations of lower weight. Observed values of the magnitudes and colors have been plotted. The lines represent the standard main sequence and giant branch of Johnson and Morgan, corrected for the uniform reddening of the cluster (Walker, 1955-56).

very sharp increases in optical luminosity (up to 6 magnitude) with a duration of 10 to 30 minutes (Figure 3). While observing UV Ceti, Lovell's group discovered a radio burst associated with the optical peak.

Observations showed that for several flare stars, it was possible to detect radio bursts of the types I, II, and III. The reason why this was possible is that the ratio of the energy available in the radio range to the energy available in the optical range is much higher for flare stars than for the Sun. That is, if we denote the energy available by W, then for the Sun,

$$\frac{W_{\text{radio}}}{W_{\text{optical}}} \approx 10^{-5},$$

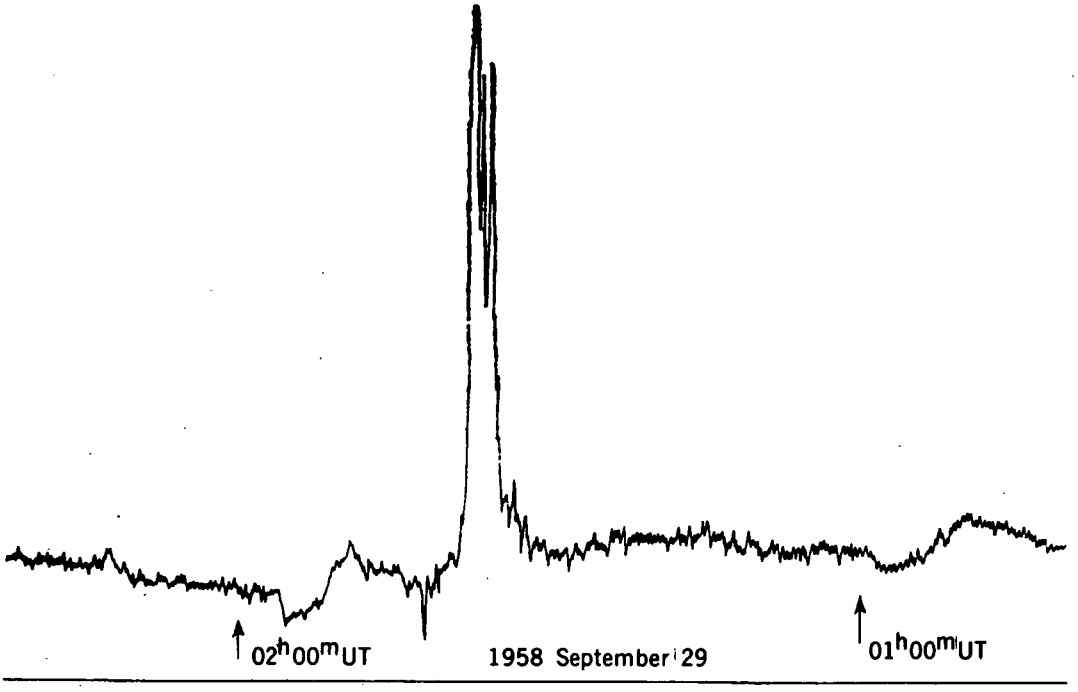


Figure 3.–The record of the radio emission obtained between 01^h and 02^h UT on September 29, 1958, with the radio telescope following the flare star UV Ceti. Time increases from right to left; radio intensity increases vertically (Lovell, 1964).

whereas for a flare star,

$$\frac{W_{\text{radio}}}{W_{\text{optical}}} \approx 10^{-2} \text{ to } 10^{-3}.$$

Because of some unknown reason, the efficiency of the flaring mechanism in these stars is 100 to 1000 times higher than that in the Sun.

Cosmic-ray activity can be discussed in connection with radio bursts. We have direct information about such activity in the case of the Sun only. At the time of a solar flare, a flux of cosmic rays is observed. For example, in the case of a solar flare with an energy in the optical range of 10^{31} ergs, the energy output for cosmic rays of more than 30 MeV is also of the order of 10^{31} ergs; that is, we have an efficient mechanism for the production of cosmic rays with an energy output comparable to that in the optical range. This corresponds to a total of 10^{35} particles with an average energy of, say, 100 MeV. (See Figure 4.)

We are interested in young stars because we assume that the proto-Sun of the young Sun was similar to a flare star. There is no direct evidence for the production of cosmic-ray particles by young stars. However, we do have indirect evidence, and we will analyze it starting with the most acceptable and ending with the most speculative.

(1) First, we find indirect evidence in the so-called Herbig-Haro objects, very faint nebulae observed by Herbig and Haro in the Orion nebula. They have a small diameter of about 100 AU and a spectrum; thus, Haro was able to classify them in a continuous sequence with the flare stars, the T-Tauri stars, and the emission-line stars, with a slow modification of the spectroscopic features in passing from one type to the next.

The Herbig-Haro objects have spectroscopic features that correspond to both relatively high energy and low temperature. That is, they show both Ca-II lines and the O-III forbidden lines. The degree of ionization is small, i.e.,

$$\frac{H^+}{H + H^+} \lesssim \frac{1}{10}.$$

It is possible to explain their spectra if one assumes that the nebulae are excited by a stream of moderately high-energy particles (1-MeV protons can account for the process very well). The stream of particles will ionize the hydrogen, and an

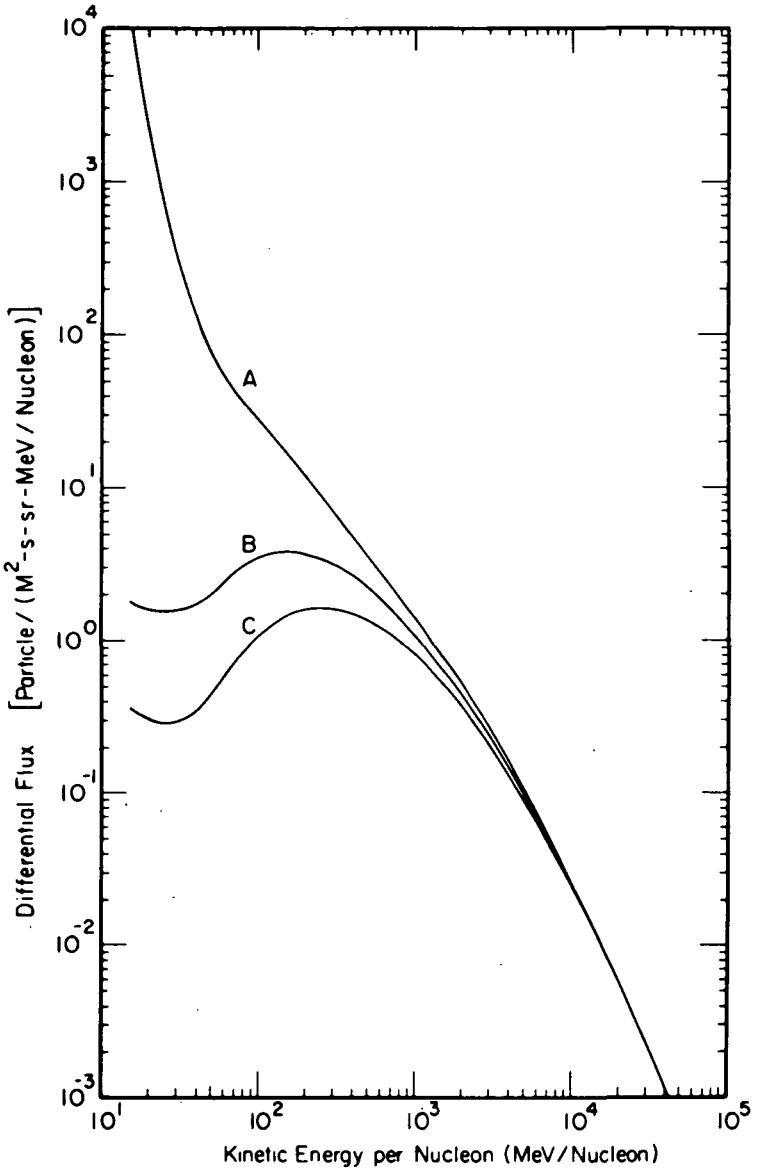


Figure 4.—Demodulated differential energy per nucleon spectrum for protons and α-particles (curve A). The demodulated energy spectra are identical as a function of energy per nucleon if the helium spectrum is multiplied by 11. Curve C represents the proton spectrum, and curve B the helium spectrum multiplied by 11 (Gloeckler and Jokipii, 1967).

equilibrium will be established between the number of ionizations and the number of recombinations. The secondary electrons produced by the ionization process have an average energy of about 30 to 40 MeV. These high-energy electrons are in sufficient number to produce the O-III line before being thermalized. Only when they are thermalized can they be used to produce the Ca-II line. Thus, such a spectrum with both high- and low-temperature features can in fact be explained by assuming that the exciting mechanism is not photons but high-energy particles. We may consider this to be the first reason for believing that there are cosmic-ray particles in the Herbig-Haro objects.

(2) The second evidence is provided by the type-III bursts. In order to produce these bursts, electrons with a velocity of the order of $1/2\ c$ must be present (which already means electrons of 0.5 MeV). Hence, type-III bursts are also associated with relatively high-energy particles.

(3) The third indirect evidence is related to the low-energy cosmic rays observed from satellites. The problem is to demodulate the effect of the solar wind, but if the methods for doing this are reliable, we must assume that the low-energy ($\approx$10-MeV) cosmic-ray flux is very high. The form of the demodulated spectrum can be explained by the fact that low-energy cosmic rays have a very short mean free path in the galaxy (100 pc); thus, they must be produced by nearby objects. They cannot be produced by pulsars, for example, because the number density of pulsars is too small and the nearest pulsar is too far away. However, if we assume that the flare stars are producing these low-energy cosmic rays (just as the Sun is), we find that the number density of nearby flare stars is high enough to explain the demodulated spectrum. So, this too is indirect evidence that the flare stars emit cosmic rays. Moreover, connected with the demodulation problem is the explanation of the X-ray spectrum, which also implies a fairly high flux of low-energy cosmic rays. On the basis of these facts, we shall assume that the proto-Sun, at an early stage of its formation, was like a flare star, a very powerful emitter of cosmic rays. After having considered the amount of energy that goes into the flares and the efficiency of production of radio waves in flare stars, we assume also that the flux of cosmic rays corresponds to an energy output of the order of one one-hundredth to one-tenth of the stellar luminosity. Naturally, this estimate is quite uncertain, but the best way to approach this value is to consider the flux of low-energy cosmic rays.

(4) Finally, there is indirect evidence connected with the abundance of lithium and beryllium on the surface of very young stars, but the question at this time is highly controversial. The possibility that these elements may form on the stellar

surface from carbon and oxygen by spallation due to stellar cosmic rays has been discussed at great length. At present it is argued whether lithium could have been produced in the galaxy by cosmic rays. The problem is not settled, so we will not discuss it.

B. Stellar Rotation

We turn now to stellar rotation. Figure 5 shows the equatorial velocity for field main-sequence stars as a function of their absolute magnitude or spectral type. The equatorial velocity drops very fast and becomes extremely small, less than can be measured, near spectral type F2 (absolute magnitude 3.8). The graph is based on a relatively old interpretation of the broadening of spectral lines. On the basis of the work of Hardorp in the last 2 years, the quantity $V \sin i$ has to be reinterpreted;

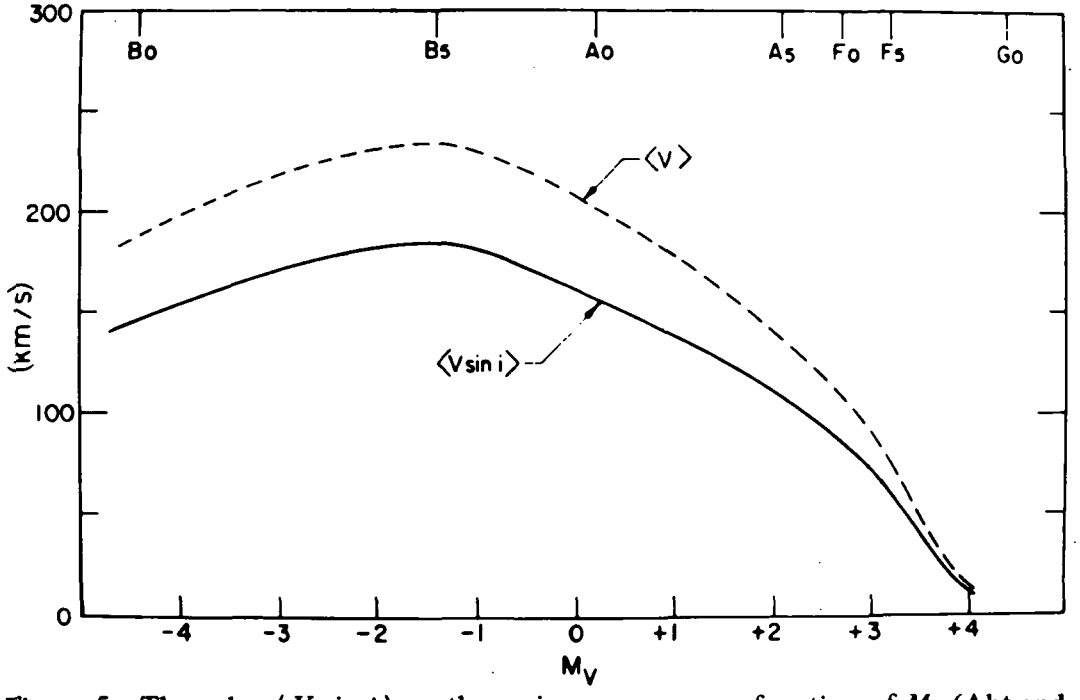


Figure 5.—The value $\langle V \sin i \rangle$ on the main sequence as a function of M_v (Abt and Hunter) (Kraft, 1969).

for A and B stars, it is larger than is shown in the figure. The limit of gravitational instability at the equator of these stars, due to the rotation, is about 350 $km \cdot s^{-1}$, and the reinterpretation shows that the velocities are actually close to the limit.

The question of how to interpret this distribution of equatorial velocity was raised years ago by Struve, and the explanation is probably related to the loss of angular momentum associated with the existence of a magnetic field at the surface of a star. Given a rotating star, there are two possible mechanisms for the loss of angular momentum:

(1) The lines of force associated with a center of activity (i.e., with a magnetic spot) can guide the motion of particles that would otherwise become lost at a distance from the surface of the Sun much larger than its radius; therefore, the efficiency in the loss of angular momentum is multiplied by $(a/R)^2$, with a being the distance at which the loss occurs and R being the radius. Thus, even a small mass loss can account for a large loss in angular momentum. For example, to bring a star like the Sun from the limit of stability (with a period of about 2 hours) to the present situation (a period of 27 days), there needs to be a mass loss of only $0.003M_{\odot}$ if $a/R \approx 10$. This, then, is a very efficient mechanism.

(2) The second possible mechanism is associated with the overall magnetic field of a star. Again, the matter is guided by the lines of force; but the mass loss, instead of being linked with flares, is linked with the stellar wind.

The efficiency of angular-momentum loss due to either of these mechanisms is related to the strength of the magnetic field and to the amount of matter that is being ejected. A point discussed extensively by O. Wilson and R. Kraft is the relation among age, equatorial velocity of rotation, and index of stellar activity. Wilson discovered that the intensity of the K-line of Ca II is an index of the amount of electromagnetic activity present at the surface of a star. He has been able to show that there exists a correlation among age, rotation, and this index. A young star seems to have a strong electromagnetic activity and loses a great deal of angular momentum. When it becomes older it rotates more slowly and has less electromagnetic activity. This fits very well with the kind of theory one can build concerning the origin of electromagnetic activity at the surface of a star.

We would like to stress a point that is very important in both the stellar-wind and the spot-activity models. In order to have loss of mass and angular momentum, there must be either electromagnetic activity or a corona which expands and carries matter and angular momentum. Both these phenomena are related to the existence of a deep hydrogen convective zone. A convective zone produces or regenerates the magnetic field and is very likely to be related to the origin of the magnetic spots; in

addition, through the production of acoustic waves, it provides a heating mechanism that generates the corona. So, whatever the mechanism of loss of angular momentum, the existence of the convective zone is fundamental.

If we look at stellar models (Baker and Temesvary), we see that stars of a spectral type later than F2 have a deep convective zone, whereas stars of a type earlier than F2 have either a very shallow convective zone or none at all. This explains very well why we have a separation between fast rotators and slow rotators at the spectral type F2. After this type, there exists an efficient mechanism for the loss of angular momentum, whereas before, the mechanism is not present.

The Sun is not a special case and fits adequately into the above picture, which solves the problem of a fast rotating Sun, and we can fully accept the idea of simultaneous formation of the Sun and the primitive nebula, according to the scheme of Laplace, as a matter of fact. The slowing down of the Sun took place later, as a phenomenon independent of the formation of the solar system and with a very different time scale. As we shall see, there are several reasons to believe that the time scale for the formation of the solar system is about 1 million years, whereas the scale for the slowing-down process is many times that period. Thus, we have formed the following picture of the proto-Sun:

(1) It is an object which has not reached the main sequence.

(2) Its radius varies from $40R_\odot$ to $5R_\odot$.

(3) Its luminosity is similar to that expected from the Hayashi track, or, in connection with the problem of infrared stars (as has been studied by Hayashi, Larson, and others), it can be fairly high in this late phase, varying from something like $500L_\odot$ to $2L_\odot$ to $5L_\odot$.

(4) It is an emitter of cosmic rays, with the power emitted either in the form of bursts or continuously, being between $0.1L_\odot$ and $0.01L_\odot$.

(5) It is ejecting matter in the form of stellar wind. The importance of this process is unknown.

(6) Its effective temperature is about 3000 K.

It is with this picture of the Sun at the early phases of solar evolution in mind that we shall try to work out the problem of the structure of the primitive nebula.

VI. THE PRIMITIVE NEBULA

How the primitive nebula was formed has very much to do with the contraction of a rotating body, but the problem does not seem to be fully solved

yet, as the reader shall see from the following discussion. Let us suppose we have a sphere of gas, rotating and contracting. For the present, we assume it has a radius $R \approx 10^4 R_\odot$ (of the order of Pluto's orbit). We do not really know the physical situation inside such a body, nor do we know very well its composition. However, we assume that its average density ρ is much lower than the present density of the Sun; i.e., $\rho \approx 10^{-12}\rho_\odot$, which is a value comparable to chromospheric densities (it is equivalent to a typical value of 10^{12} particle-cm^{-3}).

If such an object is more or less in thermodynamic equilibrium, it must be made essentially of H^2 and possibly dust. If it is optically thick, it can easily be shown that it is in a state of convection; that is, the radiative transport is unable to carry away the energy liberated by the gravitational contraction of the object and, therefore, the energy available has to be transported by convection. If so, this body can be more or less described as a polytrope; it is difficult at this point to state what the index of the polytrope is, but as the reader will see, we can solve that difficulty.

We assume then that we have a convective object with a thin radiative atmosphere. It is rotating, and we start with a velocity of rotation such that at the equator the centrifugal force equals the force of gravity, that is, if ω is the velocity of rotation,

$$\omega^2 R_{eq} = g_{eq} ,$$

g_{eq} being the gravitational acceleration at the equator. This body will rotate with the maximum angular velocity it can have if it is going to be bound by gravitation.

If we let the object contract, the hydrogen molecules will dissociate very fast. A small contraction is enough to raise the central temperature to the point at which dissociation of H^2 takes place. Now, if there exists a dissociation equilibrium

$$H^2 \rightleftharpoons H + H ,$$

the adiabatic compressibility of the gas, defined by

$$\gamma = \frac{d \log P}{d \log \rho} ,$$

decreases and can very easily become smaller than 4/3. This is a very well known case of dynamical instability. It is easy to see why.

The pressure due to gravity at the center of an object is

$$P_c \sim GM^2/R^4 .$$

If the density is

$$\rho \sim M/R^3 ,$$

it follows that the central pressure due to gravity is proportional to the density to the power 4/3; i.e.,

$$P_c \sim \rho^{4/3}$$

in a homologous configuration (one with a uniform distribution of pressure, density, and temperature).

The adiabatic relation in the gas is given by

$$P \sim \rho^{\gamma}$$

(here, P is the gas pressure). If $\gamma > 4/3$, when the sphere contracts the gas pressure increases faster than the pressure due to gravity, so equilibrium is restored, and we have a stable configuration. On the contrary, if $\gamma < 4/3$, a small contraction leads to an increase in pressure which is less than the increase in gravitational pressure, so there is nothing to prevent collapse, and a dynamic instability arises. When dissociation exists, it is not necessary to go very deep inside the sphere in order to reach a point where $\gamma > 4/3$ and the whole system becomes dynamically unstable.

The assumptions we have made and the type of reasoning we have used lead to the conclusions that the sphere of gas will collapse and that the collapse will end only when γ becomes larger than 4/3. This will occur when the H^2 molecules are completely dissociated, which takes place at a later phase of evolution of the system, when the radius has dropped to a value of the order of $100R_\odot$ (between the present orbits of Mercury and Venus). Thus, there would be a violent collapse from $10^4 R_\odot$ to $100R_\odot$, followed by a more stable situation. This collapse would take place in about a hundred years, which is the time scale for a parabolic orbit fall from Pluto to Mercury.

There is another possibility we would like to discuss. We assumed that the object is in convective equilibrium since the heat transfer cannot take place by radiation. If this is true, we can estimate the velocities that account for the energy

transport by convection. In fact, it is possible to write an expression for the convective flux, on the basis of the set of equations established by Ledoux and Walraven. We have to consider the coupled system for the generation of turbulent energy and the transfer of thermal energy.

The discussion shows that for a fast contraction, the main effects are the following: In the generation of the turbulence, it appears that, after the turbulence is established, the main source for maintaining it is the conversion of gravitational energy into turbulent motion. For the heat flux, it appears that the main effect is the conversion of turbulent energy into heat.

Let us call $\beta = -(V/r)$ the rate of homologous contraction, V the velocity of contraction, F the heat flux, and l the scale of the turbulence. We shall later assume that $l/R \approx 0.1$.

The equation of generation of turbulence simplifies to

$$\operatorname{div} \frac{3}{4} \rho V^3 = \frac{4}{3} \rho l V^2 \beta^2 .$$

The equation of heat transport reduces to

$$\operatorname{div} F = s \rho V^3 / l ,$$

where, according to the theory of dissipation of shock waves, the constant s is 2/9 if $\gamma = 5/3$, and 7/36 if $\gamma = 4/3$. The luminosity of the object is given by

$$L = \frac{GM^2}{R} \frac{3\Gamma_{\text{eff}} - 4}{5\Gamma_{\text{eff}} - 6},$$

where Γ_{eff} is the effective compressibility. To give some idea of the consistency of the assumptions, we will assume that the effective compressibility Γ_{eff} is larger than 4/3 and close to 4/3. The representation by a polytrope 3 is acceptable. We then have two ways of estimating Γ_{eff}. We obtain two expressions for the ratio of the turbulent pressure to the total pressure. The first one results from the structure of the object:

$$\frac{P_t}{P_{\text{total}}} \approx 0.12 (l/R)^{1/2} \frac{\int_0^{\xi} \xi^{4/3} \theta^2 \, d\xi}{\xi^{4/3} \theta^2} \frac{3\Gamma_{\text{eff}} - 4}{5\Gamma_{\text{eff}} - 6} = A \left(\frac{3\Gamma_{\text{eff}} - 4}{5\Gamma_{\text{eff}} - 6} \right) .$$

On the other hand, after the effects of the microscopic and macroscopic motion are averaged, we obtain

$$3\Gamma_{\text{eff}} - 4 = \frac{(3\gamma - 4)P_{\text{gas}} + P_{\text{turb}}}{P_{\text{gas}} + P_{\text{turb}}},$$

from which we derive

$$\frac{P_t}{P_{\text{total}}} = \frac{3\Gamma_{\text{eff}} - 5}{5 - 3\gamma} + 1 .$$

No physical solution ($\Gamma_{\text{eff}} > 4/3$) can be found unless $A > 7/6$. This value is reached only in the outer half of the contracting star. We therefore find a situation in which the whole object is dynamically unstable, and thus it experiences a gravitation collapse, but with an outer half which presents a very strong supersonic turbulence.

Moreover, since the object is rotating, the kinetic energy of rotation stabilizes it (every type of kinetic energy has a stabilizing effect), whereas the dissociation energy unstabilizes it. If the object is rotating at nearly the maximum angular velocity, the critical value for instability is not $\Gamma = 4/3$, but less, 1.31. This accounts for the kinetic energy of rotation in the turbulence and defines a new polytropic index, a little bit larger than 3 (i.e., 3.09), which corresponds to a polytrope that is more dense than that of index 3.

We can add that at such high turbulent velocities, the object has a very high turbulent viscosity; that is, the kinematic coefficient of viscosity, defined as

$$\nu = lv$$

(i.e., the product of a characteristic length and a velocity), can be very large—of the order of 10^{15} or more—since v is comparable to the speed of sound and l is of the order of millions of kilometers. With such a high kinematic coefficient of viscosity, the object definitely rotates as a solid body over a distance of the order of $(vt)^{1/2}$. For a time scale of the order of 300 years, this yields a solid body rotation over approximately 30 million kilometers. In fact, it is easy to see that the relative depth of solid-body rotation remains constant and of the order of one-third of the radius of the contracting object.

We are left with an object that is rotating and contracting as a whole. It is in gravitational collapse from 10 000 to 100 solar radii but with the outer third rotating like a solid body. The dynamics of such an object is certainly not simple.

To make things simple (but not necessarily correct), we shall follow a former approach of Schatzman (1967) which assumes the law of conservation,

$$d(KMR^2\omega) = R^2\omega\, dM ,$$

where KR^2 is the square of the radius of gyration, dM is the fraction of the mass left behind, and ω is the Keplerian velocity of rotation, determined by the balance of gravitation and centrifugal force at the equator, given by

$$GM/R^2 = \omega^2 R .$$

From this, we derive an expression for the mass as a function of the radius,

$$M = M_{\text{fin}}(R/R_{\text{fin}})p ,$$

where

$$1/p = 2/K - 3$$

and M_{fin} and R_{fin} are the final mass and radius, respectively.

The picture that emerges from these considerations is that of an object that rotates as a solid body with the Keplerian angular velocity ω and contracts because of a secular, rather than a dynamic, instability while maintaining its characteristic shape and leaving mass behind along the equator; the object can be described by a polytrope of index 3.09. This is nothing more than the old Laplace picture. The only improvement is that a knowledge of the internal structure of the object allows us to estimate the total amount of mass left behind during the contraction.

The amount of mass $\nabla M/M$ left behind depends on the polytropic index. The more concentrated the object is, the less mass it leaves behind, since the angular momentum of the central body is greater. Some values of $\nabla M/M$ for different values of the polytropic index follow:

Polytropic Index	$\nabla M/M$
1.5	0.4
3	0.094
3.09	0.089

According to this theory, the mass of the primitive nebula is a little less than 10 percent of the solar mass. The mass of the terrestrial planets is about

$$\frac{2}{3 \times 10^5} M_\odot$$

(this includes Mercury, Venus, Earth, Mars, and the asteroids). We can assume that the terrestrial planets were formed from the solid particles left behind in the contraction, that is, that the gaseous components that were left behind escaped. If we are given the composition

H, He	99%
C, N, O, Ne, Mg	1%
Si and heavier	negligible

and we exclude from it the gases that certainly have escaped, we are left with approximately only 10^{-4} of the mass. Then, the amount of matter from which the terrestrial planets were formed was

$$M_\odot \frac{2 \times 10^4}{3 \times 10^5} = 0.066 M_\odot .$$

The inclusion of the mass of the outer planets yields a value close to $0.089 M_\odot$ that corresponds to the polytropic index 3.09. So, as a first approximation, the process through which matter is left behind during contraction could have produced a primitive nebula of about the right mass for the formation of the planets.

VII. PHYSICAL CONDITIONS IN THE PRIMITIVE NEBULA

We now have a picture of a central body (the proto-Sun) surrounded by a disk of dust and gas; what we would like to know are the physical conditions within the disk. We must take into account two effects: heating by the radiation coming from the central star, and the flux of cosmic rays emitted by the proto-Sun.

If we assume that the proto-Sun was surrounded, like the present Sun, by a spherical turbulent magnetic field that behaves like a diffusing shield, we can assume that a fraction of the cosmic rays produced by the proto-Sun were reflected and fell back onto the nebula where they acted as a heating mechanism. However, we do not know the radius of the diffusing shield, and we must make several assumptions. If $4\tau R^2\phi$ is the total number of particles of average energy W_0, we can write

$$4\pi R^2 \phi W_0 = \xi L ,$$

where L is the luminosity of the Sun at the time of the primitive nebula and ξ is a fraction smaller than unity. One problem in the study of the heating process is to explain what causes the opacity. We can assume only that dust particles were in the same abundance found now in interstellar space. One can compute the surface density in the disk (from the mass-radius relation for a polytrope), and the result (in atom-cm^{-2}) is roughly

$$S = 10^{28} \frac{1}{a_{\mathrm{AU}}^2} ,$$

where a_{AU} is the distance from the center in astronomical units (the deviation from this simple law is negligible). This is the number of atoms in a column parallel to the axis of rotation of the disk and with a surface area of 1 cm^2; but then, the number of dust particles is so large that the optical thickness of the system (in a direction parallel to the axis of rotation) is very large (10^5). Hence, radiation penetrates only a small distance, and even cosmic rays with an energy of $\approx$100 MeV do not go very far; thus, there would be an isothermal inner layer and heating in the outer layers. The effective (outer layer) temperature of such a nebula is given by

$$2\sigma T_{\mathrm{eff}}^4 = \xi L / 4\pi a_{\mathrm{AU}}^2 ;$$

e.g., for $\xi = 0.1$ and $L = 10^{34}$,

$$T_{\text{eff}} = a_{\text{AU}}^{-1/2} \times 236 \text{ K} .$$

The inner temperature can also be computed as a straightforward diffusion problem. Typical values are 70 K for the outer layer ($a_{\text{AU}} \approx 11$ AU) and 200 K for the central, isothermal part ($a_{\text{AU}} \approx 2$ AU). They correspond to $\xi = 0.1$ and a time scale of evolution of about 2×10^6 years.

Finally, we wish to point out that for the isothermal part of the nebula, there is an equation of state relating densities and pressures:

$$\rho = \rho_0 \exp\left|-A\frac{S^2}{a^3}\right|$$

(according to Schatzman, 1967), where $A = GM\bar{\mu}/2RT$, ρ_0 is the density at the equator, a is the distance from the center, and S is the distance above the equatorial plane. This equation, of course, defines a scale height

$$h^2 = a^3/A ;$$

the ratio between the thickness of the nebula and the radius is given by

$$h/a = 3 \times 10^{-3} a_{\text{AU}}^{1/2} T_e^{1/2} ,$$

where T_e is the temperature in the equatorial plane at the distance a_{AU} (in astronomical units). For the typical values obtained previously, h/a is very small; i.e., the nebula is flat.

VIII. EVAPORATION OF HYDROGEN IN THE NEBULA

The problem of the evaporation of hydrogen arises from very simple considerations concerning the abundance of deuterium in the solar system. We know that on the Earth the present ratio of deuterium to hydrogen, D/H, is about 2×10^{-4}. On the other hand, from radio-astronomy considerations, mainly undetected spectral lines, Weinnel has given an upper limit for this ratio in

interstellar matter:

$$(D/H)_{\text{interstellar}} < 5.5 \times 10^{-5},$$

which is definitely much smaller than the same ratio on the Earth. This difference should be explained.

One can make a very simple physical argument for the formation of deuterium. If the original ratio was of the order of 10^{-5} or even less, enough additional deuterium must have been produced to yield the observed ratio. The only way in which deuterium can be formed is by spallation of helium by protons. This process yields almost all the products possible; that is,

$$He^4 + p \rightarrow H^3, He^3, D, \text{ or } 4p .$$

It has an energy threshold of the order of 20 MeV, and the cross sections for the different possible products are of the same order of magnitude, i.e., 10 mb (10^{-26} cm^2).

Let us see how much energy is needed to produce the deuterium observed. Assume the deuterium is produced in a nebula that has a mass of $0.05M_{\odot}$ (10^{32} g), that is, just about half that of the primitive nebula we have considered previously. In such a nebula, there are about 2×10^{55} helium atoms. Suppose a flux of cosmic rays is causing the spallation. These cosmic rays shower into a large area of the nebula; the showers are not very dense and we must assume that the stopping of the cosmic rays takes place in a cold plasma. The cross section for stopping in a cold plasma is much larger than the cross section for the spallation, which means that the efficiency of the mechanism is very small. The total number of deuterons produced per square centimeter of the nebula is given by

$$\int \phi \frac{\sigma_{sp}}{\sigma_{stop}} \left(\frac{He}{H}\right) dt ,$$

where ϕ is the flux of cosmic rays per square centimeter and σ_{sp} and σ_{stop} are the cross sections for the spallation and stopping processes, respectively. The ratio of the latter two quantities gives the probability that a cosmic-ray particle may trigger a spallation process before being stopped. Obviously, we must multiply this by the ratio of the helium and hydrogen abundances. Finally, we integrate over the time period in which the process takes place.

Then, we may calculate how much energy is needed to produce this amount of deuterium since each cosmic-ray particle has a probability $\sigma_{sp}/\sigma_{stop}$ of causing a

reaction. The product of the average energy of the cosmic rays times the number of particles necessary to produce one deuteron will give us the average energy needed per particle to produce one deuteron, i.e.,

$$W = \bar{E}\frac{\sigma_{\text{stop}}}{\sigma_{\text{sp}}},$$

$\bar{E}$ being the average energy of the cosmic-ray particles. The required number of deuterons is

$$2 \times 10^{-4} \times 4 \times 10^{55} \approx 10^{52},$$

i.e., the observed value of the ratio D/H times the number of hydrogen atoms in the nebula (H = 2 He $\approx 4 \times 10^{55}$). If 100 MeV is the average energy of the cosmic-ray particles and 10^4 to 10^5 is the value of the ratio $\sigma_{\text{stop}}/\sigma_{\text{sp}}$, the amount of energy necessary to form one deuteron is

$$W \approx 10^6 - 10^7 \text{ MeV} \approx 1 \text{ to } 10 \text{ ergs}.$$

Hence, in order to produce 10^{52} deuterons, we need 10^{52} to 10^{53} ergs. In the Sun, there are $2 \times 10^{33} \times 6 \times 10^{23} \approx 10^{57}$ protons (the mass of the Sun times the number of protons per gram), so there must be

$$\frac{10^{52} \text{ to } 10^{53}}{10^{57}} = 10^{-4} \text{ to } 10^{-5} \text{ ergs per solar proton},$$

which amounts to 10 to 100 MeV per solar proton. During the contracting phase, the energy available per proton is of the order of 1 to 10 keV; hence, we see, even in this gross approximation, that we need 10 to 1000 times the energy available in order to obtain the observed amount of deuterium. In a nebula which has the solar composition, to produce that much deuterium by spallation only is then out of the question. The only way would be a process in which the hydrogen could escape faster than the helium, i.e., to evaporate hydrogen from the nebula faster than helium (at least in the region of the terrestrial planets). Such a process can take place. If it is possible to change the value of the ratio He/H from 1/2 to approximately 1000, it may then be feasible to have the deuterium formed by spallation of the remaining helium and thus to obtain the correct value for D/H.

Let us consider more closely the physical conditions in the primitive nebula, including the effect of the radiation coming from the Sun. The nebula is being heated by the Sun, mainly by thermal radiation, but an appreciable fraction of the heating is also caused by cosmic-ray radiation. At a distance of 1 or 2 AU, the temperature is about 200 K (at the equatorial plane). The temperature along the vertical extension of the nebula can be found in the following way: Take a frame of reference at a distance a from the center and let it rotate with the Keplerian velocity ω. In such a rotating frame of reference, the gravity and the centrifugal force balance each other, and the attraction of the Sun gives a net vertical component of apparent gravity, that is, a force that will pull particles toward the equatorial plane. The acceleration of a particle falling toward the equatorial plane is

$$g_z = -\omega^2 z ,$$

z being the vertical distance of the particle from the equatorial plane and ω the angular velocity of the Keplerian motion at the distance a. This result has been quoted repeatedly in a number of papers that have appeared during the last 25 years, and it is obtained only as long as one can neglect the mass of the nebula in comparison with the solar mass. With this, one can calculate the scale height of the nebula,

$$h^2 = \frac{2kT}{\mu} \frac{a^3}{GM} ,$$

and its density,

$$\rho = \rho_0 \exp(-z^2/h^2) .$$

If we take, for example, the expression for the effective temperature of the nebula as given by Safranov,

$$T_{\text{eff}} = (2/3\pi)^{1/4} (R_\odot/a)^{3/4} T_{\text{eff}_\odot} (1 - A)^{1/4} ,$$

where A is the albedo of the nebula (of the order of 0.27) and $T_{\text{eff}_\odot}$ is the effective temperature of the proto-Sun, we can deduce the central temperature which is due to the effect of radiative transfer. The formula shows that the temperature varies essentially as $a^{3/4}$. Then, $h^2 \sim a^{9/4}$, and $h/a \sim a^{1/8}$, which means that from the

Earth to Pluto the ratio of weight to distance is practically a constant (a factor of $40^{1/8}$, i.e., about 60 percent of the change in h/a); therefore, the nebula is practically flat. In the outer layers, the temperature is lower than in the center, which means there is an even smaller scale height near the surface and that there is a point where the density drops very sharply. Thus, we can picture the outer layers as constituting the atmosphere-like section of the primitive nebula.

We have assumed the Sun to be a flare star, and we know this implies cosmic rays; it also implies ultraviolet radiation. Let us consider what takes place when the nebula is irradiated by this radiation. As we shall see, there is no need to assume a constant UV flux but only stochastic flares with a frequency high enough for the time scale between two flares to be shorter than the relaxation time for the recombination of hydrogen.

In interstellar matter, the UV radiation produces an effect discussed 35 years ago by Strömgren. When a bright star embedded in neutral hydrogen emits UV radiation, a sphere of ionized hydrogen, known as a "Strömgren sphere", is formed which has a thin boundary (compared to the size of the sphere) in which the ionization drops sharply. This boundary, the region where the hydrogen recombines, is a strong emitter of the H_α line of hydrogen. Systematic H_α photographs have repeatedly confirmed the existence of these spheres around bright stars.

The case we are discussing is slightly different. The UV radiation from the Sun ionizes the surface of the nebula: What are the physical conditions induced by this process? When the absorbed photons ionize the hydrogen, the heat balance of the region is changed; the ionized region radiates less per unit mass than a neutral region, and this raises its temperature. We know that the temperature of ionized hydrogen is of the order of 5000 K. What is interesting about it is that the thermal velocity corresponding to 5000 K is comparable to the escape velocity at that distance from the Sun, that is, we do not have to go very far in the tail of the Maxwell distribution function to find velocities large enough for evaporation to take place. In other words, a large fraction of the ionized hydrogen will have a large enough velocity to leave the nebula. Thus, we have a mechanism through which hydrogen atoms can leave. Another very interesting aspect is that the velocity of the helium will always be half that of the hydrogen. (It does not matter whether or not the helium is ionized since it will be thermalized to the hydrogen temperature through collisions.) This means that the time scale for the loss of helium will be twice that for the loss of hydrogen. In summation, simply by looking at what happens on the surface layers of the nebula, we are able to conclude that hydrogen will escape twice as fast as helium.

If at a given distance the concentration of hydrogen is given (as a function of time) by

$$|\mathrm{H}| \sim e^{-t/\tau_{\mathrm{H}}},$$

τ_{H} being some characteristic time, the concentration of helium will be given by

$$|\mathrm{He}| \sim e^{-t/2\tau_{\mathrm{H}}},$$

and the hydrogen-to-helium ratio will be

$$|\mathrm{H}|/|\mathrm{He}| \sim e^{-t/2\tau_{\mathrm{H}}}.$$

If τ_{H} is not too large, this may be a mechanism to separate helium from hydrogen during the formation of the nebula. Then, the problem is to find the value of τ_{H} and the structure of the region in question more precisely.

There is a very simple way to look into the structure of the ionized boundary of the primitive nebula. Its thickness and density will be given by both the mean free path of the photons within it and the mean free path of the escaping protons. One could also calculate the rate of evaporation in this fashion; however, there is a simpler way. Given a certain flux L_γ from the Sun, we can calculate how much of it enters the nebula. More protons cannot be escaping from the nebula than there are photons entering it; this sets an upper limit to the number of escaping atoms, which yields a lower limit for the evaporation time scale.

Let us take a differential length da at a distance a from the center of the Sun. The number of incoming photons in a solid angle $d\Omega$ will be

$$\frac{L_\gamma}{h\nu}\frac{d\Omega}{4\pi}.$$

If we denote by dn/da the number of protons escaping per square centimeter-second, the total number of escaping protons will be

$$2\pi a\,da(dn/da).$$

By our previous argument, we can equate the above two quantities. Noting that

$$\frac{d\Omega}{4\pi} = \frac{1}{2}\frac{dh}{da}\frac{da}{a},$$

we can write dn/da as a certain flux per square centimeter divided by a characteristic time scale τ_H; that is,

$$dn/da = S/\tau_H .$$

We recall that $h \sim a^{9/8}$; thus,

$$h = h_0(a/a_0)^{9/8} ,$$

where a_0 and h_0 are constants that depend on the units chosen. From these relations, we deduce the value

$$\tau_H = \frac{8}{9}\frac{4\pi a_0^2 S_0}{(L_\gamma/h\nu)}\left(\frac{a_0}{h_0}\right)\left(\frac{a_0}{a}\right)^{1/8} ,$$

which, assuming the energy of the ionizing photon is 13 eV (2×10^{-11} ergs), is of the order of

$$\tau_H = 2.10^5(a_0/a)^{1/8}(L_\odot/L_\gamma) \text{ years} ;$$

this relation is nearly independent of distance. If we assume that the UV flux is of the same order as $L_\odot$ at Earth's distance, we obtain a time of evaporation there of about 10^5 years. If we assume what seems to us a more reasonable estimate, $L_\gamma \approx (1/10)\, L_\odot$, we obtain a time scale for the evaporation of the order of 1 to 2 million years, which seems to be a reasonable value.

We must remember that at the same time the spallation process is taking place, the helium is also evaporating. In order to gain the factor of 1000 necessary to produce the observed amount of deuterium, we need a time of the order of $7\tau_H$, that is, several million years. Clearly, it is difficult to decide what is the exact value of the time of escape since we do not know how many ionizing photons have been coming from the Sun during the phase in question. Certainly, a better analysis of the situation within flare stars is needed in order to get information we can use to find out how much evaporation has taken place in the Sun.

We are left nevertheless with a difficulty which must be considered. The evaporated hydrogen would have filled the space surrounding the nebula, but since we cannot detect it now, we must propose a mechanism to account for its absence.

A possible one could be the solar wind. This provides a flux of particles leaving the Sun at relatively high velocities (small mass loss) which can sweep away the hydrogen and helium particles. It does not have to be a continuous process but can proceed by bursts. Let us see whether this process is energetically possible. If we assume that the mass of the nebula is a few percent of the solar mass ($\approx 0.1\, M_\odot$), we arrive at a value an order of magnitude from the escape velocity at the Earth's distance (the escape velocity is smaller at larger distances). We find it is sufficient that a mass of $10^{-2} M_\odot$ go into the solar wind in 10^6 years in order to sweep away the helium and hydrogen evaporated from the primitive nebula. This is a small amount compared to both the energy needed for the separation of hydrogen and helium and the cosmic-ray energy needed to produce the deuterium.

IX. FORMATION OF THE PRIMARY BODIES IN THE NEBULA

We review now the problem of the formation of the primary bodies in the primitive nebula as it was treated by McCrea and Williams (1965). The discussion is centered on the behavior of dust particles in the cloud. Essentially, one considers the effect of a weak gravitational field on the dust particles (the field due to a large sphere of gas which has a mass comparable to that of the Earth but is much more extended). The argument is based entirely on the friction mechanism, and the conclusions are mainly the following: The small dust particles experience such a large friction that they do not fall but remain embedded in the gas. The larger particles, on the other hand, can fall at a very low speed. Moreover, the question is whether their mass stays constant or whether other particles may stick to them. For densities of the order of those considered above (i.e., 10^{16} particles-cm^{-3}), the major effect seems to be the growth of the larger particles along their fall. The larger particles sweep the smaller ones, which increases their own mass, whereupon they fall faster and sweep more particles until they arrive near the center of the cloud with a relatively large mass. We take the ratio of particles of dust to hydrogen in the cloud (in number, not in mass) to be of the order of 10^{-13}.

The formation of the primary bodies is interesting enough to justify going into a little more detail. This will enhance our understanding as well as clarify the physical assumptions that are being made. To study the physics of the dust particles, we shall assume, with McCrea and Williams, that they look very much like whiskers,

or filaments. This is substantiated by what we know about the mass-to-volume ratio of the micrometeorites that seem to have a relatively large volume for a small mass, which can be explained only if they are made of filaments. We shall also assume that although these dust particles may contain a fair amount of iron, silicates, and so on, their density is of the order of unity. It is very simple to transfer the analysis made by McCrea and Williams for the case of a sphere to our case of a flat nebula because, just as in a homogeneous sphere, the gravity in the nebula is proportional to the distance; instead of being proportional to the distance from the center, it is proportional to the distance from the equatorial plane. As we stated previously, at a height z over the equator the gravitational force is given by $-g^2 z$. A dust particle that starts with a radius r_0 and is sweeping other particles will have the radius

$$r = r_0 + \alpha x ,$$

after traveling a distance x. The proportionality constant α is given by

$$\alpha = \frac{1}{4} \frac{k\rho}{\sigma} ,$$

where k is the mass fraction of dust embedded in the gas, ρ is the density of the gas, and σ is the density of the material of the dust particle. For simplicity, we assume a uniform density of the nebula over a layer of thickness $2h$, with h being the scale height introduced previously. We have then

$$h - x = z ,$$

x being the distance over which the dust particle has traveled while accreting other dust particles.

To determine the force exerted on the dust grain by the medium, we have to introduce the viscosity μ. If we assume our case is analogous to that of a particle moving through a fluid, we can use Stokes' expression for the viscous force,

$$P = 6\pi\mu r \frac{dx}{dt} ,$$

which is proportional to the radius of the particle and to its velocity. The equation of motion of a particle that changes its mass m while moving is

$$\frac{d}{dt} m \frac{dx}{dt} + P - mg = 0 ,$$

where mg is the gravitational force. Once the motion is established, the acceleration is very small and we can assume, with McCrea and Williams, that the acceleration term is negligible and write

$$6\pi\mu r \frac{dx}{dt} = \text{gravitational force} .$$

If we set

$$m = \frac{4}{3}\pi\sigma(r_0 + \alpha x)^3$$

and

$$g = \omega^2 z = \omega^2(h - x) ,$$

the equation of motion becomes

$$6\pi\mu(r_0 + \alpha x)\frac{dx}{dt} = \frac{4}{3}\pi\sigma(r_0 + \alpha x)^3\omega^2(h - x) .$$

From it, we can obtain both a solution for the motion and a solution for the radius. The result is simple, particularly if we can assume that the initial radius is much smaller than the radius after the end of accretion, i.e., that the particle grows enormously. After a straightforward integration, we find

$$\frac{4}{3}\pi\sigma r_H^2\omega^2\sqrt{\pi/kTm_H}\, t = \frac{1}{a^2h^2}\left[\frac{\alpha h}{\gamma_0} + \log\frac{\alpha h}{(1 - x/h)r_0}\right] ,$$

with $r_0 \ll \alpha h$, where we have substituted for μ its expression and the subscript H refers to the hydrogen molecule. It is not worth going into the detailed computation of this equation. The main point is that the logarithmic term is always negligible except when x becomes similar to h, but we can consider anything that close to the equator to be at the equator itself. So, neglecting this term and substituting μ for its expression

$$\mu = \frac{1}{6\pi}\sqrt{kTm_H/\pi}\,\frac{1}{r_H^2} ,$$

we find that the time needed for a particle to fall into the equator is

$$t = \frac{9}{2}(\mu/\omega^2)(4/kr_0 s) ,$$

where the surface density $s = h\rho$ and we have substituted for the expression given previously.

At 1000 K, μ can be taken as 5×10^{-5} g-cm^{-1}-s^{-1}, k as 10^{-2} (1 percent), and the surface density as 10^4 g-cm^{-2}. At the Earth's distance, the value of t is then

$$t = 6/r_0 \text{ years} .$$

After taking into account that ω, s, and the viscosity change with distance, we get an extra factor $(a/a_0)^5$; i.e.,

$$t = (6/r_0)(a/a_0)^5 \text{ years} ,$$

a_0 being the astronomical unit. This is the result that has been underlined by McCrea and Williams: The time of fall is terribly short. If we take

$$r_0 \approx 100\ \mu\text{m} = 10^{-2}\ \text{cm} ,$$

this time is 600 years. Even if we take $r_0 = 10\ \mu$m, it is only 6000 years. A radius of 1 μm is out of the question, because then the expression for the viscous force would not be the same, and the particle would not fall at all. This means that as soon as a particle becomes bigger than a minimum size, it will fall rapidly toward the equator, sweeping material on its way. It is interesting to determine the radius of the particle and its mass from the quantity αh. The radius is of the order of 0.5 m, which means that, at least in the Earth's region, there would be planetesimals that have a weight of about 1 ton, lie very close to the equatorial plane of the primitive nebula, and were formed in a very short time.

There is, to be sure, the question of what takes place at the distance of Jupiter; then, the factor $(a/a_0)^5$ becomes a very large number (5^5, or roughly 3000), and the corresponding time is of the order of 1 million years, which is just at the limit of possibility for forming planetesimals. At longer distances, those of Saturn and Uranus, such a process will not take place at all; the time scale is much too long. Apparently this is due to two facts: The gravity is too weak and the mass (or the surface density) is too small for the process to proceed. This is a rather new aspect of the physics of the nebula at large distances. To go into more detail would require following the thermal history more closely. Up to this point, the nebula has been considered to be at a constant temperature, but if the temperature drops, the scale height becomes smaller, the density increases, and the viscosity changes, and we no longer know what is taking place.

X. ANGULAR MOMENTUM AND THE LAW OF PLANETARY DISTANCES

We turn now to the mechanical problem considered by O. Yu Schmidt a few years ago as a way of expressing the law of planetary distances. As we mentioned previously, the idea is to consider the angular momentum p and to take a momentum distribution $f(p)$. The angular momentum of a planet will be determined by the average angular momentum in a ring of the nebula. If we call the boundaries of the ring n and $n + 1$, the angular momentum of the nth planet will be given by

$$p_n = \frac{\int_{q_n}^{q_{n+1}} pf(p)\,dp}{\int_{q_n}^{q_{n+1}} f(p)\,dp},$$

the q's being convenient limits of integration at the boundaries n and $n + 1$. The idea of Schmidt is to write

$$q_n = \frac{p_n + p_{n-1}}{2},$$

or in other words, that the boundaries of the ring over which we integrate are themselves defined by the average of the angular momentum in the ring. This is the starting point. To get an idea of what it all leads to, assume that

$$f(p) \sim p^{\lambda},$$

λ being a parameter. Then, one obtains

$$p_n = \frac{\lambda + 1}{2(\lambda + 2)} \frac{(p_{n+1} + p_n)^{\lambda+2} - (p_n + p_{n-1})^{\lambda+2}}{(p_{n+1} + p_n)^{\lambda+1} - (p_n + p_{n-1})^{\lambda+1}},$$

This is certainly an oversimplified picture because there is no reason to believe that the accretion has taken place in a ring with sharp boundaries. Nevertheless, it is interesting to see how the law of planetary distances could be obtained from a simple mass distribution.

There are easy cases that can be considered. Schmidt has taken the one in which $\lambda = 0$. This corresponds to a linear distribution in angular momentum and gives

$$p_n = \frac{1}{4}(p_{n+1} + 2p_n + p_{n-1}) ,$$

a very simple relation whose solution is

$$p_n = a + bn ,$$

a and b being constants. This corresponds to

$$\sqrt{r_n} = a + bn ,$$

which, as we mentioned earlier, is the expression for the law of planetary distances that gives the best fit with the numerical values of the semimajor axes of the orbits of the planets.

The question is how λ relates to the mass density. Surface density σ is related to the angular momentum by

$$\sigma \sim p^{6\nu-4} ,$$

where $1/\nu = (2/K) - 3$, K being the moment of inertia. This yields

$$\lambda = 6\nu - 1 .$$

The value ν is a very small quantity that depends on the polytropic index of the contracting object; for the polytropic index 3,

$$\lambda = -0.88 ,$$

which is relatively far from the value $\lambda = 0$. It is not too easy to calculate a solution p_n for any arbitrary value of λ, but it turns out, at least from one numerical experiment, that p_n is not affected too much by the value of λ. In summation, we might say that as a first approximation (keeping in mind that it rests on an assumption concerning the limits of the region of accretion), the value thus obtained for p_n is not incompatible with the law of planetary distances.

XI. SPALLATION AND THE ORIGIN OF THE LIGHT ELEMENTS

Naturally, the problem of the origin of the light elements is different from that of the origin of the heavy ones, which are always assumed to be preexistent to the formation of the solar system (including the radioactive elements formed in the last supernova explosion, which also took place before the formation of the solar system). By light elements, we mean essentially lithium, beryllium, and boron. There are some slight difficulties with regard to Li^6 and Li^7 in particular for the following reason. When spallation takes place (much above the threshold for the formation of light elements) by means of the reactions

$$p + C, \quad p + N, \quad \text{and} \quad p + O,$$

which give various products, including

$$Li^6, Be^6 \text{ (radioactive)} \rightarrow Li^6,$$

$$Li^7, B^7 \text{ (radioactive)} \rightarrow Li^7,$$

and so on, the average cross sections are such that the ratio Li^6/Li^7 (which equals the inverse ratio of the spallation cross sections for the formation of Li^6 and Li^7) is about 1/2. On the other hand, what we observe on the Earth is

$$Li^6/Li^7 \approx 1/12.$$

The value of this ratio on the Sun's surface is not known.

If there is a universal mechanism for lithium formation (for example, if lithium was formed before the solar system), it could be possible for the Li^6/Li^7 ratio to be of the order of 1/2, since lithium might have been produced by spallation in interstellar space by cosmic rays. However, we would then have to explain how the Li^6 was partly destroyed during the formation of the solar system.

One simple explanation might be that the lithium got mixed inside the proto-Sun during its formation and that the Li^6 was then destroyed by hydrogen burning, that is, by the reaction

$$Li^6 + p \rightarrow He^3 + He^4,$$

which takes place faster than

$$Li^7 + p \rightarrow He^4 + He^4 .$$

However, for this reaction to take place, the temperature must become as high as 2×10^6 K. At the time of the formation of the primitive nebula, the temperature at the center of the proto-Sun was never as high as that; even at the time when the Sun was the size of Mercury's orbit, the temperature never exceeded about 10^5 K, which is much too low for lithium burning. Furthermore, it is very difficult to imagine that the material of the primitive nebula may have got inside the Sun at a later phase and that it was ejected back into the nebula. Physically, this would be extremely hard to believe, both for hydrodynamical and for energetical reasons. Also, if we look at an object like a T-Tauri star, matter appears to be ejected rather than absorbed. In these stars, the lithium abundance is very large and (in the few known cases) the ratio Li^6/Li^7 seems to be 1/2.

Another explanation is that lithium was formed in the primitive nebula by the same spallation process which has been assumed in order to explain the formation of deuterium, except that it could have been the spallation of carbon, nitrogen, and oxygen instead of the spallation of He^4. Then, the reason why the ratio Li^6/Li^7 is 1/12 and not 1/2 would be found essentially in the shape of the energy spectrum of the cosmic rays showering the primitive nebula.

Before studying the problem of the Li^6/Li^7 ratio, one should calculate whether the cosmic-ray flux that is supposed to have been able to produce the observed D/H ratio would also have been sufficient to produce the observed value of the lithium ratio. With the same flux producing the observed deuterium, it is possible to estimate how much lithium was formed. The D/H ratio is

$$\frac{D}{H} = \frac{1}{H} \phi_{cr} \frac{\sigma_D}{\sigma_{el}} \frac{He}{H} ,$$

where ϕ_{cr} denotes the number of cosmic rays per second per square centimeter of the nebula integrated over the time in which the process takes place. We can write a similar relation for the lithium:

$$\frac{Li}{H} = \frac{1}{H} \phi_{cr} \frac{\sigma_{Li}}{\sigma_{el}} \frac{C,N,O}{H} ,$$

where σ_{Li} denotes the cross section for the spallation of carbon, nitrogen, and oxygen to produce lithium.

If we take the ratio of the two expressions, we obtain the lithium to deuterium ratio, which contains only the spallation cross sections for lithium and deuterium and the abundance of C, N, and O, relative to He:

$$\frac{\text{Li}}{\text{D}} = \frac{\text{C,N,O}}{\text{He}} \frac{\sigma_{\text{Li}}}{\sigma_{\text{D}}} .$$

The cross-section ratio gives a factor of 2 to 3, and the ratio of abundances is about 10^{-2} to 10^{-1}, so the Li/D ratio has a value of 0.3 to 0.1. This order of magnitude seems to be compatible with the amount of lithium observed at the surface of the Earth. The author feels that there is fair chance that at least a fraction of the lithium on the Earth was formed by the same spallation process which is assumed to have been the origin of the deuterium.

XII. CONCLUSION

As one can see, the questions raised by this theory of the origin of the solar system by far outnumber the answers given here. It would be easy to draw a very long list of various questions that deserve to be considered in more detail. Certainly, the first question would be that of the thermal history of the primitive nebula. A finer analysis of the nuclear processes that have taken place in the nebula is required, including the way in which the energy spectrum of the cosmic-ray particles has been changed by their motion within the primitive nebula. It would be worth getting a better understanding of how hydrogen was separated and, finally, of how condensation took place.

The author's feeling is that a theory of the origin of the solar system based on these kinds of considerations is likely, by its own logic, to provide, for example, the physical conditions the geochemists need in order to explain the special features of the Earth, the asteroids, and other bodies in the solar system.

REFERENCES

Gloeckler, G., and Jokipii, J. R., II, *Astrophys. J. (Letters)* 148:L45, 1967.
Kraft, R. P., "Stellar Rotation" in *Stellar Astronomy,* H. Y. Chiu et al., eds., Gordon and Breach Science Publishers Inc., New York, 1969, p. 341.
Lovell, B., *Observatory* 84:195, 1964.
McCrea, W. H., and Williams, I. P., *Proc. Roy. Soc. A* 287:143, 1965.
Schatzman, E., *Ann. Astrophys.* 30(6):963, 1967.
Schmidt, Otto, "A Theory of the Origin of the Earth", Lawrence and Wishart, London, 1959, p. 49.
Walker, Merle F., "Studies of Extremely Young Clusters I. NGC 2264", *Astrophys. J.* Suppl. Ser. 2:376, 1955-56.

BIBLIOGRAPHY

Schatzman, E., *Bull. Acad. Roy. Belg.* 35:1141, 1949.
Schatzman, E., *Ann. Astrophys.* 25:18, 1962.
Schmidt, O. Yu, "Four Lectures on the Theory of the Origin of the Earth", Moscow, 1957.

CHAPTER 12
EVOLUTION OF PLANETARY ATMOSPHERES*

S. I. Rasool
Institute for Space Studies
Goddard Space Flight Center

I. INTRODUCTION

The present composition of the Earth's atmosphere is 78 percent nitrogen, 21 percent oxygen, and 1 percent argon, with traces of carbon dioxide, water vapor, and ozone. The atmospheres of Mars and Venus, on the other hand, are predominantly composed of carbon dioxide, whereas those of Jupiter and Saturn contain mainly hydrogen and helium, with small amounts of methane and ammonia. Such a wide variety in the composition of the atmospheres of the planets is most intriguing when one considers that all nine planets were probably formed at the same time and out of the same chemically homogeneous mixture of gas and dust, that is, the primitive solar nebula.

The most likely explanation for this diversity in composition seems to be that the planetary atmospheres have undergone important evolutionary changes during their long history of about 4.5 billion years. If all planets acquired a substantial amount of atmosphere at the time of their accretion, these atmospheres then went through profound transformations because of the escape of lighter gas into space and the continued replenishment of the atmosphere by outgassing from the interior and by chemical reactions of extreme complexity between the atmospheric gases and crustal rocks. The degree to which each of these processes has influenced the evolutionary history of a planet depends essentially on its size, mass, internal structure, and distance from the Sun.

*To be published in *The Origin of Life in the Universe*, edited by Cyril Ponnamperuma, North-Holland Publishing Co., Amsterdam, 1971.

II. EARLY HISTORY OF THE EARTH'S ATMOSPHERE

The first, and probably the most important, mechanism by which a planet acquires an atmosphere is the capturing of gaseous atoms and molecules into its gravitational field at the time of its accretion. The gross composition and approximate amount of such an atmosphere can be estimated with reasonable confidence because the planet and its atmosphere should contain elements in the same relative abundances as they were present in the primitive solar nebula out of which the planets accumulated. The elemental composition of the primitive Sun has been estimated by Cameron (1968) (Figure 1), with due consideration of the fact that the composition in the outer regions of a contracting proto-star is different from that in its interior. From this figure, it is evident that hydrogen is the most abundant element, followed by helium. Carbon, nitrogen, oxygen, and neon, though each is about 1000 times less abundant than hydrogen, compose the second most important group of elements in the solar nebula. Elements like silicon, magnesium, and iron, which today make up most of the solid Earth, were present only in minute amounts (<1 percent) in the primordial mix from which the planets accumulated.

If the temperature at planetary distances in the primitive solar nebula was between 100 and 300 K, as suggested by Urey (1959), a thermodynamic calculation indicates that a planet, at the time of its formation, would acquire a gaseous envelope with an initial composition similar to that given in Table 1.

However, the total extent of this gaseous envelope will be extremely massive because it is clear (from Figure 1) that hydrogen is approximately 400 times more abundant (by weight) than the group of elements magnesium, silicon, and iron which, along with oxygen, constitute the entire solid Earth. In other words, by this process of acquisition of an atmosphere, the proto-Earth and also Mars and Venus would have had to be 300 to 400 times more massive than their present value and essentially composed of gaseous H_2 and He. This is exactly the situation at the present time on Jupiter and Saturn. As deduced from spectroscopic observation, their compositions are precisely the same as that of the primitive solar nebula, shown in Table 1. On the other hand, the composition of the Earth is vastly different. Table 2 shows that, relative to the primitive solar nebula, not only is the Earth deficient in hydrogen and helium (which being the lightest gas can be assumed to have escaped from the gravitational field of the earth) by several orders of

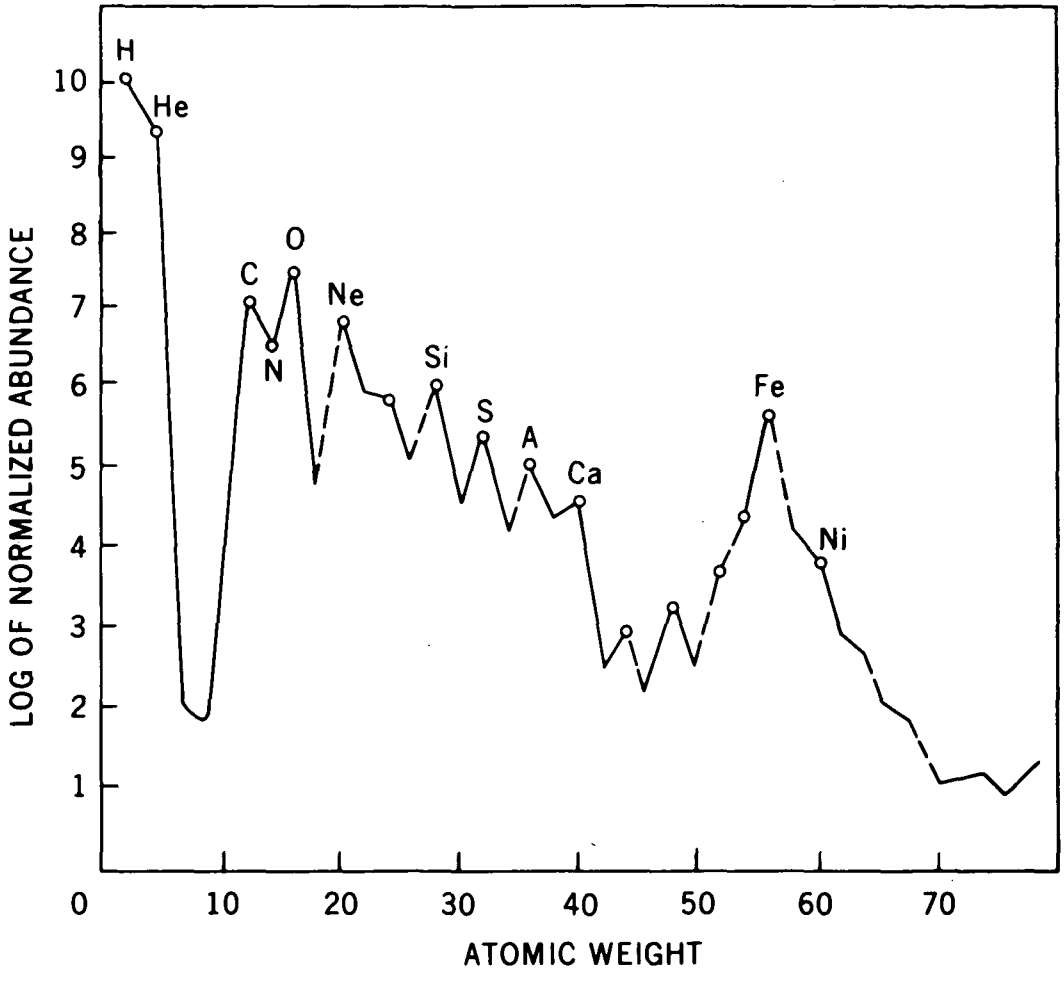


Figure 1.—Abundances of elements in the primitive solar nebula (after Cameron, 1968).

magnitude, but it is also deficient in carbon, nitrogen, and cosmically abundant rare gases (Ne, A^{36}, Kr, and Xe) by as much as a factor of 10^6. A comparison of the deficiency factors of these elements with those of nonvolatiles, such as sodium, magnesium, aluminum, and silicon, strongly suggests that the Earth lost only those

Table 1.–Possible initial composition of a planetary atmosphere at the time of the planet's formation.

Gas	Percent by Weight
H_2	63.5
He	34.9
H_2O	0.6
Ne	0.34
CH_4	0.26
NH_3	0.11
A^{36}	0.15

Table 2.–Comparison of the compositions (given in atoms per 10 000 atoms of silicon) of the Earth and solar system.

Element	Whole Earth* *a*	Solar System** *b*	Deficiency Factor log (*b*/*a*)
H	250	2.6×10^8	6.0
He	3.5×10^{-7}	2.1×10^7	13.8
C	14	135 000	4.0
N	0.21	24 400	5.1
O	35 000	236 000	0.8
Ne	1.2×10^{-6}	23 000	10.3
Na	460	632	≈0
Mg	8 900	10 500	≈0
Al	940	851	≈0
Si	10 000	10 000	0
A^{36}	5.9×10^{-4}	2 280	6.6
Kr	6×10^{-8}	0.69	7.1
Xe	5×10^{-9}	0.07	7.1

*After Rubey (1955), Ringwood (1964), and Mason (1966).

**A. G. W. Cameron, "The Origin of Planetary Atmospheres", selections from the TRW Space Technology Laboratories Lecture Series, presented on March 3, 1964.

elements which are volatile at a temperature of a few hundred degrees. It therefore appears that in its very early history, Earth became completely devoid of its gaseous envelope. How then did our planet acquire its present atmosphere and oceans?

Geochemists have presented convincing evidence that during geological history, both the atmosphere and oceans of the Earth developed slowly through a number of well-defined processes (Mason, 1966). Additions to the Earth's sphere during geological times include the following:

(1) Gases released in volcanic emanation and the crystallization of magmas (mainly H_2O and CO_2, with small amounts of N_2 and traces of HCl, HF, H_2S, and SO_2).

(2) Oxygen produced by photochemical dissociation of water vapor.

(3) Oxygen produced by photosynthesis.

(4) Helium from the radioactive breakdown of uranium and thorium.

(5) Argon from the radioactive breakdown of K^{40}.

(6) Additions by solar winds.

Losses during geological times include the following:

(1) Hydrogen and helium by gravitational escape.

(2) Carbon dioxide by the formation of coal and petroleum and by organic burial.

(3) Carbon dioxide by the formation of calcium and magnesium carbonates.

(4) Oxygen by oxidation of H_2 to H_2O, free iron to ferrous and ferric iron, sulphur to sulphates, and so forth.

(5) Nitrogen by formation of oxides in the air and nitrifying bacteria in the soil.

The most important process by which gases have been supplied to the atmosphere and oceans is volcanic activity, which has probably been effective during the entire period of 4.5 billion years and has been the main source of water, carbon dioxide, and nitrogen. Oxygen was produced later by photosynthesis, and the traces of argon which are today present in the atmosphere have been released slowly as radioactive products during geological history. Simultaneous with these additions to the atmosphere has been a considerable loss of H_2 and He to space by gravitational escape, of CO_2 to the crust to form carbonates, and of O_2 to form oxides. Though the significance of these evolutionary processes has been fairly well established, several important questions have yet to be answered. What has been the history of the volatiles now present at the surface of the Earth? Have the carbon, nitrogen, oxygen, and hydrogen always been in the form of CO_2, N_2, H_2O, and H_2, or did carbon and nitrogen combine with hydrogen early in the Earth's history to form

CH_4 and NH_3? Under what atmospheric conditions did life originate on Earth, and how did the appearance of life change the atmosphere? These are some of the basic questions that must be answered in order to paint a coherent picture of the evolution of the Earth's atmosphere.

Opinions on these questions are many and varied; sometimes they are almost diametrically opposed. The Oparin-Urey theory (Oparin, 1938; Urey, 1952) of the origin of life on the Earth, supported by the laboratory experiments of Miller (1953) and more recently of Ponnamperuma and Klein (1970) and others, suggests a primitive atmosphere composed mainly of CH_4, with small amounts of NH_3, H_2, and H_2O vapor. On the other hand, the school of thought represented by Abelson (1966), and supported also by laboratory experiments on the synthesis of amino acids, holds that the early atmosphere of the Earth was made of CO_2, CO, H_2, and H_2O vapor.

Geologists are also divided on the subject. Holland (1962) has presented a model for the evolution of the atmosphere in which, during a very early stage after the formation of the Earth and at the commencement of outgassing, the major components of the volcanic emanation were CH_4 and H_2 rather than CO_2 and H_2O. This was so because oxygen was deficient in the volcanic "melt", having been removed by free iron, which was more abundant in the crust at that time than now. Under these conditions, the atmosphere of the Earth would be largely composed of H_2 and CH_4, with small amounts of NH_3 and H_2O, *provided* free hydrogen did not escape as rapidly as it does today. Rubey (1955), on the other hand, believes that the early atmosphere was probably made of CO_2 and N_2 because not enough hydrogen was available to keep CH_4 from converting into CO_2. Holland's model (Holland, 1962) is supported by the calculations of McGovern (1969) and the author, who have investigated the thermal properties of model primitive atmospheres of the Earth. They find that in a 99 percent CH_4, 1 percent H_2 atmosphere, the average exospheric temperature may be as low as 650 K (compared with the present-day value of 1500 K), making the escape of hydrogen a relatively slow phenomenon (Figure 2). However, Abelson (1966) has argued that if methane were abundant in the primitive atmosphere, the earliest rocks should contain unusual amounts of organic matter, which apparently is not the case.

Despite the disagreement over the composition of the primitive atmosphere, it is almost certain that it was devoid of free oxygen. How and when did free oxygen become a major constituent of the atmosphere? There are two main sources of production of free oxygen: first, the dissociation of H_2O vapor in the upper atmosphere and the subsequent escape of hydrogen, and second, photosynthetic

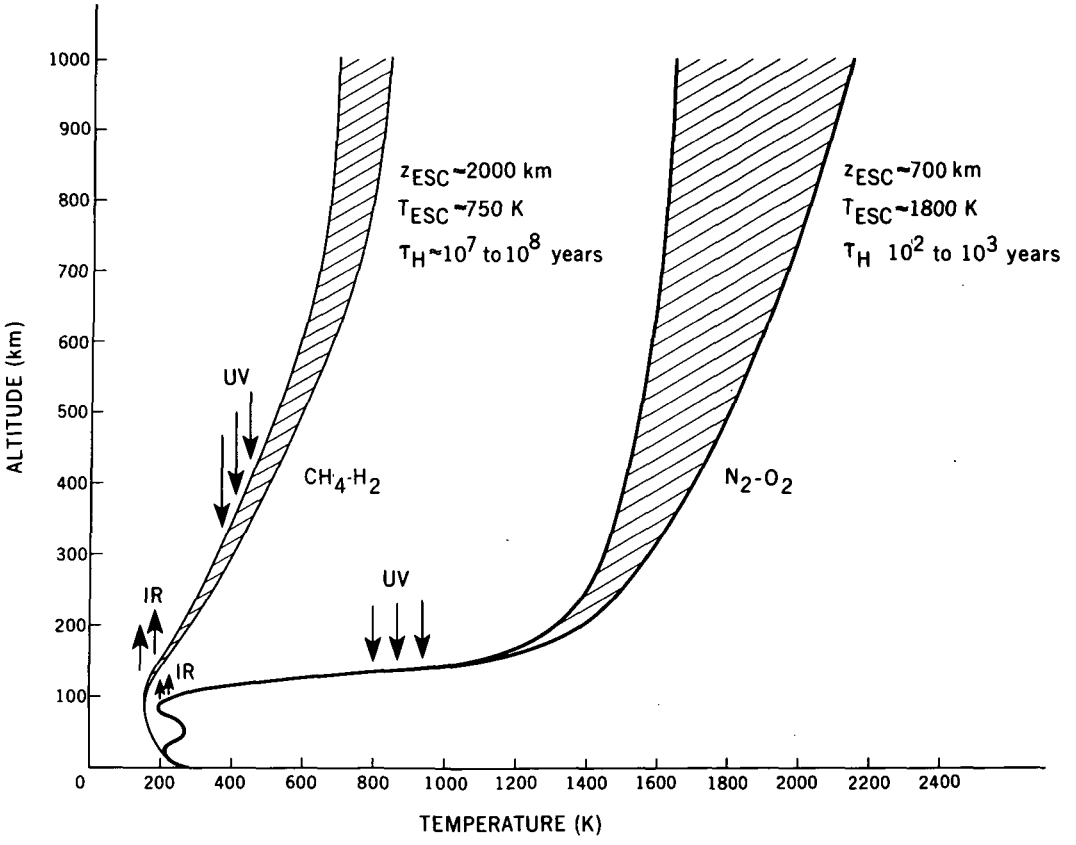


Figure 2.–Vertical distribution of temperature calculated for a methane-hydrogen primitive atmosphere. Also shown is the temperature profile in the present-day atmosphere of the Earth.

fixation of CO_2 and H_2O to produce carbohydrates and oxygen. During the prebiological history of the Earth, however, the photodissociation of H_2O was the only source of free oxygen. Several calculations have been carried out to determine the amount of O_2 that could have been produced by this process. The results are discordant, and estimates of the abundance of free oxygen in the primitive atmosphere range between 10^{-3} and 10^{-1} of the present amount. The principal reason for this uncertainty is the difficulty in obtaining a reliable estimate of the escape flux of hydrogen from such an atmosphere. At the same time, the question of the abundance of O_2 in the primitive atmosphere is of considerable importance

because an equilibrium amount in the atmosphere as large as 10^{-1} of the present-day value would be amply sufficient to oxidize CH_4 and NH_3 rapidly into CO_2 and N_2. All theoretical attempts in calculating the abundance of O_2 in the Earth's atmosphere have neglected the presence of NH_3. It is quite possible that even a small amount of NH_3, which is a strong absorber of ultraviolet radiation, will inhibit the dissociation of water vapor, and thus the primitive atmosphere of Earth may have been protected against oxidation for the first several million years.

III. MARS

Viewed through a telescope, Mars presents a most fascinating and colorful picture. More than half the planet appears to be made of a reddish rocky material; about one-third is comprised of dark areas like the maria of the Moon. The poles are covered with a bright whitish material whose extent waxes and wanes with the seasons. Some early optimistic observers also noticed "canals", which immediately led to speculation that Mars was a haven for life flourishing in those darker regions believed to be the remains of ancient oceans. Later observations seemed to support these ideas when the Greek and French astronomers reported observing seasonal changes in the intensity of the dark regions. A wave of darkening would spread across the planet when the polar ice started to melt. It was believed that the additions of water every spring rejuvenated the Martian flora and fauna, resulting in the observed changes in color.

However, in the last decade, intense activity in space exploration and ground-based astronomy has provided us with new information that forces us to completely revise our long-held notions regarding the surface and the atmosphere of Mars. It now appears that the planet, instead of being a haven for life, is more like the Moon, a rugged surface covered with craters: The dark areas may be only the sloping parts of the same desolate terrain; the canals are nothing but linear hills or chains of craters; and even the polar caps, which were believed to be a reservoir of water to feed the planetary canal system, are instead composed of frozen carbon dioxide at a temperature of 148 K.

This extremely inhospitable nature of the planet, as revealed by spacecraft observations, is not confined to the surface alone. The atmosphere of Mars also appears to be entirely different than that of the Earth. It was expected that, like Earth, Mars had an abundance of nitrogen in the atmosphere, with small amounts of oxygen and carbon dioxide–a breathable atmosphere.

The evidence for the unexpectedly different nature of the Martian environment has been obtained almost entirely by the exploration of the planet by Mariner spacecraft. The latest results from the Mariner experiments suggest that not only is the total amount of atmosphere on Mars only one one-hundredth of that on Earth, but also its composition is entirely different: almost pure CO_2 with little or no nitrogen.

The first spacecraft (Mariner 4), which flew past Mars in 1964 at a distance of 20 000 km, returned a large amount of new information. Twenty-two historic pictures taken from a distance of 20 000 to 40 000 km revealed the most unexpected feature of the Martian surface, namely, the existence of craters. The largest of the craters (300 km X 300 km) visible in the first photographs showed little erosion, indicating the primeval nature of the surface and suggesting that the planet has not had a heavy atmosphere or large oceans since early in its history.

As for the atmosphere, a team of scientists from NASA's Jet Propulsion Laboratory devised a brilliant experiment performed on board Mariner 4 to determine the surface density and the atmospheric structure of Mars. This experiment, which is described in more detail in the section on Venus, revealed quite unexpectedly that the total atmospheric pressure may be only 6 mb (contrasted to 1000 mb for Earth) and that the atmosphere may be largely composed of CO_2.

In July and August 1969, a much more detailed study of the Martian surface and atmosphere was accomplished by Mariners 6 and 7, which flew past Mars as close as 2000 km. These spacecraft were equipped with high-resolution TV cameras capable of photographing large portions of the planet with a resolution of about 300 m, and with a variety of other instruments designed to measure the surface and atmospheric temperatures, composition of the atmosphere, and structure of the ionosphere and the upper atmosphere. Both missions were successful.

Figure 3 shows a section of typical Martian terrain photographed by Mariner 6, covered with craters of different sizes. Careful analysis of about 200 photographs of similar resolution reveals that three principal types of terrain exist on Mars: (1) the cratered terrain, which resembles the surface of the Moon (the absence of tectonic activity and extensive erosion suggest that this terrain may be as old as the planet itself); (2) the chaotic terrain, which consists of small ridges and depressions, especially in the region of Meridiani Sinus; and (3) the featureless terrain, which is devoid of any relief (the best example of this type of terrain is in the region of Hellas).

The existence of three different types of terrain strongly suggests that though the major part of the planetary surface may be primeval and has probably not

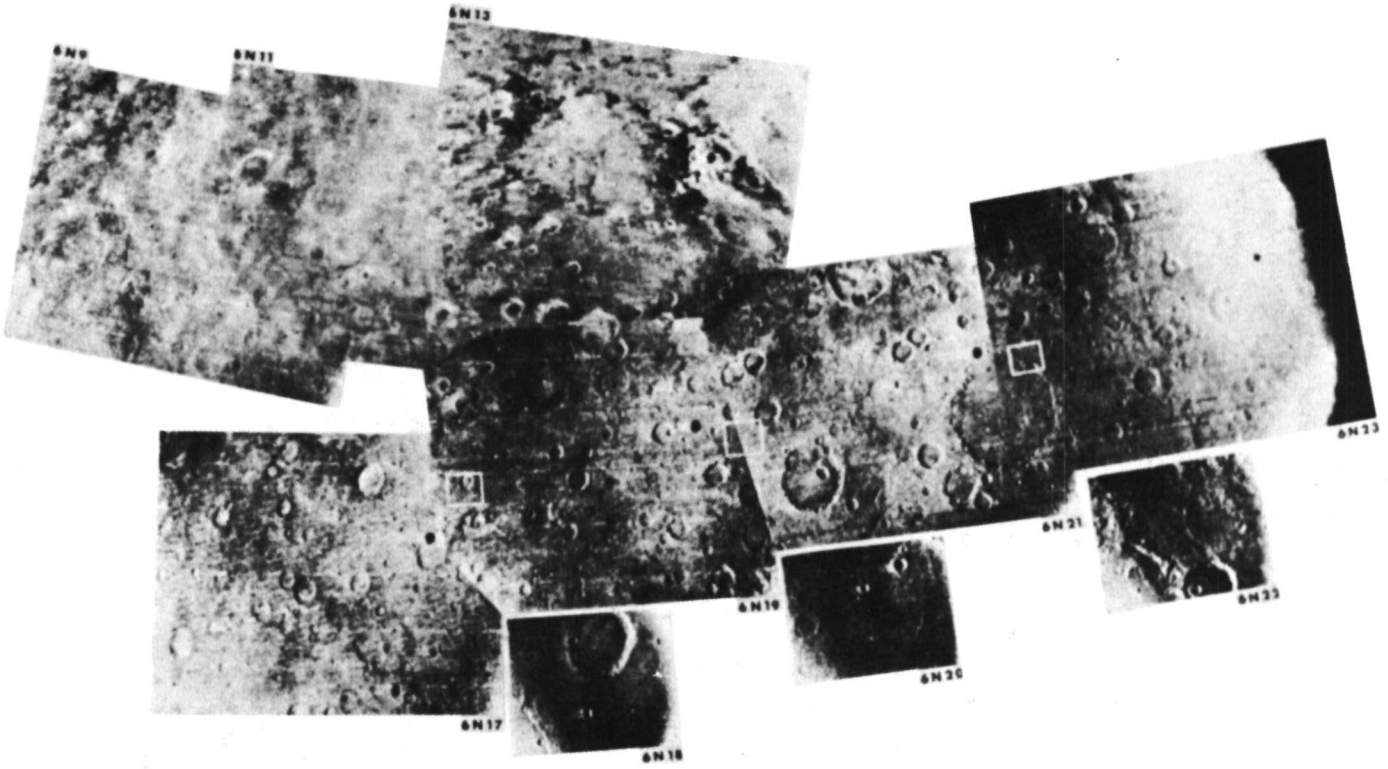


Figure 3.—A mosaic of four pictures of Mars taken by Mariner 6 covering an area 4000 km across and 700 km wide, parallel to 15° S latitude.

Table 3.—Surface pressures and temperatures on Mars, from Mariner occultation experiments.

Occurrence	Latitude	Longitude	Martian Local Time	χ (deg)	Surface Pressure (mb)	Surface Temperature (K)
Mariner 4 entry	50.5° S	177.0° E	13:00	67	4.9	160
Mariner 6 entry	3.7° N	4.3° W	15:45	57	5.4	254
Mariner 7 entry	58.2° S	30.3° E	14:30	56	5.0	221
Mariner 4 exit	60.0° N	34.0° W	00:30	104	7.6	(240)
Mariner 6 exit	79.3° N	87.1° E	22:10	107	8.5	152
Mariner 7 exit	38.1° N	148.3° W	03:10	130	8.0	209

undergone any substantial modification since its early history, there do exist on Mars regions where evidence for some surface activity is overwhelming. Mars may not be as "dead" as the Moon.

Mariner 7 extensively photographed the southern polar cap. Figure 4 is an extraordinary view of the cap as seen from above the south pole. Craters are visible throughout the region, indicating that the material that covers the cap may be only a few meters thick. Also, an infrared radiometer carried by Mariner 7 made precise measurements of the surface temperatures on Mars at different locations, including the south polar cap. Here, the temperature was found to be 148 ± 2 K, which is exactly the condensation temperature of CO_2, implying that the material covering the two poles is largely solid CO_2.

The successful occultation experiments on both spacecraft also provided new information regarding the temperature structure of the atmosphere. Temperature and pressure profiles are now available at four different points on the planet, two on the dayside and two on the nightside (Table 3 and Figure 5). The results are close to the theoretical predictions of several authors who had suggested that the day-to-night and equator-to-pole variations in surface temperatures should be large because of the thinness of the atmosphere. In the atmosphere, the temperature decreases adiabatically with the altitude at the equator, whereas at the pole the atmosphere seems to be extremely cold and isothermal.

In summary, therefore, Mars appears to be a cold, dry planet, its surface a heavily cratered terrain dating back billions of years, and its atmosphere a thin

Figure 4.—Southern polar cap of Mars as photographed by Mariner 7.

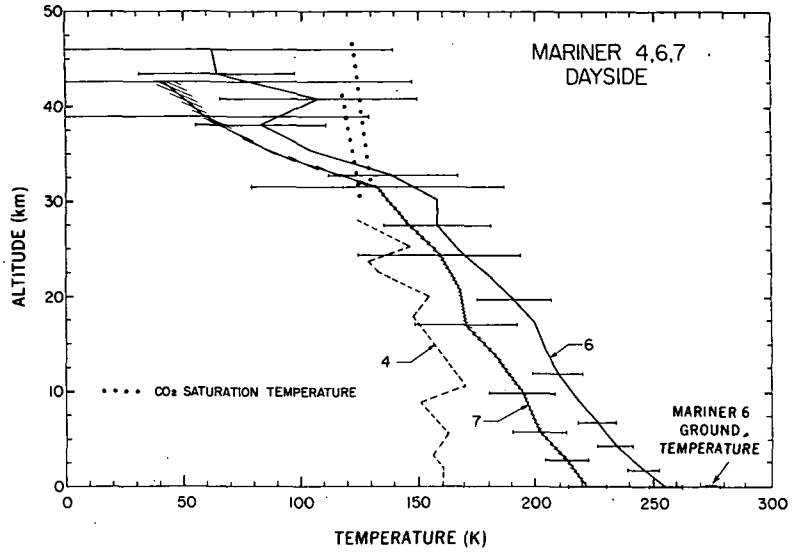


Figure 5a.–Dayside temperature profile in the middle atmosphere of Mars.

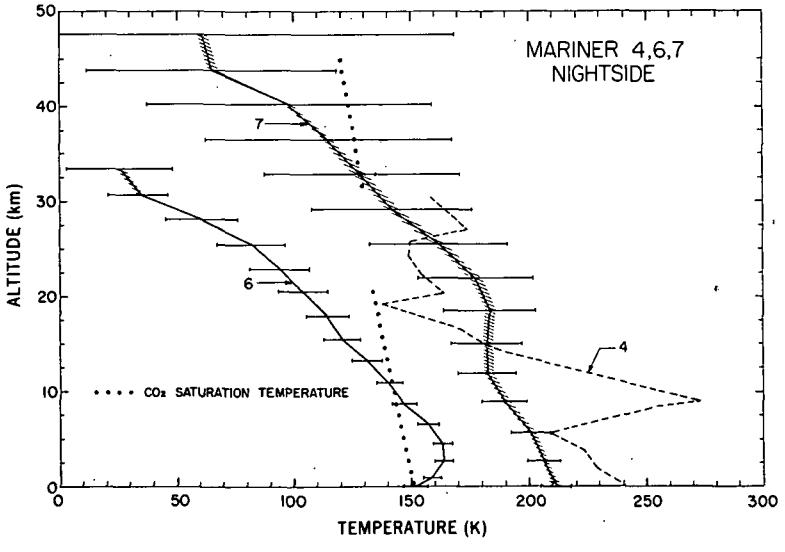


Figure 5b.–Nightside temperature profile in the lower atmosphere of Mars.

envelope of almost pure CO_2. The important question that emerges from this description of the surface and atmosphere of Mars is whether the new data on the planetary environment have increased or decreased the chances that life exists or has ever existed on Mars.

The answer is fairly straightforward. The new and more ample data from Mariners 6 and 7 have certainly decreased the likelihood that Martian life ever reached the high state of evolution on the Earth. The TV pictures show that most of the surface of Mars is covered with old craters, and there is no evidence in these pictures that the terrain has been modified by large-scale tectonic activity or that liquid oceans once covered the planetary surface.

Another discovery from the recent Mariner mission that is relevant to the problem of life on Mars is that the polar caps are covered with CO_2 ice rather than H_2O ice. Both of these observations force us to conclude that oceans of water have been absent on Mars from the very beginning. On the Earth, oceans have played a vital role in the chemical and biological evolution of life. It is now generally accepted that chemical evolution (i.e., the synthesis of such organic compounds as amino acids and nucleotides, which are the building blocks of all forms of life) essentially took place in the oceans. The random stirring and mixing of what is called the "primordial soup" over millions of years is the only way the simple organic compounds could have joined together to form the complex chains of proteins, DNA, and RNA. Without an enclosed fluid medium like the oceans, the probability that living cells would evolve from simple inorganic compounds seems to be near zero. For this reason, the new evidence of the absence of oceans on Mars during the major part of its history strongly implies that life could not have gone through all the critical stages of evolution as it did on the Earth.

However, the possibility still remains that a *partial* evolution toward life did take place on Mars during the brief period in its early history when oceans might have existed. Perhaps the first few steps along the path of chemical evolution were taken, and then the process was stopped when, for some reason, the water disappeared. If this is confirmed by future analysis of the surface material on Mars, it will have the greatest impact on human philosophy. It will mean that given the proper conditions, evolution toward life can take place independently on a planet. Since there are certainly billions of planets in our galaxy alone, the probability of existence of "intelligent" life elsewhere in our galaxy will suddenly increase a millionfold. On the other hand, a negative result on Mars will only mean that we must look elsewhere in the solar system for the existence of extraterrestrial life.

IV. VENUS

Venus, the closest planet to the Earth, is the third brightest object in the sky. It has been observed and studied for centuries and yields only to the Sun and the Moon in attracting attention. Yet, for all its popular appeal, it has until recently remained an enigma to the astronomers—a planet shrouded in mysteries.

The main reason for this lack of information has been that the planet Venus is permanently enveloped by a veil of clouds, and consequently no surface features have ever been observed, even through the most powerful telescopes. On October 18, 1967, however, a Russian spacecraft, Venera 4, succeeded for the first time in the long history of planetary research in penetrating the veil and unravelling many of the mysteries of the Venusian atmosphere. Only a day later, on October 19, an American probe, Mariner 5, flew past Venus at a distance of 4100 km and explored the atmosphere of the planet by radio waves, obtaining precise values of the distribution of atmospheric density with height up to an altitude of several hundred kilometers. After combining the results of these two entirely different types of measurements, we now possess more information on the atmosphere of Venus than was available on the atmosphere of the Earth 25 years ago.

The two spacecraft measured the temperature in different ways, but their data led to the same conclusion: The surface of Venus is extremely hot (700 K), and there is no reasonable chance of finding life on its surface.

How were these temperature measurements made, and how sure are we of their accuracy? Venera 4 contained a simple thermometer, a barometer, and several gas analyzers which made direct measurements of temperature, pressure, and composition at different points in the atmosphere as the capsule descended toward the surface. The first measurement was made at a point where the temperature in the atmosphere was 300 K. At this point, gas analyzers indicated that the atmosphere was composed mainly of CO_2 and contained only very small amounts of water vapor (0.1 percent). At approximately the same altitude, a parachute opened, and the spacecraft began its slow descent toward the surface. Temperature was measured at frequent intervals throughout the ensuing parachute descent of the capsule and was found to be increasing at a rate of 10 K every kilometer. At the point where the temperature had attained a value of 550 K and the pressure about 20 times that of the Earth's atmosphere, the signals from the spacecraft stopped abruptly. It was first believed that the signals ceased suddenly because the capsule had reached the surface of Venus; however, a critical analysis of these data, in comparison with that of Mariner 5, subsequently indicated that at the time when signals

from Venera 4 stopped, the capsule was still about 16 miles above the surface. The temperature and pressure values at the ground were therefore probably much higher than those indicated.

Although Mariner 5 did not send a probe to the surface of Venus, it acquired information on conditions deep in the atmosphere by a radio propagation experiment. This experiment depended on the fact that the Mariner 5 trajectory carried the spacecraft behind Venus. As a result, signals from the spacecraft to the Earth were blocked for a period of about 20 minutes. Prior to the passage of the spacecraft behind Venus, and again just after its emergence on the other side of the planet, the signals from Mariner 5 traversed the Venus atmosphere en route to the Earth. As the beam penetrated the atmosphere, refraction caused the path of propagation to deviate from a straight line and the velocity of propagation to vary from the speed of light in free space. In addition, because the density of the atmosphere decreases with altitude, the power in the beam was spread over a greater angular width, which caused the signal power received at Earth to decrease. During this period, these effects were observed as changes in the frequency, phase, and signal strength received at the tracking stations on Earth. The sign of the phase change differed for passage through the neutral atmosphere and through the charged-particle ionosphere, respectively; hence, it was possible to construct separate density profiles for the atmosphere and the ionosphere. From the rate of change of density with altitude, it was possible to deduce the temperature and pressure profiles of the atmosphere if the composition was known. As mentioned earlier, the Soviet probe provided this last data: The Venus atmosphere is 90 percent CO_2. With this additional information, temperature and pressure distributions in the atmosphere of Venus could be obtained from the Mariner 5 data.

This method should, in principle, determine atmospheric conditions down to the surface of the planet. However, because of the high density of the Venus atmosphere, the radio signals from the spacecraft, as they passed through the lowest layers of the atmosphere, were bent or refracted through such a large angle that they never reached the Earth. At the lowest depth probed, the atmospheric temperature was 480 K and pressure about 7 atm, and both were still increasing. Conditions at and near the surface were not determined by this experiment.

However, more recently, Veneras 5 and 6 (identical to Venera 4) probed the atmosphere of Venus to lower depths and determined the conditions down to a level where the temperature was 600 K and the pressure approximately 26 atm. It has now been established that even this level in the atmosphere is at least 10 km above the surface, and the temperature and pressure, if they continue to increase

with depth at the same rate as above, will reach values as high as 720 K and 100 atm at the surface (Figure 6).

These extremely rigorous conditions on the surface of Venus—a sizzling 700-K temperature, and a heavy atmosphere 100 times more massive than that of the Earth and composed mainly of noxious carbon dioxide instead of nitrogen and life-sustaining oxygen as on Earth—are most intriguing for a student of planetary evolution. Still more puzzling is the fact that no more than 0.1 percent of water vapor was observed in the lower atmosphere of Venus. If condensed to a liquid, this amount of water would cover the surface of Venus to a depth of only 10 cm. The amount of water in the oceans of the Earth, however, is much greater: If spread uniformly over the surface of the globe, it would form a layer about 3 km deep. According to current theories on the origin of planets, Venus and Earth condensed out of the same homogeneous material about 4.5 billion years ago. If so, Venus' surface should also be covered by an ocean 3 km deep; but because of the 700-K temperature of the surface, this amount of water should be present in the form of steam rather than liquid. Venera 4 showed that most of this water is missing.

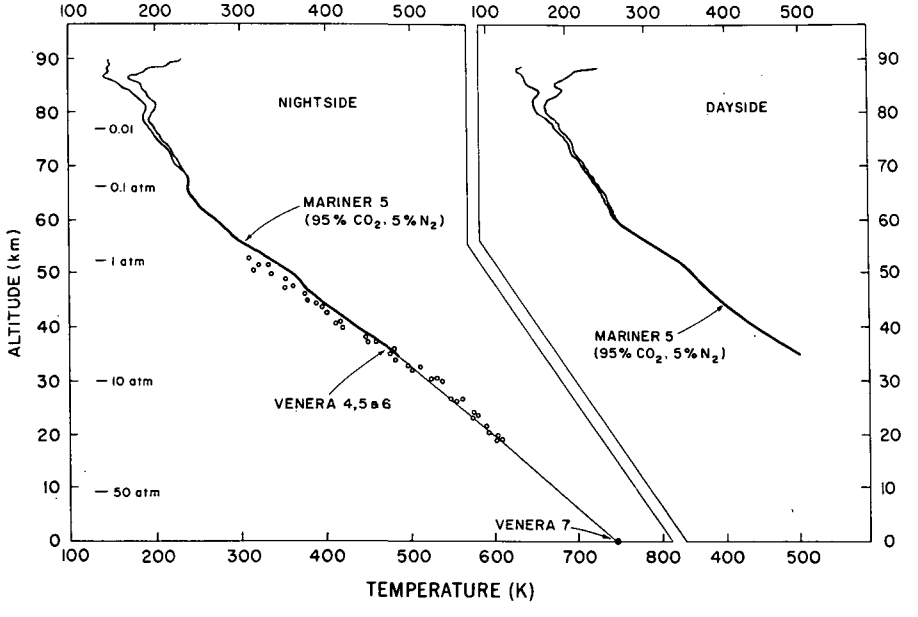


Figure 6.—Temperature distribution in the lower atmosphere of Venus as a function of altitude from actual measurements by Mariner 5 (solid lines) and the three Soviet probes, Veneras 4, 5, and 6 (circles).

Why did two planets, identical in size and weight, probably formed at the same time out of similar materials, and situated at comparable distances from the Sun, evolve along different paths? Why does one planet offer an excellent climate for life, whereas the other offers conditions hostile to terrestrial organisms?

The answer to this problem requires the resolution of three important questions: Why is the surface of Venus so hot? Why does the atmosphere contain such large quantities of carbon dioxide instead of N_2 and O_2? What happened to the oceans of water on Venus?

The measurement of the abundance of carbon dioxide on Venus is of great interest because it helps answer the first question: Why is Venus so hot? According to information radioed back to Earth from the spacecraft, the atmosphere of Venus consists primarily of a heavy layer of carbon dioxide, about 70 000 times more than is in the atmosphere of the Earth. The dense atmosphere of carbon dioxide acts as an insulating blanket that seals in the planet's heat and prevents it from escaping into space. The trapped heat raises the surface temperature to a far higher value than it would have otherwise. This effect is caused by the absorbing properties of a planetary atmosphere and can best be explained by taking an example of the Earth.

The solar radiation reaching the Earth has a value of 2 cal-cm^{-2}-min^{-1}. Part of this energy (about 30 percent), however, is directly reflected back to space by the clouds present in the Earth's atmosphere. Only 70 percent of the solar energy, therefore, penetrates through the atmosphere and reaches the surface. This energy is sufficient to heat the ground to a temperature of only 255 K. The Earth's surface itself radiates, but since the temperature is not too high, the radiation is in the far-infrared region of the electromagnetic spectrum—a dimly glowing object. The atmosphere of the Earth, however, contains small quantities of water vapor, carbon dioxide, and ozone. These gases have a property of absorbing the far-infrared radiation with great efficiency. As a matter of fact, only 10 percent of the infrared actually gets through the Earth's atmosphere. The atmosphere, having absorbed all this radiation, radiates in all directions, partly toward the surface and partly toward space. The radiation toward the surface increases the ground temperature by about 30 K, from 255 K to the observed value of 288 K. This phenomenon is called the greenhouse effect of the atmosphere, being an allusion to a greenhouse where the glass cover acts like the atmosphere, transparent to the solar radiation but opaque to the radiation from the interior.

On Venus, a greenhouse effect of about 500 K is required to explain the observed temperature. Calculations based on the insulating properties of carbon dioxide show that the temperature of Venus could easily be raised by 500 K as a

result of the greenhouse effect of a very heavy layer of CO_2, as was reported to be present on Venus.

The high temperature on the surface of Venus can, therefore, be understood from the abundance of carbon dioxide in the atmosphere. The question one then asks is why carbon dioxide is so abundant on both Venus and Mars and not on the Earth. This problem is related to the evolutionary history of the three planets, and in the following section, we will attempt to discuss it in more detail.

V. EARLY HISTORY OF THE ATMOSPHERES OF VENUS AND MARS

Table 4 summarizes the surface parameters and the abundances of major volatiles on Venus, the Earth, and Mars. Several interesting features are noteworthy. First, the amount of CO_2 in the atmosphere of Venus is approximately equal to the amount buried in the crust of the Earth in the form of carbonates. Second, nitrogen, which is at the present time the major constituent of the Earth's atmosphere, is only 1 to 5 percent of the total amount of CO_2, not only on Earth but also on Mars and Venus. Third, the large quantities of water that make up the oceans of the Earth are practically absent on Venus but may be present in a frozen state below or on the surface of Mars. Fourth, the amount of free oxygen in the atmosphere of the Earth is small in relation to the total amount of oxygen in the crust, but it is even less abundant in the Venusian and Martian atmospheres. Finally, comparison of the abundances of the four gases in the atmosphere of Mars with those of Venus and the

Table 4.—Surface temperatures and pressures and surface and atmospheric abundances on Venus, the Earth, and Mars.

Planet	Temperature (K)	Pressure (atm)	Abundance ($g\text{-}cm^{-2}$)							
			Oxygen		Water		Nitrogen		Carbon Dioxide	
			Crust	Atmosphere	Oceans	Atmosphere	Crust	Atmosphere	Crust	Atmosphere
Venus	700	75	?	<10	0	≈100	?	<3000	?	70 000
Earth	300	1	8×10^6 (total)	200	300 000	≈1	2000(?)	800	70 000	≈1
Mars	230	0.01	?	≈0.01	?	≈0.01	?	<1	?	≈70

Earth suggests that if the origin of these gases is the same (the outgassing from the interior), the volcanic activity on Mars has been about 1000 times less effective than that on either the Earth or Venus.

In order to investigate the possible evolutionary paths which the atmospheres of Venus, the Earth, and Mars may have followed to arrive at their present diversity in composition, one can make three basic assumptions: (1) At some time in the early history of the solar system, the terrestrial planets (Mercury, Venus, the Earth, and Mars) completely lost their primordial atmospheres; (2) the present atmospheres of the latter three of these planets (Table 4) have developed mostly from the degassing of the planetary interiors (free oxygen in the Earth's atmosphere being an exception related to the presence of life); and (3) the major constituents of the outgassing from the interior for all three planets are essentially the same (i.e., water vapor and carbon dioxide, with $H_2O/CO_2 = 4$ and with nitrogen accounting for less than 1 percent of the volcanic emanations).

At the beginning of the outgassing, when the planets are more or less devoid of an atmosphere, their ground temperatures T_G are essentially determined by the amount of solar radiation absorbed by the surface and are equal to the effective temperatures of the planet T_e. For a rapidly rotating planet and for a given planetary albedo A_p, we have

$$T_G = T_e = \frac{S_p}{4\sigma}\,(1\text{-}A_p),$$

where S_p is the solar constant.

As the atmosphere accumulates, the ground temperature begins to exceed the effective temperature because of the additional heating of the surface by the atmospheric greenhouse effect. The magnitude of the greenhouse effect can be calculated by solving the equation of radiative transfer for a given atmosphere. However, the transfer problem becomes quite complicated when allowances are made for the strong frequency dependence of the infrared molecular absorption coefficient, especially for a gas such as water vapor. Therefore, for the first estimates of ground temperature, we assume a gray atmosphere and use the Eddington approximation for the solution of the radiative transfer equation.

We start the greenhouse calculations for Venus at time zero when the outgassing has just commenced and the total amount of atmosphere is less than 10^{-3} mb. At this point we assume that planetary albedo is determined only by the surface, and, by analogy with the Moon and Mercury, we assume it to be equal to 7

percent. The average ground temperature at this time is equal to T_e, and for an albedo of 7 percent, it will be 330 K for fast rotation and 390 K for slow rotation. As the atmosphere accumulates, the greenhouse effect increases the ground temperature, provided the albedo remains constant. Because the initial surface temperature T_e is already quite high, H_2O remains in the atmosphere as vapor and accelerates the greenhouse effect. The increase in ground temperature as a function of the buildup of H_2O-CO_2 pressure in the atmosphere is shown in Figure 7. Each set of curves indicates the range of possible temperatures, which stems from the uncertainties in the greenhouse calculations and the effect of convection. However, in the case of Venus, by the time the atmosphere has accumulated to a total pressure of 10^{-1} atm, the ground temperature has already risen to 430 K for T_e = 330 K and is greater than 500 K for T_e = 390 K. One important implication of this result is that the temperature on the surface of Venus always remained above the boiling point of water at the pressures involved. This is illustrated by the position of the phase diagram of water in Figure 7. The temperature curves for Venus miss the liquid phase of water by a wide margin. It may be noted here that the calculated Venus temperatures are so much higher than the boiling point of water that even if the albedo of the planet were 30 to 40 percent instead of 7 percent, water would not condense at the surface.

These initial high temperatures and the complete absence of liquid water from the surface had a significant bearing on the accumulation of CO_2 in the atmosphere of Venus. As the CO_2 is degassed from the interior into the atmosphere, its partial pressure should be buffered by reactions with the crust, such as

$$CaMgSi_2O_6 + CO_2 = MgSiO_3 + CaCO_3 + SiO_2.$$

These atmosphere-crust reactions are, however, temperature dependent, and at high temperatures, large quantities of CO_2 can accumulate in equilibrium with the silicates. Figure 7 also shows the plots of CO_2 partial pressures as a function of temperature for reactions with $CaSiO_3$, $CaMgSi_2O_6$, and $MgSiO_3$. If at any time the amount of CO_2 in the atmosphere is higher than the equilibrium value at that temperature, the situation is unstable, and CO_2 should be removed from the atmosphere and deposited into the crust as carbonates. However, the atmosphere-crust reaction will proceed at a rapid rate *only* if liquid water is present at the surface to facilitate the contact. These conditions should be satisfied simultaneously in order to remove CO_2 from the atmosphere effectively. In the case of Venus, the

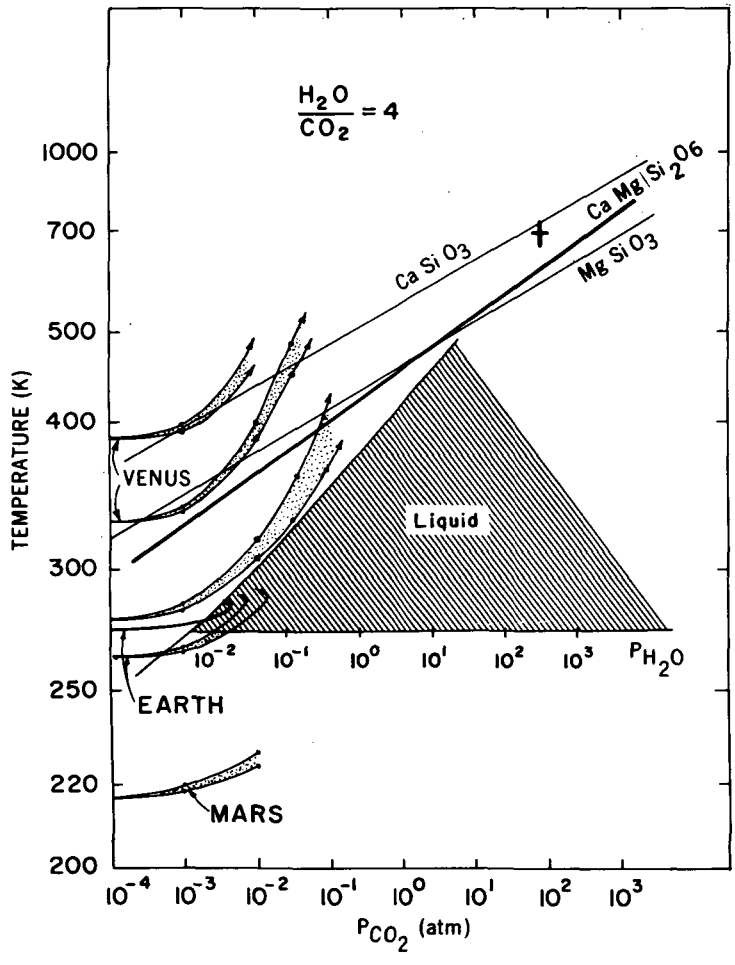


Figure 7.–Plot of increase in surface temperatures on Venus, the Earth, and Mars caused by the greenhouse effect of an H_2O-CO_2 atmosphere during the evolution of the three planets. The initial temperatures on these planets equal the effective temperatures for a planetary albedo of 7 percent for Mars, for 7 and 20 percent for the Earth, and for 7 percent with two different rates of rotation for Venus. The phase diagram for water is shown, and the region in which water can exist as a liquid is represented by the hatched area. Also plotted are the equilibrium values for the partial pressures of CO_2 as a function of temperature for three different silicate reactions. (After Rasool and de Bergh, 1970.)

temperature was always high enough that the amount of CO_2 in the atmosphere never substantially exceeded the equilibrium pressure at that temperature. The only exception is for the case of T_e = 330 K and the reaction with $CaSiO_3$. However, the absence of liquid water would have impeded the reaction from proceeding rapidly, and the temperature should have soon built up to a value at which the atmospheric CO_2 would be in equilibrium with the crust. In this way, the CO_2 continued to accumulate in the atmosphere to the present value of approximately 75 atm at a temperature of 700 K.

The important question which still remains is the absence of large quantities of water vapor in the present atmosphere of Venus. According to this model, the Venus atmosphere should contain approximately 300 atm of H_2O. However, only about 10^{-1} atm of water appears to be present today (Table 4). As was mentioned earlier, most of the water could have escaped from the Venusian atmosphere in the early stages when its exospheric temperature would have been quite high (>3000 K), and the escape of hydrogen and even oxygen would have been very rapid.

If the present amount of H_2O in the Venus atmosphere ($\approx 10^{-1}$ atm) is the equilibrium value between the outgassing and loss of hydrogen to space, it would imply that the escape of water from Venus commenced at $P_{H_2O} \approx 10^{-1}$ atm. At this point, however, according to Figure 7, the surface temperature is already greater than 430 K, and CO_2 in the atmosphere is in equilibrium with the silicates. From this time to the present, if the partial pressure of water never exceeded 10^{-1} atm, the increase in the surface temperature to the present value of 700 K would have been principally governed by the greenhouse effect of CO_2 alone. Consequently, the temperature would increase at a slower rate than is shown by the arrows in Figure 7 and would probably follow the CO_2 equilibrium curve for $CaSiO_3$.

When similar calculations are carried out for the Earth, a completely different evolutionary pattern emerges, explaining why the temperature on the surface of the Earth remains a comfortable 290 K and why almost all of the CO_2 is in the crust and H_2O is in the oceans. For the Earth, the initial value of T_e for an albedo of 7 percent is 275 K. As the water vapor and CO_2 atmosphere begins to accumulate on the Earth, the ground temperature increases, as on Venus, but soon temperature and pressure conditions become such that liquid water can condense at the surface. (This is shown in Figure 7 where the curve for Earth enters the hatched area for the liquid-water phase.) From this point on, the Earth follows an evolutionary path completely different from that of Venus. With the volcanic steam condensing into liquid, the amount of water vapor remaining in the atmosphere is small, and the increase in ground temperature is determined by the accumulation of CO_2 alone.

However, the total amount of CO_2 soon becomes greater than the equilibrium value with the silicates. The reactions with crustal rocks proceed rapidly and are considerably accelerated by the presence of liquid water on the surface. Due to the erosion of rocks by liquid water, fresh silicates are brought into contact with the CO_2 in the atmosphere and in the oceans. The CO_2 therefore never accumulates in excess of the equilibrium value of about 10^{-4} atm at 290 K. At the same time, the volcanic steam continues to condense at the surface, slowly building up the oceans to their present depth of approximately 3 km. Nitrogen, the inert gas which constitutes only about 1 percent of the volcanic gases, accumulates to make up the bulk of the present atmosphere.

This chain of events in the case of the Earth began only because T_e was 275 K and because water vapor, the major constituent of volcanic emanations, was able to condense out of the atmosphere to mark the beginnings of the oceans. As is clearly evident from Figure 7, the initial temperature of the planet is an extremely critical parameter in determining which evolutionary path a planetary atmosphere will follow. In fact, a calculation for an initial temperature of 280 K indicates that the surface temperature increase would have been rapid enough for the Earth to miss the liquid phase of water at its surface. A runaway greenhouse effect would have made the conditions on the Earth as hostile as on Venus. This situation could have occurred on the Earth if it were closer to the Sun by only 6 to 10 million kilometers.

However, it is interesting to note that for a planet further away from the Sun, like Mars, the chances are very meager that a runaway greenhouse would ever take place. When the initial temperature of a planet is less than 273 K, the volcanic steam freezes at the surface, and only CO_2 accumulates in the atmosphere (as is the case on Mars today). However, when large quantities of CO_2 have accumulated on the planet, the greenhouse effect due to CO_2 alone will raise the surface temperature above 273 K, melting the frozen water and thereby initiating the transfer of CO_2 from the atmosphere into the crust. When this evolutionary stage is reached on Mars, conditions on the surface may become very similar to those on the Earth today: water in a liquid state, CO_2 in the sediments, and an atmosphere consisting mainly of N_2.

This evolutionary path for Mars is completely different from the usual suggestions that Mars, in its early history, had oceans and a heavier atmosphere of which the present one is a remnant. According to the present model of the evolution, oceans have been absent on Mars from the very beginning but may

eventually accumulate by melting when the atmospheric pressure becomes large enough that the greenhouse effect raises the temperature of the planet above 273 K.

VI. JUPITER

The recent findings on Mars and Venus discussed earlier mark the beginning of the accumulation of basic data for the understanding of the history of the terrestrial planets. However, to resolve the age-old problem of the origin and evolution of the solar system as a whole, it is the exploration of Jupiter that will eventually provide information of prime significance. This is so not only because Jupiter is the largest planet—several times more massive than the other eight planets combined—but also because it presents such puzzling aspects of far-reaching importance that their eventual solution will have direct bearing on our understanding of the primitive environments from which the planets were formed and the life on Earth originated.

Perhaps the most interesting aspect of Jupiter is that its present atmosphere seems to be composed of the same gases—hydrogen, methane, and ammonia—out of which the first living organisms are believed to have been synthesized on the Earth about 4 billion years ago. Is it possible that similar initial steps along the path of life are occurring now on Jupiter? To answer this question, one needs to know the exact composition, temperature, and pressure conditions that exist at various levels in the atmosphere and at the surface. This lead us to the other puzzling aspect of Jupiter.

It has recently been found that Jupiter may have a source of heat in the interior almost four times more intense than the Sun at that distance. What is the source of this energy? Is it that Jupiter is still contracting toward its final size and is thereby releasing gravitational energy? If so, what about the other giant planets? Are they all, 4.5 billion years after their birth, still in the process of accumulation and do not yet have a solid surface?

A. Composition

Jupiter and the other giant planets differ markedly from the terrestrial planets in regard to their composition, both in their interiors and in their atmospheres. Though they are much larger and hundreds of times more massive than the Earth, they have surprisingly low density. In general, the density is about that of water. For

Jupiter, it is 1.33 g/cm^3, and for Saturn only 0.71 g/cm^3. This should be compared with the density of the Earth, which is 5.5 g/cm^3. The low density of the giant planets is puzzling because the pressure of their great mass should compact them to a higher density than that of the Earth. The explanation of this apparent paradox is connected with the composition of the giant planets. In contrast to the terrestrial planets which are made up of Fe, Ni, and silicates, the major planet seems to be composed mainly of hydrogen and helium, the lightest of all the elements. In fact, the density of Jupiter is almost exactly the same as that of the Sun, indicating that the ratio of hydrogen and helium to other heavier elements may be about the same on Jupiter as on the Sun. This is what one would expect if Jupiter condensed out of a contracting solar nebula which had the same composition as the Sun has today. Jupiter, being so massive, did not lose any of the bases during its long history and should therefore reflect the composition of the material out of which it was formed (Table 1).

Spectroscopic observations of Jupiter have already determined the presence of methane, ammonia, and hydrogen above the clouds of Jupiter, and it is believed that substantial amounts of water are present below. Methane and ammonia are easy to detect because of their strong absorption bands in the near infrared, and careful analysis of the band structure can also give the concentration of these gases in the atmosphere. As early as 20 years ago, Kuiper (1952) successfully measured the amounts of methane and ammonia on Jupiter, and their concentration appears to be roughly the same as mentioned above.

Hydrogen and helium, however, pose special problems. Neither of them produces absorption bands in the far infrared, as do methane and ammonia. Helium, being a rare gas, is completely inert and under ordinary conditions cannot be detected by spectroscopic techniques employed in optical astronomy. It does, however, produce emission lines in the ultraviolet which can be observed from the Earth only if the measurements are made from above the atmosphere. Hydrogen, on the other hand, under special conditions produces absorption lines in the visible part of the spectrum. Molecular hydrogen has a quadrupole moment and could produce a vibration-rotation spectrum which could be detected from the Earth if sufficiently large amounts were present on Jupiter. C. C. Kiess, C. H. Corliss, and H. K. Kiess (1960) were the first to detect four such lines of hydrogen at around 8200Å. A comparison of their strengths with the theoretical values of their intensities can give an estimate of the amount of hydrogen on Jupiter above the reflecting level of 8200Å photons. Because of several inherent problems in this technique of measurement, the estimate of hydrogen on Jupiter cannot be made with a precision

greater than a factor of 3. A recent evaluation of this problem suggests that at a level in the atmosphere where hydrogen exerts a pressure of 1.8 atm, the total pressure is between 2 and 2.8 atm. If the other gas is helium, its abundance may be 6 to 30 percent, close to the computed solar composition.

B. Thermal Structure

Situated at 5.2 AU from the Sun, Jupiter receives 27 times less solar energy than the Earth: only 12 500 erg-cm^{-2}-s^{-1} distributed over the planetary globe. In addition, the albedo of Jupiter is 0.45, and, therefore, only 55 percent of this energy, or 7000 erg-cm^{-2}-s^{-1}, actually enters the planet, all of which should eventually be returned to space in the form of thermal radiation. The recent measurements of this energy, however, suggest that the flux emitted by Jupiter is of the order of 30 000 erg-cm^{-2}-s^{-1}, about four times higher than the expected value. How accurate are these measurements, and what could be the source of this excess energy?

If it is assumed that Jupiter behaves as a blackbody, the effective radiating temperature for the amount of solar energy reaching the planet will be 103 K. At this low temperature, the radiation will be mainly at wavelengths between 10 and 100 μm, with a maximum at 30 μm. Measurement of this radiation is difficult to make from the ground because of the absorbing properties of the Earth's atmosphere. The Earth's atmosphere contains molecules of CO_2 and H_2O which absorb the far infrared with great efficiency. There are, however, "windows" at $\lambda < 20$ μm where the atmospheric absorption is small, which allows ground-based astronomers to measure energy from different celestial sources. Depending on the gases present in the atmosphere of Jupiter, such measurements made at different wavelengths will provide information on temperature at varying depths in the atmosphere. Figure 8 shows a recent attempt in this direction.

On the basis of these measurements, one can derive a temperature profile of the atmosphere above the clouds, and several such models are shown in Figure 9. The best fit to the data is model 3, which corresponds to a distribution of gases in the atmosphere (shown in Figure 10). Regarding the structure of the Jovian atmosphere below the visible clouds, the most comprehensive study to date is by Lewis (1969), who has carried out thermodynamical calculations for an NH_3-H_2O-H_2S system. The temperature is assumed to increase with depth adiabatically ($\approx$2 K-km^{-1}). As the temperature increases, ammonia and water

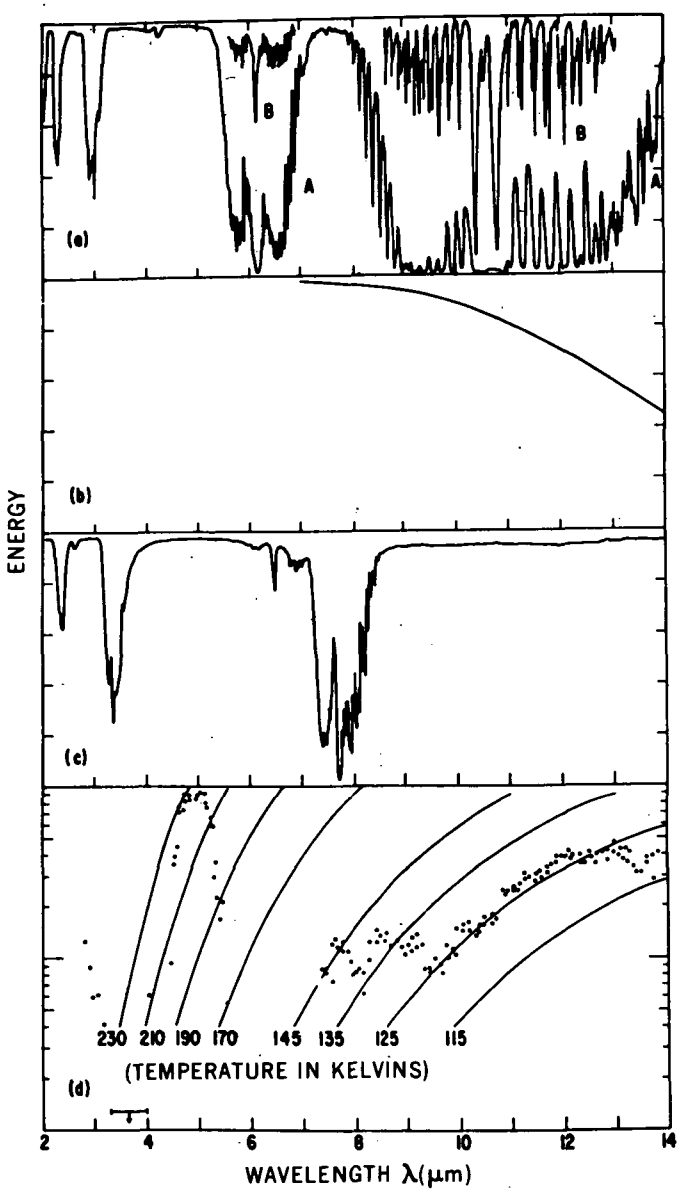


Figure 8.—(a) Absorption spectra of ammonia from 2 to 14 μm for two different values of pressure (A and B); (b) absorption spectrum of hydrogen from 8 to 14 μm for a pressure of 33 atm and a temperature of 85 K; (c) room-temperature absorption spectrum of methane from 2 to 14 μm; and (d) 2.8- to 14- μm absorption spectrum of Jupiter. (After Gillett, Low, and Stein, 1969.)

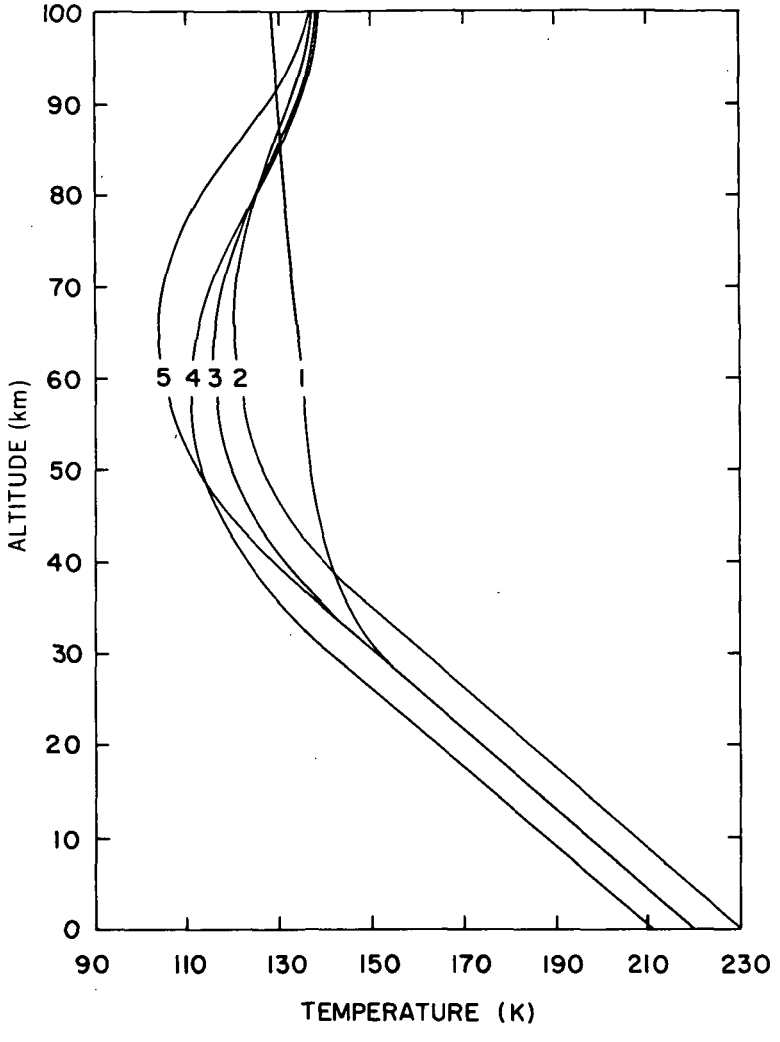


Figure 9.–Temperature distribution in the atmosphere of Jupiter above the cloud levels for five different models. Model 3 agrees best with the data shown in Figure 8. (After Hogan, Rasool, and Encrenaz, 1969.)

become important constituents. Several cloud layers are formed, the topmost (at $T = 150$ K) being composed of NH_3 crystals, with lower ones composed of NH_4SH and an aqueous NH_3 solution at temperature levels of 220 and 310 K, respectively (Figure 11).

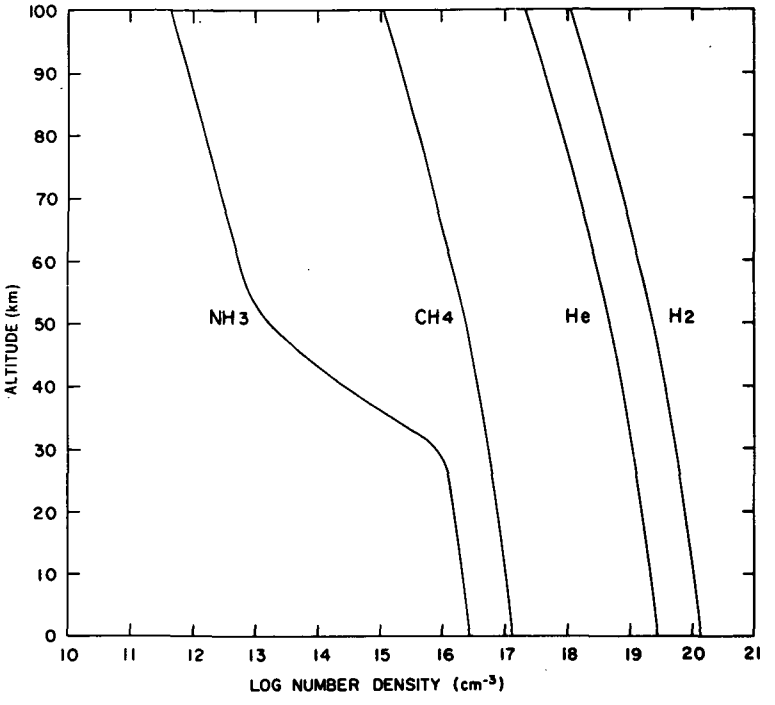


Figure 10.–Density distributions for H_2, He, CH_4, and NH_3, corresponding to thermal model 3 shown in Figure 9. (After Hogan, Rasool, and Encrenaz, 1969.)

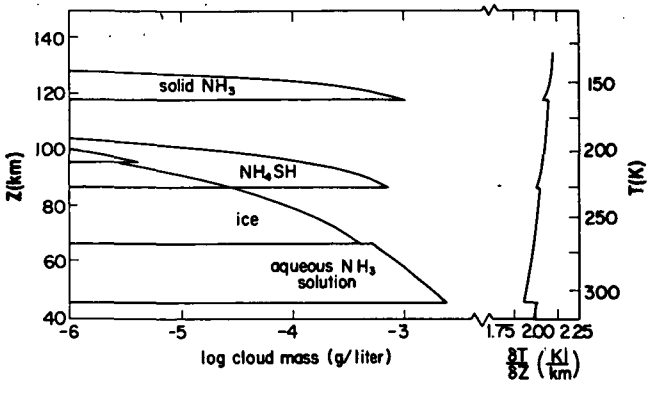


Figure 11.–Cloud masses and wet adiabatic lapse rate vs. altitude and temperature, respectively, for NH_3-H_2O and NH_3-H_2S clouds in a solar-composition model of the Jovian atmosphere. The prominence of the NH_4SH and aqueous NH_3 clouds is noteworthy. (After Lewis, 1969.)

C. Color and Life

The orange-and-blue coloration of the cloud bands of Jupiter and the presence of a Giant Red Spot have long fascinated the optical astronomers. A recent, new interpretation of these features has made the problem all the more exciting. It has been proposed that the visible, colorful "surface" of Jupiter is the seat of intense prebiological activity, where the first living organisms are being synthesized.

Three arguments have been advanced in favor of this hypothesis. First, the atmosphere of Jupiter is composed precisely of those gases (hydrogen, methane, ammonia, and water vapor) that have supposedly played a critical role in the events that have led to the development of life on the Earth.

Second, Ponnamperuma and Woeller (1969) have demonstrated that when a simulated Jovian atmosphere, at temperatures as low as 150 K, is exposed to ultraviolet radiation, it not only produces complex organic molecules like amino acids and nucleotides, but the color of the resulting products is very similar to the yellowish-orange red of the Jovian clouds and of the famous Red Spot.

Third, an ultraviolet spectrum of Jupiter, first taken by T. P. Stecher (1965) from a rocket, indicated a significant absorption centered at $\lambda = 2600$Å. None of the known major atmospheric constituents of Jupiter can account for this feature. However, C. Sagan et al. (1967) have pointed out that the absorption features observed on Jupiter match closely with those of adenine, which is a basic constituent of both RNA and DNA and, therefore, one of the most important chemicals in biological systems. In fact, laboratory experiments of Ponnamperuma and Woeller (1969) have convincingly shown that in electron irradiation of methane, ammonia, and water, the largest single nonvolatile compound formed is adenine. In addition, the production of adenine is enhanced when hydrogen is deficient, as it seems to be at the cloudtops of Jupiter.

How can one test this extremely interesting hypothesis? First and foremost in this respect is, of course, the ultraviolet spectroscopy of Jupiter from rockets and Earth-orbiting satellites. High-resolution spectra of Jupiter in the 1800Å to 3000Å interval, when matched with laboratory spectra of organic molecules, as suggested by Sagan et al. (1965), should provide highly significant information on this problem. Subsequent experiments in the infrared and a search for HCN by close-range flyby missions should further clarify the question. The actual resolution of the problem will probably have to await *in situ* exploration of the Jovian atmosphere or return of the samples to the Earth. It is difficult to predict the timetable involved for such experiments. Because of the long travel to Jupiter and

the crushing force of gravity of the planet, the *in situ* exploration and return of samples is perhaps another 20 years away, but numerous experiments of great value can certainly be performed from high-altitude aircraft, balloons, rockets, and Earth-orbiting satellites as early as next year.

REFERENCES

Abelson, P. H., *Proc. Nat. Acad. Sci. U.S.A.* 55:1365, 1966.

Cameron, A. G. W., "A New Table of Abundances of the Elements in the Solar System", in *Origin and Distribution of the Elements,* L. H. Ahrens, ed., Pergamon Press, Inc., New York, 1968, pp. 125-143.

Gillett, F. C., Low, F. J., and Stein, W. A., "The 2.8-14μ spectrum of Jupiter", *Astrophys. J.* 157:925-934, 1969.

Hogan, J. S., Rasool, S. I., and Encrenaz, T., "The Thermal Structure of the Jovian Atmosphere", *J. Atmos. Sci.* 26:898-905, 1969.

Holland, H. D., in *Petrologic Studies: A Volume in Honor of A. F. Buddington,* A. E. Engel, H. James, and B. F. Leonard, eds., Geological Society of America, New York, 1962, p. 447.

Kiess, C. C., Corliss, C. H., and Kiess, H. K., "High Dispersion Spectra of Jupiter", *Astrophys. J.* 132:221, 1960.

Kuiper, G. P., "Planetary Atmospheres and Their Origin", in *The Atmospheres of the Earth and Planets,* G. P. Kuiper, ed., University of Chicago Press, Chicago, 1952, Chap. 12.

Lewis, J. S., "Geochemistry of the Volatile Elements on Venus", *Icarus* 11:367-385, 1969.

McGovern, W. E., "The Primitive Earth: Thermal Models of the Upper Atmosphere for a Methane-Dominated Environment", *J. Atmos. Sci.* 26:623-635, 1969.

Mason, B., *Principles of Geochemistry,* John Wiley & Sons, New York, 1966, 3rd ed.

Miller, S. L., "A Production of Amino Acids Under Possible Primitive Earth Conditions", *Science* 117:528-529, 1953.

Oparin, A. I., *Origin of Life,* S. Morgulis, trans., Macmillan Co., New York, 1938.

Ponnamperuma, C., and Woeller, F., "Organic Synthesis in a Simulated Jovian Atmosphere", *Icarus* 10:386-392, 1969.

Ponnamperuma, C., and Klein, H. P., "The Coming Search for Life on Mars", *Quart. Rev. Biol.* 45:235, 1970.

Rasool, S. I., and de Bergh, C., "The Runaway Greenhouse and the Accumulation of CO_2 in the Venus Atmosphere", *Nature* 226, 1970.

Ringwood, A. E., in *Advances in Earth Science,* P. M. Hurley, ed., MIT Press, Cambridge, 1964.

Rubey, W. W., "Development of the Hydrosphere and Atmosphere, With Special Reference to Probable Composition of the Early Atmosphere", in *Crust of the Earth,* A. Poldervaart, ed., Geological Society of America, New York, 1955, p. 631.

Sagan, C., Lippincott, E. R., Dayhoff, M. O., and Eck, R. V. "Organic Molecules and the Coloration of Jupiter", *Nature* 213:273-274, January 21, 1967.

Stecher, T. P., "The Reflectivity of Jupiter in the Ultraviolet", *Astrophys. J.* 142:1186-1190, 1965.
Urey, H. C., "On the Early Chemical History of the Earth and the Origin of Life", *Proc. Nat. Acad. Sci. U.S.A.* 38:351-363, 1952.
Urey, H. C., "The Atmospheres of the Planets", *Handbuch der Physik* 52:363, 1959.

BIBLIOGRAPHY

The problem of the formation of the planetary atmospheres is discussed in–
Urey, H. C., *The Planets, Their Origin and Development,* Yale University Press, New Haven, 1962.

The early history of the Earth's atmosphere and the problem of the origin of life is discussed in–
Brown, H., "Rare Gases and the Formation of the Earth's Atmosphere", in *The Atmospheres of the Earth and Planets,* G. P. Kuiper, ed., University of Chicago Press, Chicago, 1952, p. 258.
Miller, S. L., and Urey, H. C., *Science* 130:245, 1959.

Mariner 6 and 7 results on Mars are summarized in–
Leighton, R. B., "The Surface of Mars", *Scientific American,* May 1970.
JPL, *Mariner-Mars 1969: A Preliminary Report,* NASA SP-225, Washington, D.C.
Rasool, S. I., Hogan, J., Stewart, R., and Russell, L. "Temperature Distributions in the Lower Atmosphere of Mars From Mariner 6 and 7 Radio Occultation Data", *J. Atmos. Sci.* 27:841, 1970.

Venus' Mariner 5 and Venera 4 results are extensively covered and critically evaluated in–
Jastrow, R., and Rasool, S. I., ed., *The Venus Atmosphere,* Gordon and Breach, Science Publishers, Inc., New York, 1969.

Venera 5 and 6 results are given in–
Avduevsky, V. S., "A Tentative Model of the Venus Atmosphere Based on the Measurements of Veneras 5 and 6", *J. Atmos. Sci.* 27, July 1970.

The early history of the atmosphere of Mars and Venus is discussed in–
Sagan, C., "Origins of the Atmospheres of Earth and Planets", in *International Dictionary of Geophysics,* K. Runcorn, ed., Pergamon Press, New York, 1967.

The most up-to-date observations and theoretical calculations of the Jovian atmosphere are contained in–
Icarus, November 1969.
J. Atmos. Sci., September 1969.

Page intentionally left blank

CHAPTER 13
HISTORY OF THE LUNAR ORBIT*

Peter Goldreich
Department of Astronomy and Institute of Geophysics and Planetary Physics
University of California, Los Angeles

I. INTRODUCTION

In spite of its close proximity to the Earth, we probably know less about the Moon's origin than about the origin of most other solar system satellites. In large part this difficulty may be traced to the observation that the Moon no longer moves on the orbit in which it was formed. This fact has been common knowledge ever since Darwin's pioneering work in which it was shown that frictionally retarded tides continuously transfer angular momentum from the Earth's spin to the lunar orbit. The principal consequence of this transfer is to increase both the lunar semimajor axis and the length of day. From observations of the secular acceleration of the Moon's longitude, the present rate of tidal interaction may be deduced. As MacDonald (1964) has emphasized, "If the present rate of tidal interaction is appropriate for all times, the orbital elements of the Moon have undergone major changes in the last 2×10^9 years". It then becomes natural to inquire whether from the present state of the Earth-Moon system its state at past epochs may be determined. This is the question that we shall attempt to answer.

We begin by pointing out several fairly obvious difficulties. Suppose for the moment that we are able to derive a set of dynamical equations governing the evolution of the Earth-Moon system. When we integrate these equations back in time, how will we recognize the state in which the Moon was formed (or captured)? In other words, what will prevent us from overshooting this state and continuing our integration back to states through which the Earth-Moon system has never passed? In general, we shall have no means of knowing when we have reached the state in

*Research supported under NASA Nsg 216-62. A similar version of this paper appeared in *Reviews of Geophysics* 4(4):411-439, 1966.

which the Moon was formed. Only if this state were in some sense special could it be recognized when the integrations pass through it. Fortunately, the orbits in which most other major solar system satellites move are sufficiently special that we would immediately recognize if our integrations took the Moon back into such an orbit.

A second difficulty involves the derivation of the dynamical equations governing the evolution of the Earth-Moon system. Although we are now quite confident that tidal friction is responsible for the observed secular acceleration of the Moon, we are still unable to locate unambiguously the tidal energy sink. Whether tidal dissipation occurs mainly in the shallow seas or instead in the solid Earth is as yet an unresolved question. The more difficult task of determining the instantaneous tidal torque between the Earth and Moon is also largely unaccomplished. Even should both these uncertainties be resolved in the next few years (and there is some hope they might be*), we would still be ignorant of the situation as it pertained to the distant past. Thus, when we write expressions for the tidal torque, they are somewhat uncertain and may be in need of considerable revision when the torque is observationally determined. Additional dynamical uncertainty arises from our neglect of complex many-body gravitational interactions. For example, it may be that the direct gravitational interaction between the Moon and Sun produces a secular change in the Moon's semimajor axis. Existing proofs can rule this possibility out only in first- and second-order perturbation theory. Although we are ignorant of many aspects of direct gravitational interactions, this does not preclude them from being of importance. Unfortunately, the art of celestial mechanics is not yet sufficiently advanced to provide us with any answers to these difficulties.

In view of the uncertainties just described, what information can we hope to obtain by developing a set of equations governing the tidal evolution of the Earth-Moon system and then integrating them back in time? This question is best answered by a brief description of the current investigation.

Suppose that we are able to classify all satellites (other than the Moon) into two groups, those formed by accretion in orbit about a planet and those formed in orbit about the Sun and later captured by a planet. Further, let us assume that all satellites comprising the first class were formed moving in a unique type of orbit. When we integrate backward to find the past states of the Earth-Moon system, two possibilities may arise. It is possible that the past orbit of the Moon about the Earth (as predicted by our computation) may at some stage coincide with the unique state in which these other satellites are believed to have been formed. Suppose, in addition, we were able to demonstrate that this unique state would not have arisen,

*W. H. Munk, 1965, personal communication.

given a slightly different set of present parameters for the Earth-Moon system. Then we might well conclude that the Moon was also formed by accretion in this unique type of orbit. Moreover, we would have strong indirect evidence for asserting that our dynamical equations were essentially correct. On the other hand, the calculated past orbit of the Moon may never coincide with the unique states in which the close satellites are found. The situation then admits several conflicting interpretations. One possibility is that our understanding of either the tidal torque or the secular effects arising from direct gravitational interactions is incorrect and that this discrepancy is reflected by the failure of our calculations to return the Earth-Moon system to this unique state. Another interpretation, which relies on the validity of our calculations, would assert that the Moon had a mode of formation that differs from the one we have been considering. Finally, it is also possible that not only did the Moon originate in some other state from the one we are considering but also our calculations may be incorrect.

The problem of deriving dynamical equations governing the evolution of the Earth-Moon system and then integrating them back in time has been tackled by several authors beginning with Darwin (1879, 1880). More recent calculations have been performed by Gersternkorn (1955), MacDonald (1964), and Kaula (1964). MacDonald was the first to use an electronic computer , a practice continued in this investigation. Unlike Darwin, MacDonald neglected effects arising from interactions between the Earth's oblate figure and the Sun and Moon and from interactions between the Sun and Moon as well. Because these interactions play a central role in our treatment, we can expect significant differences between our results and his.

Section II describes the principal dynamical features of the Earth-Moon-Sun system in the absence of tidal forces. Equations are derived describing the precessional motions of the Moon's orbital plane and the Earth's equator plane. The solutions of these equations are described, and a scheme for numerically integrating them on a computer is devised. Five invariants of the precessional equations are derived.

Section III discusses the origin of natural satellites in light of the dynamical development given in Section II.

Section IV considers the slow changes in the conserved quantities produced by frictionally retarded tides. Both Darwin's and MacDonald's forms for the tidal torque are given.

Section V contains the results of numerical integrations of the equations derived in Sections II and IV. Comparison is made between the results obtained by using Darwin's and MacDonald's forms for the tidal torque.

Section VI attempts to view our results in light of existing theories of the origin of the Moon. Several new problems, both observational and theoretical, are suggested, which, if solved, would lead to better understanding of the dynamical evolution of the Earth-Moon system.

Because this paper follows two others which contain extensive discussion of the mechanism of tidal friction (MacDonald, 1964, and Kaula, 1964), we shall not dwell very long on this matter. It is therefore recommended that the reader who is unfamiliar with the mechanics of tidal friction first acquaint himself with the contents of the above-cited papers. The points that we emphasize will be concerned with departures or disagreements between the present treatment and the work of MacDonald and Kaula.

Finally, we must mention that our method of treating the dynamical evolution of the Earth-Moon system is rigorously applicable only for a circular lunar orbit. The error introduced by treating the present lunar eccentricity (the current value is $e = 0.0549$) as zero is undoubtedly small. The lunar eccentricity, however, is also subject to change due to tidal friction. Fortunately, indications are that the lunar eccentricity is increasing at the present time, and therefore it was smaller in the past. Calculations of the tidal rate of change of lunar orbital eccentricity show that tides raised on the Earth by the Moon tend to increase the eccentricity, whereas tides raised on the Moon by the Earth act to decrease it. The results of these calculations are uncertain because the rate of tidal dissipation in the Moon is unknown. If, however, the value of the specific dissipation function $(1/Q)$ appropriate for the Earth also applies to the Moon, the lunar orbital eccentricity would be increasing (Goldreich, 1963; MacDonald, 1964; Kaula, 1964). Thus, the approximation made in neglecting the lunar orbital eccentricity becomes even better in the past.

Our method of calculating the past state of the Earth-Moon system is based on the existence of three distinct time scales for dynamical change.

The short time scale is determined by the revolution periods of the Earth about the Sun and the Moon about the Earth, or, equivalently, by the year and current month.

The intermediate time scale is set by the precessional motions of the lunar orbit plane and the Earth's equator plane. As we shall prove in Section III, the relative precessional motion of the lunar orbit plane and the Earth's equator plane is periodic. At present, there is an 18.6-year precession of the normal to the lunar orbit plane about the normal to the ecliptic, whereas the Earth's spin axis precesses about this normal in approximately 27 000 years. The relative motion of the equatorial and lunar planes is periodic, with a period slightly in excess of the lunar precessional period.

The rate at which the frictional tides alter the state of the Earth-Moon system defines the long time scale. At present, this time scale is calibrated in billions of years.

Our plan of solution calls for successive averaging of the complete equations of motion (including tidal torques) over the short and then the intermediate time scales. These averaged equations are then integrated back a short interval on the long time scale. The equations of motion appropriate to this new state of the Earth-Moon system are then reaveraged on the short and intermediate time scales, and once again we step the averaged equations back on the tidal time scale. The first step in this procedure (i.e., averaging on the short time scale) is performed analytically, whereas the calculations on the intermediate and long time scales require the use of a large computer.

Our method of integration, by its very nature, requires that the three time scales be well separated. This requirement must be satisfied at all stages of the calculation, past as well as present. This condition is violated only when the Moon is very close to the Earth; then the precessional period becomes shorter than 1 year. Fortunately, in this configuration the solar influence on the Earth-Moon system is negligible, and since the month is still very short compared with the precessional period, our method of averaging remains satisfactory.

II. PRECESSIONAL EQUATIONS

Here we shall be concerned with motion on the short and intermediate time scales. Consideration of the tidal torques will be deferred until Section IV.

We model the Earth's principal moments of inertia by those of a rotating fluid with the current internal density distribution. The Earth's spin axis is taken to be coincident with the direction of its rotational angular momentum. Within this approximation the Earth's axis will exhibit no free nutation, this being assumed negligible. The forced nutation (at present 18 years) produced by the lunar torque will, however, be present. We partially allow for the finite strength of the Earth by using the secular Love number k_s instead of the corresponding fluid Love number k_f. The principal moments of inertia about the rotation axis and about an axis in the equatorial plane are

$$C = I[1 + (2k_s R_\oplus^5/9GI)\Omega_\oplus^2] \tag{1}$$

and

$$A = I[1 - (k_s R_\oplus^5/9GI)\Omega_\oplus^2], \tag{2}$$

where $\Omega_\oplus$ is the Earth's spin angular velocity, $R_\oplus$ is its equatorial radius, I is the moment of inertia of the equivalent sphere, and G is the gravitational constant. The value of k_s calculated from the present values of $\Omega_\oplus$ and C is (Munk and MacDonald, 1960)

$$k_s = 0.947 . \tag{3}$$

The torques that produce the precession of the equinoxes arise from the attraction of the Moon and Sun for the Earth's equatorial bulge. In a similar fashion, torques acting on the Moon arising from the attraction of the Sun and the Earth's oblate figure cause a precession of the lunar orbit plane.

The disturbing potential felt by the Moon due to the Earth's figure may be written as

$$R_1 = \frac{\mu}{r}\left[\frac{A_2}{r^2} P_2(\sin\theta) + \frac{A_3}{r^3} P_3(\sin\theta) + \cdots\right] \tag{4}$$

(Kozai, 1959) for an axially symmetric Earth. In this expression, $\mu = G(M + m)$, where M is the mass of the Earth and m is the mass of the Moon. The Earth-Moon distance is given by r; the P_n (sin θ) are Legendre polynomials of order n, with θ denoting latitude. The coefficients A_n measure the Earth's deviation from sphericity. In our model, $A_3 = 0$ because of reflection symmetry about the equator plane, and the next nonvanishing coefficient is A_4. We shall retain only the A_2 term. The truncated version of R_1 now becomes

$$R_1 = -\frac{\mu}{r}\left[+\frac{2}{3}\frac{JR_\oplus^2}{r^2} P_2(\sin\theta)\right] , \tag{5}$$

where we have replaced A_2 by the more usual expression $-2/3\,JR_\oplus{}^2$. In terms of the principal moments of inertia C and A,

$$J = \frac{3}{2}(C - A)/MR_\oplus^2 . \tag{6}$$

The disturbing potential (felt by the Moon) due to the Sun takes the form (Brouwer and Clemence, 1961),

$$R_2 = \frac{\mu}{r}\left(\frac{M_\odot}{M + m}\right)\left[\frac{r^3}{r_\odot^3} P_2(\cos S) + \left(\frac{M - m}{M + m}\right)\frac{r^4}{r_\odot^4} P_3(\cos S) + \cdots\right] . \tag{7}$$

Here, $r_\odot$, is the Earth-Sun separation, S is the angle between the Earth-Moon and Earth-Sun center lines, and $M_\odot$ is the mass of the Sun. As was done with R_1, we shall retain only the P_2 term in R_2.

At this stage we shall neglect the orbital eccentricity of both the Moon about the Earth and the Earth about the Sun. Thus r and $r_\odot$ are replaced by the corresponding semimajor axes a and $a_\odot$. Before we can average R_1 and R_2 on the short time scale and thus obtain their secular parts, we must first express the angles θ and S in terms of the Keplerian elements of the lunar and solar orbits. We shall describe the lunar orbit by the previously defined semimajor axis a, the inclination to the equator plane ϵ, and the angle in the orbit measured from the ascending node on the equator plane Φ. In addition, we shall make use of two auxiliary angles: the inclination of the orbit to the ecliptic I, and the angular position in the orbit measured from the ascending node on the ecliptic Φ'. For the solar orbit the elements other than $a_\odot$ are the obliquity of the Earth's equator to the ecliptic γ and the position in orbit as measured from the Moon's ascending node on the ecliptic u.

It is now a simple exercise in spherical trigonometry to verify that

$$\sin\theta = \sin\epsilon \sin\Phi \tag{8}$$

and

$$\cos S = \cos^2\frac{I}{2}\cos(\Phi' - u) + \sin^2\frac{I}{2}\cos(\Phi' - u)\,. \tag{9}$$

Then, R_1 may be written as

$$R_1 = \frac{2}{3}\frac{\mu}{a^3}JR_\oplus^2\left(\frac{1}{2} - \frac{3}{4}\sin^2\epsilon + \frac{3}{4}\sin^2\epsilon\cos 2\Phi\right). \tag{10}$$

If one averages R_1 over one orbit period of the Moon, the cos 2Φ term vanishes, leaving the secular part of R_1:

$$\bar{R}_1 = \frac{2}{3}\frac{\mu}{a^3}JR_\oplus^2\left(\frac{1}{2} - \frac{3}{4}\sin^2\epsilon\right). \tag{11}$$

In a similar fashion, by squaring cos S as given by Equation 9 and substituting the resulting expression into Equation 7, we find, setting $\beta = \sin I$,

$$R_2 = \mu\left(\frac{M_\odot}{M+m}\right)\frac{a^2}{a_\odot^3}\left[\frac{1}{4} - \frac{3}{8}\beta^2 + \left(\frac{3}{4} - \frac{3}{8}\beta^2\right)\cos 2(\Phi' - u) + 3\frac{\beta^2}{8}\cos 2\Phi' + 3\frac{\beta^2}{8}\cos 2u\right]. \tag{12}$$

The terms in Equation (12) containing a cosine vanish upon averaging on the short time scale. The secular part of R_2 which then remains is given by

$$\bar{R}_2 = \mu\left(\frac{M_\odot}{M+m}\right)\frac{a^2}{a_\odot^3}\left(\frac{1}{4} - \frac{3}{8}\sin^2 I\right). \tag{13}$$

The secular part of the disturbing potential felt by the Sun due to the Earth's figure is obtained by replacing ϵ by γ, a by $a_\odot$, and μ by GM in Equation 11. This potential, which we denote by R_3, is

$$\bar{R}_3 = \frac{2GMJR_\oplus^2}{3a_\odot^3}\left(\frac{1}{2} - \frac{3}{4}\sin^2\gamma\right), \tag{14}$$

where we have neglected $M/M_\odot$ compared with unity in replacing μ by GM.

Our next step will be the derivation of the mutual torques between the Earth, Sun, and Moon from the secular disturbing functions $\bar{R}_1$, $\bar{R}_2$, and $\bar{R}_3$. The geometry involved is illustrated in Figures 1 and 2. The unit vectors a, b, and c are chosen normal to the equator, lunar orbit, and ecliptic planes, respectively. Slow variations in the orientation of the ecliptic plane relative to inertial space are neglected throughout our investigation, so that c is a fixed vector in inertial space. We denote the mean motions of the Moon and Sun about the Earth by n and $n_\odot$. In Figure 1, the horizontal plane x_1x_2 coincides with the lunar orbit, whereas the ecliptic lies in the horizontal plane y_1y_2 of Figure 2.

The torques between the Earth, Moon, and Sun are derived by differentiating the expressions for $\bar{R}_1$, $\bar{R}_2$, and $\bar{R}_3$ with respect to ϵ, I, and γ and multiplying by the appropriate mass (i.e., either m or $M_\odot$). If we denote by $\mathbf{L}_{\odot\oplus}$ the secular torque produced by the Earth's equatorial bulge on the lunar orbit, we have

$$\mathbf{L}_{\odot\oplus} = -\frac{\mu m}{a^3}JR_\oplus^2 \sin\epsilon\cos\epsilon\,\frac{\mathbf{a}\times\mathbf{b}}{|\mathbf{a}\times\mathbf{b}|}, \tag{15}$$

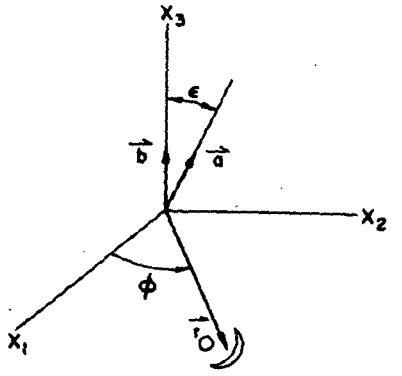


Figure 1.–Coordinate system used to describe the Moon's motion.

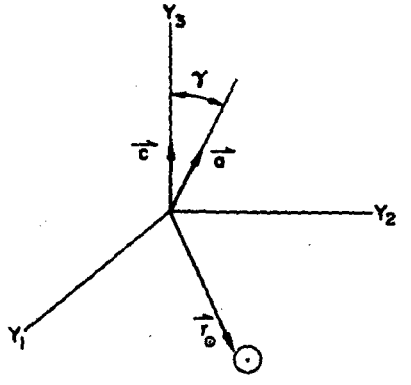


Figure 2.–Coordinate system used to describe the Sun's motion.

where $\mathbf{L}_{\odot\oplus}$ is seen to lie along the Moon's ascending node on the equator plane. Writing $\sin\epsilon = |\mathbf{a} \times \mathbf{b}|$, we obtain

$$\mathbf{L}_{\odot\oplus} = -\frac{\mu m}{a^3} J R_\oplus^2 (\mathbf{a} \cdot \mathbf{b})(\mathbf{a} \times \mathbf{b}) . \tag{16}$$

In an analogous fashion, the secular torque exerted by the equatorial bulge on the Sun may be written as

$$\mathbf{L}_{\odot\oplus} = -\frac{GMM_\odot}{a_\odot^3} J R_\oplus^2 (\mathbf{a} \cdot \mathbf{c})(\mathbf{a} \times \mathbf{c}) . \tag{17}$$

Lastly, using $\bar{R}_2$ we find for the secular torque exerted by the Sun on the Moon

$$\mathbf{L}_{\odot\odot} = +\frac{3}{4}\mu\left(\frac{M_\odot m}{M+m}\right)\frac{a^2}{a_\odot^3}(\mathbf{b}\cdot\mathbf{c})(\mathbf{b}\times\mathbf{c})\,. \tag{18}$$

To complete the set of mutual torques we have the Sun's torque on the Earth $\mathbf{L}_{\oplus\odot} = -\mathbf{L}_{\odot\oplus}$, the Moon's torque on the Earth $\mathbf{L}_{\oplus\odot} = -\mathbf{L}_{\odot\oplus}$, and the Moon's torque on the Sun $\mathbf{L}_{\odot\odot} = -\mathbf{L}_{\odot\odot}$.

Let us denote by H and h the scalar angular momenta in the Earth's spin and in the lunar orbital motion, respectively. The small intrinsic spin angular momentum of the Moon is neglected throughout. Then, the precessional equations of motion may be written as

$$\frac{dH\mathbf{a}}{dt} = \mathbf{L}_{\oplus\odot} + \mathbf{L}_{\oplus\odot}$$

and (19)

$$\frac{dh\mathbf{b}}{dt} = \mathbf{L}_{\odot\oplus} + \mathbf{L}_{\odot\odot}$$

Setting

$$\mathbf{L}_{\oplus\odot} = -\mathbf{L}_{\odot\oplus} = L(\mathbf{a}\times\mathbf{b})(\mathbf{a}\cdot\mathbf{b})\,,$$

$$\mathbf{L}_{\oplus\odot} = -\mathbf{L}_{\odot\oplus} = K_1(\mathbf{a}\times\mathbf{c})(\mathbf{a}\cdot\mathbf{c})\,, \tag{20}$$

and

$$\mathbf{L}_{\odot\odot} = -\mathbf{L}_{\odot\odot} = K_2(\mathbf{b}\times\mathbf{c})(\mathbf{b}\cdot\mathbf{c})\,,$$

where

$$L = \frac{\mu m}{a^3}JR_\oplus^2\,,$$

$$K_1 = \frac{\mu M_\odot}{a_\odot^3}JR_\oplus^2\,, \tag{21}$$

and

$$K_2 = \frac{3}{4}\mu\left(\frac{M_\odot m}{M+m}\right)\frac{a^2}{a_\odot^3}$$

we obtain

$$\frac{dH\mathbf{a}}{dt} = L(\mathbf{a}\times\mathbf{b})(\mathbf{a}\cdot\mathbf{b}) - K_1(\mathbf{c}\times\mathbf{a})(\mathbf{a}\cdot\mathbf{c})$$

and (22)

$$\frac{dh\mathbf{b}}{dt} = -L(\mathbf{a}\times\mathbf{b})(\mathbf{a}\cdot\mathbf{b}) + K_2(\mathbf{b}\times\mathbf{c})(\mathbf{b}\cdot\mathbf{c})\,.$$

The approximations made in deriving Equations 22 are briefly recalled. We have ignored the eccentricity of both the Moon and the Earth orbits. The Earth has been modeled by a fluid rotating such that its angular velocity $\Omega_\oplus$ is always parallel to its angular momentum $H\mathbf{a}$. Finally, we have averaged the equations of motion on the short time scale. This last step is equivalent to replacing the Sun and Moon by rotating tori in their respective orbit planes. In fact, it is a simple matter to derive Equations 22 once this relation is realized. The reasoning behind our more formal (and lengthy) derivation of these equations is that the apparatus we have developed in this section will be used later in less obvious circumstances, and it is well to get accustomed to it now.

Let us derive a set of scalar equations from the vector relations given by Equations 22. We begin by proving that H and h are conserved quantities. Taking the dot product of the first equation with $\mathbf{a}$ yields

$$\mathbf{a} \cdot \frac{dH\mathbf{a}}{dt} = 0 = \mathbf{a} \cdot \mathbf{a}\frac{dH}{dt} + H\mathbf{a} \cdot \frac{d\mathbf{a}}{dt} = \frac{dH}{dt}; \tag{23}$$

thus,

$$\frac{dH}{dt} = 0 .$$

Similarly, dotting $\mathbf{b}$ into the second equation, we find

$$dh/dt = 0 . \tag{24}$$

Dotting $\mathbf{c}$ into Equations 22 and using Equations 23 and 24, we arrive at

$$H\frac{d(\mathbf{a} \cdot \mathbf{c})}{dt} = L(\mathbf{a} \cdot \mathbf{b})(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} \tag{25}$$

and

$$h\frac{d(\mathbf{b} \cdot \mathbf{c})}{dt} = -L(\mathbf{a} \cdot \mathbf{b})(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} . \tag{26}$$

Forming the combination $hH\, d(\mathbf{a} \cdot \mathbf{b})/dt$, we find

$$\frac{d(\mathbf{a} \cdot \mathbf{b})}{dt} = \left[\frac{K_2}{h}(\mathbf{b} \cdot \mathbf{c}) - \frac{K_1}{H}(\mathbf{a} \cdot \mathbf{c})\right](\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} . \tag{27}$$

Another constant of the motion may be derived by adding Equations 25 and 26. The new constant Λ is the component of total angular momentum in the Earth-Moon system that is normal to the ecliptic:

$$\frac{d\Lambda}{dt} = \frac{d}{dt}[H(\mathbf{a}\cdot\mathbf{c}) + h(\mathbf{b}\cdot\mathbf{c})] = 0 . \tag{28}$$

The conservation of Λ arises because the external (solar) torques on the Earth-Moon system lie in the ecliptic. The fourth and final constant of the motion can be derived by multiplying Equation 25 by $2(\mathbf{a}\cdot\mathbf{c})/H$, Equation 26 by $2(\mathbf{b}\cdot\mathbf{c})/h$, and Equation 27 by $2(\mathbf{a}\cdot\mathbf{b})$, and adding the resulting equations. This yields

$$\frac{d\chi}{dt} = \frac{d}{dt}[K_1(\mathbf{a}\cdot\mathbf{c})^2 + K_2(\mathbf{b}\cdot\mathbf{c})^2 + L(\mathbf{a}\cdot\mathbf{b})^2] = 0 . \tag{29}$$

From its form, χ is seen to be a sum of potential energies. This is not surprising, since an energy integral must exist, and the kinetic energies are separately conserved as a consequence of the constancy of H and h.

To summarize, in the absence of tidal friction we have obtained four conserved quantities, H, h, Λ, and χ. These constants will all be shown to vary on the long time scale when tidal torques are included.

The triple scalar product $(\mathbf{a}\times\mathbf{b})\cdot\mathbf{c}$ appears on the right-hand side of Equations 25 through 27. It may be expressed in terms of $(\mathbf{a}\cdot\mathbf{c})$, $(\mathbf{b}\cdot\mathbf{c})$, and $(\mathbf{a}\cdot\mathbf{b})$ as

$$|(\mathbf{a}\times\mathbf{b})\cdot\mathbf{c}|^2 = 1 - (\mathbf{a}\cdot\mathbf{c})^2 - (\mathbf{b}\cdot\mathbf{c})^2 - (\mathbf{a}\cdot\mathbf{b})^2 + 2(\mathbf{a}\cdot\mathbf{c})(\mathbf{b}\cdot\mathbf{c})(\mathbf{a}\cdot\mathbf{b}) . \tag{30}$$

Before discussing the solutions of the precessional equations, we shall first define a more compact notation. Setting

$$x = (\mathbf{a}\cdot\mathbf{c}) , \quad y = (\mathbf{b}\cdot\mathbf{c}) , \quad z = (\mathbf{a}\cdot\mathbf{b}) , \quad w = (\mathbf{a}\times\mathbf{b})\cdot\mathbf{c} , \tag{31}$$

we can write the relevant scalar equations as

$$\Lambda = Hx + hy , \tag{32}$$

$$\chi = K_1x^2 + K_2y^2 + Lz^2 , \tag{33}$$

$$\frac{dx}{dt} = \frac{L}{H}zw , \tag{34}$$

$$\frac{dy}{dt} = -\frac{L}{h} zw \,, \tag{35}$$

$$\frac{dz}{dt} = \left(\frac{K_2}{h} y - \frac{K_1}{H} x\right) w \,, \tag{36}$$

and

$$w^2 = 1 - x^2 - y^2 - z^2 + 2xyz \,. \tag{37}$$

The three first-order equations (Equations 34 through 36) together with Equation 37 may be solved simultaneously for x, y, and z as functions of time. Their solution is made possible by the constancy of H and h. If we preferred, we could use the conservation laws for χ and Λ to eliminate two of the three dependent variables (x, y, z) and obtain a single first-order differential equation. This procedure turns out to be less favorable than it might at first appear, mainly because of the difficulty in choosing the correct branches of the square roots and because of the relatively long machine time required to compute them. In fact, as we see from Equation 37, the values of x, y, and z specify only the magnitude but not the sign of w, whose evaluation also involves computation of a square root. Once again, it is advantageous to avoid taking the square root in our computations. We bypass this difficulty by evaluating w not from Equation 37, but instead from the equation obtained by taking its time derivative. Although this procedure increases from three to four the number of simultaneous differential equations to be solved, it still results in a reduction in computing time. Thus, we shall always compute w from the equation

$$\frac{dw}{dt} = \frac{L}{H} z(yz - x) - \frac{L}{h} z(xz - y) + \left(\frac{K_2}{h} y - \frac{K_1}{H} x\right)(xy - z) \,. \tag{38}$$

In the numerical calculations which we shall describe in Section V, we numerically integrate the simultaneous Equations 34, 36, and 38. Equation 32 is used to eliminate y in favor of x. As mentioned previously, we chose not to use Equations 33 and 37 to eliminate w and x.

In all cases of interest to us, the solutions of these equations are periodic. The nature of the solution is best understood in terms of the triple scalar product w. From the definition of w we know that it vanishes if and only if a, b, and c are

coplanar. Furthermore, from Equations 34 to 38, we see that changing the sign of w while keeping x, y, and z fixed is tantamount to a change in sign of t, or, in other words, to time reversal. Thus, the time interval between consecutive zeros of w is just the half-period. In the cases of interest to us, it is a simple matter to verify that w does actually attain both positive and negative values. The crux of the proof in this case rests on the fact that z is always positive in our calculations. Then, from Equations 34 and 35, we note that x and y will vary monotonically unless w changes sign. Since $|x|$ and $|y|$ must always remain less than unity, we see that w must change sign.

We delay further consideration of the details of integrating these equations until after the development of the equations governing tidal friction because the initialization of the precessional equations and the solution of the tidal equations are intimately related. In the next section we shall devote some time to discussing several general features of the solutions of the precessional equations. The information we shall require may be obtained entirely from the integrals of the motion H, h, χ, and Λ and bears on the general subject of the origin of natural satellites.

III. ORIGIN OF NATURAL SATELLITES

A glance at the orbital elements of the natural satellites reveals that they may be classified into two groups. The first group is composed of satellites moving on equatorial or near-equatorial orbits. With the exception of the Moon and Triton (Neptune 1), it includes all lunar-sized satellites. Relevant orbital parameters for the two groups are listed in Table 1. All data have been taken from Allen (1963). It is to be noted that the satellites of the first group generally occupy smaller orbits than those of the second. This trend is especially striking when one considers only those satellites orbiting a particular major planet.

We shall now show that for each planet there is a critical distance such that a satellite orbit lying well within this distance will maintain a nearly constant inclination to the planet's equator plane. This nearly constant inclination is maintained in spite of the precessional motions of both the orbit and the equator planes. For orbits much larger than the critical one, the satellite orbit plane no longer maintains constant inclination to the planet's equator, but rather it holds a nearly constant inclination to the plane of the planet's orbit around the Sun. To prove these statements and derive an expression for the critical radius, we need use

Table 1.—Orbital parameters for satellites.

Equatorial or Near-Equatorial		Nonequatorial	
Satellites	a/R	Satellites	a/R
Earth		Earth	
		Moon	60.27
Mars		Mars	
1 Phobos	2.76		
2 Deimos	6.92		
Jupiter		Jupiter	
1 Io	5.90	6	160.7
2 Europa	9.40	7	164.4
3 Ganymede	14.99	8	326
4 Callisto	26.36	9	332
5	2.54	10	164
		11	313
		12	290
Saturn		Saturn	
1 Mimas	3.11	Iapetus	59.67
2 Enceladus	3.99	Phoebe	216.8
3 Tethys	4.94		
4 Dione	6.33		
5 Rhea	8.84		
6 Titan	20.48		
7 Hyperion	24.83		
Uranus		Uranus	
1 Ariel	8.08		
2 Umbriel	11.25		
3 Titania	18.46		
4 Oberon	24.69		
5 Miranda	5.49		
Neptune		Neptune	
		Triton	15.85
		Neried	249.5

only χ and Λ, the energy and angular momentum integrals of the precessional motion (Equations 32 and 33). In this section, notation previously reserved for the Earth and Moon will be used to denote planets and satellites generally. Eliminating x in favor of y in the equation for χ, we obtain

$$\chi - K_1\left(\frac{\Lambda}{H}\right)^2 = Lz^2 - \frac{2K_1\Lambda hy}{H^2} + K_1\left(\frac{h}{H}\right)^2 y^2 + K_2 y^2 . \tag{39}$$

The left-hand side of this equation is a constant. Hence, the condition that z (and thus the satellite's inclination to the planet's equator) remain nearly fixed during the precessional motion is that L be much larger than the remaining terms on the right-hand side. Using Equation 21, we find

$$\frac{K_1 h}{K_2 H} = \frac{4J}{3}\frac{M}{m}\left(\frac{R_\oplus}{a}\right)^2 \frac{h}{H} = \frac{4J}{3\alpha}\frac{D}{T_s}, \tag{40}$$

where D and T_s are the planet's spin period and the satellite's orbit period, respectively, and α is defined so that the planet's rotational moment of inertia $C = \alpha M R_\oplus{}^2$. For all satellite orbits, $K_1 h/K_2 H \ll 1$. Furthermore, for all satellites except the Moon, $h < H$ and $\Lambda < H$, and thus the terms on the right-hand side of Equation 40 having K_1 as coefficient may be neglected with respect to the K_2 term. Even for the Moon, where $h/H \approx 5$, it is easily verified that the K_1 terms may be dropped.

Having neglected the K_1 terms, we see that the critical distance from the planet is determined by the value of a at which $K_2 = L$. Physically, this condition is equivalent to determining the value of a at which the torques on the satellite orbit due to the planet and the Sun are equal. Again using Equation 21, we obtain

$$\frac{K_2}{L} \approx \frac{3}{4J}\frac{M_\odot}{M}\frac{a^5}{R_\oplus^2 a_\odot^3} . \tag{41}$$

The critical value of a as determined from the equality of K_2 and L is given by

$$\left(\frac{a}{R_\oplus}\right)_c \approx \left(\frac{4J}{3}\frac{M}{M_\odot}\right)^{1/5}\left(\frac{a_\odot}{R_\oplus}\right)^{3/5} . \tag{42}$$

Values of $(a/R_\oplus)_c$ are listed in Table 2 for planets having known satellites. It should be mentioned that for Jupiter and Saturn, the large satellites act like the equatorial bulge in forcing other satellites to maintain constant inclinations to the equator

plane. Thus, $a/R_\oplus$ for these planets is really somewhat larger than the value listed in Table 2.

Because the ratio K_2/L varies as the fifth power of $a/R_\oplus$, we see that satellites that are closer or farther than the critical radius by factors as small as 2 are already almost completely under the domination of either the planetary or the solar torque.

A comparison of the critical values of $a/R_\oplus$ with the 31 satellites listed in Table 1 shows that with the single exception of Triton, the equatorial satellites are those lying within $(a/R_\oplus)_c$ and the nonequatorial satellites are those lying outside $(a/R_\oplus)_c$. Of course, it is not surprising that satellites lying outside $(a/R_\oplus)_c$ are not found in equatorial orbits. Indeed, even if one of these satellites were placed in an equatorial orbit, it would soon precess off it. The surprising result is that, with the single exception of Triton, the 20 satellites found closer than the critical distance from their planets all move in equatorial orbits. This fact certainly calls for an explanation.

Although it is known that tidal torques will move nonequatorial satellites [within $(a/R_\oplus)_c$] toward equatorial orbits, there are several cases of these orbits for which tidal changes of inclination have definitely been negligible. For example, the satellites Titan, Hyperion, and Deimos fall into this category. We are then led to believe that equatorial satellites reflect a condition of origin, not evolution. This conclusion further suggests that these satellites were formed by accretion from a

Table 2.–Initial values of $(a/R)_c$ for various planets.

Planet	a/R Critical
Earth*	10
Mars	13
Jupiter**	32
Saturn**	43
Uranus†	84
Neptune	70

*For the Earth, $(a/R_\oplus)_c$ has been computed using the current figure. When the Moon was at $10R_\oplus$, $(a/R_\oplus)_c \approx 17$, owing to the increased oblateness of the Earth.

**The critical values of a/R for Jupiter and Saturn become somewhat larger when account is taken of the effects of their satellites. Thus, Ganymede increases $(a/R)_c$ for Jupiter to 38, and Titan increases $(a/R)_c$ for Saturn to 57.

†The value of J for Uranus is unknown; we have used $J = 0.017$ as an estimate.

thin equatorial disk that formed about their planets. The formation of a disk of this type would be the natural result of inelastic collisons between particles orbiting the planet. It is to be emphasized that the equatorial plane is unique in its ability to maintain such a disk. A disk of particles, if placed in any other plane, would rapidly disperse into a band as the individual particle orbits precessed at different rates (Goldreich, 1965). Of course, beyond $(a/R_\oplus)_c$, an equatorial disk could not form. We might expect that here a disk might form in the planet's orbit plane and that as a consequence some satellites would have been formed there. The absence of satellites in a planet's orbit plane beyond $(a/R_\oplus)_c$ implies that the material from which satellites formed did not extend out to these distances.

Triton, the one exception to our rule, moves on a near-circular retrograde orbit inclined by 20 degrees to Neptune's equator. It has been suggested by Lyttleton (1936) that Triton's orbit is the result of a near encounter which ejected Pluto. A more recent evaluation of the evidence for this hypothesis has been given in Goldreich and Soter (1966).

Let us return now to a discussion of the origin of the Moon. As we can see, the present distance to the Moon places it far outside the critical distance from the Earth. We know, however, that the Moon was closer to the Earth in the past, and possibly at one time it was even inside the critical radius. On the basis of our examination of other satellite orbits, we strongly suspect that if the Moon had formed within the Earth's critical radius, it would have formed on an equatorial orbit. Of course, the critical radius was larger when the Moon was closer to the Earth because the Earth was spinning faster then. We shall find that the Moon's orbit would have maintained a constant (to within a degree) inclination to the Earth's equator if the Moon were ever closer than about $10R_\oplus$. By means of numerical calculations, we attempt to determine whether this inclination was ever exactly zero.

IV. EQUATIONS OF TIDAL FRICTION

We shall derive the equations of motion that are valid on the long time scale. The principal tidal changes are brought about by the frictionally retarded lunar tide raised on the Earth. Additional smaller changes are produced by tides raised on the Earth by the Sun. The tide raised by the Earth on the synchronously rotating Moon produces only radial forces and hence does not lead to any secular changes (within our approximation of a circular lunar orbit). Accordingly, this tide will be neglected. The effects due to the tides raised by the Sun on the Moon and by the Earth and Moon on the Sun are easily seen to be negligible. Thus, our calculations will only include effects arising from the solar- and lunar-induced Earth tides.

Denoting the total tidal torques acting on the Earth and Moon by $\mathbf{T}_\oplus$ and $\mathbf{T}_\bigcirc$, we may augment Equations 19 to obtain equations valid on the long time scale:

$$\frac{dH\mathbf{a}}{dt} = \mathbf{L}_{\oplus\bigcirc} + \mathbf{L}_{\oplus\odot} + \mathbf{T}_\oplus \tag{43a}$$

and

$$\frac{dh\mathbf{b}}{dt} = \mathbf{L}_{\bigcirc\oplus} + \mathbf{L}_{\bigcirc\odot} + \mathbf{T}_\bigcirc . \tag{43b}$$

We are interested here in deriving equations governing the slow tidally induced changes in quantities that were constant on the intermediate time scale. We may simplify these derivations by neglecting the torques $\mathbf{L}_{\oplus\bigcirc}$, $\mathbf{L}_{\oplus\odot}$, $\mathbf{L}_{\bigcirc\oplus}$, and $\mathbf{L}_{\bigcirc\odot}$ from the outset. We begin by dotting the first of the resulting abridged equations by **a** and the second by **b** to obtain

$$dH/dt = \mathbf{T}_\oplus \cdot \mathbf{a} \tag{44}$$

and

$$dh/dt = \mathbf{T}_\bigcirc \cdot \mathbf{b} . \tag{45}$$

Next, dotting both equations by **c** yields

$$H\,dx/dt = \mathbf{T}_\oplus \cdot \mathbf{c} - x\mathbf{T}_\oplus \cdot \mathbf{a} \tag{46}$$

and

$$h\,dy/dt = \mathbf{T}_\bigcirc \cdot \mathbf{c} - y\mathbf{T}_\bigcirc \cdot \mathbf{b} . \tag{47}$$

Of course, in deriving Equations 46 and 47, we have neglected terms arising from the precessional torques. Next, dotting **b** into Equation 43a and **a** into Equation 43b and then adding, we find

$$\frac{dz}{dt} = \frac{\mathbf{T}_\oplus \cdot \mathbf{b}}{H} + \frac{\mathbf{T}_\bigcirc \cdot \mathbf{a}}{h} - z\left(\frac{\mathbf{T}_\oplus \cdot \mathbf{a}}{H} + \frac{\mathbf{T}_\bigcirc \cdot \mathbf{b}}{h}\right) . \tag{48}$$

From Equations 44 to 47 it now follows that

$$d\Lambda/dt = (\mathbf{T}_\oplus + \mathbf{T}_\bigcirc) \cdot \mathbf{c} . \tag{49}$$

An equation for da/dt may be derived from Equation 45 for dh/dt. Since $h = m(\mu a)^{1/2}$, it follows that

$$da/dt = 2a\mathbf{T}_\bigcirc \cdot \mathbf{b}/h . \tag{50}$$

Using Equations 1, 5, and 22, we observe that

$$\frac{dK_1}{dt} = \frac{2K_1}{H}\frac{dH}{dt} = \frac{2K_1}{H}\mathbf{T}_\oplus \cdot \mathbf{a}\,,$$

$$\frac{dK_2}{dt} = \frac{2K_2}{a}\frac{da}{dt} = \frac{4K_2\mathbf{T}_\odot \cdot \mathbf{b}}{h}\,,$$

and

$$\frac{dL}{dt} = \frac{2L}{H}\frac{dH}{dt} - \frac{3L}{a}\frac{da}{dt} = \frac{2L}{H}\mathbf{T}_\oplus \cdot \mathbf{a} - \frac{6L\mathbf{T}_\odot \cdot \mathbf{b}}{h}\,. \tag{51}$$

Hence, $d\chi/dt$ is given by

$$\frac{d\chi}{dt} = \frac{2K_1}{H}x\mathbf{T}_\oplus \cdot \mathbf{c} + \frac{2K_2}{h}y(\mathbf{T}_\odot \cdot \mathbf{c} + y\mathbf{T}_\odot \cdot \mathbf{b}) + 2Lz\left(\frac{\mathbf{T}_\oplus \cdot \mathbf{b}}{H} + \frac{\mathbf{T}_\odot \cdot \mathbf{a} - 4z\mathbf{T}_\odot \cdot \mathbf{b}}{h}\right). \tag{52}$$

The components of $\mathbf{T}_\oplus$ and $\mathbf{T}_\odot$ are most conveniently resolved along the directions of the $x_1x_2x_3$ and $y_1y_2y_3$ axes, which were described in Figures 1 and 2. We shall denote the unit vectors along these axes by $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$ and $\mathbf{f}_1$, $\mathbf{f}_2$, and $\mathbf{f}_3$, respectively. Expressing these unit vectors in terms of **a**, **b**, and **c**, we have

$$\mathbf{e}_1 = \frac{\mathbf{a} \times \mathbf{b}}{\sin\epsilon}\,,$$

$$\mathbf{e}_2 = \frac{\mathbf{a} - \cos\epsilon\mathbf{b}}{\sin\epsilon}\,, \tag{53}$$

and

$$\mathbf{e}_3 = \mathbf{b}\,;$$

and

$$\mathbf{f}_1 = \frac{\mathbf{a} \times \mathbf{c}}{\sin\gamma}\,,$$

$$\mathbf{f}_2 = \frac{\mathbf{a} - \cos\gamma\mathbf{c}}{\sin\gamma}\,, \tag{54}$$

and

$$\mathbf{f}_3 = \mathbf{c}\,.$$

Writing either torque $\mathbf{T}_\oplus$ or torque $\mathbf{T}_\odot$ as

$$\mathbf{T} = T_1\mathbf{e}_1 + T_2\mathbf{e}_2 + T_3\mathbf{e}_3 + T_1'\mathbf{f}_1 + T_2'\mathbf{f}_2 + T_3'\mathbf{f}_3 ,$$

we find

$$\begin{aligned} \mathbf{T}\cdot\mathbf{a} &= T_2(1-z^2)^{1/2} + T_3 z + T_2'(1-x^2)^{1/2} + T_3' x , \\ \mathbf{T}\cdot\mathbf{b} &= T_3 - T_1'\frac{w}{(1-x^2)^{1/2}} + \frac{T_2'(z-xy)}{(1-x^2)^{1/2}} + T_3' y , \\ \mathbf{T}\cdot\mathbf{c} &= T_1\frac{w}{(1-z^2)^{1/2}} + T_2\frac{(x-yz)}{(1-z^2)^{1/2}} + T_3 y + T_3' . \end{aligned} \tag{55}$$

and

We may save ourselves some extra labor by noticing that the T_1 and T_1' terms in Equations 55 all contain w as a factor. When we average these dot products (multiplied by functions of x, y, and z) over a precessional period, the T_1 and T_1' terms will all vanish, since w has odd parity relative to x, y, and z. Thus, we shall drop these terms from here on.

We shall employ expressions for the tidal torques that were developed by Kaula (1964) and MacDonald (1964). However, we shall not repeat the derivations they gave; the reader who is unfamiliar with these two papers may prefer to skim through them before reading the rest of this section. In Section V, we shall compare and contrast the results obtained from two different forms of the tidal torque. Since MacDonald's expressions for the tidal torques are simpler, we begin with them. Because it proves to be very difficult, if not impossible, to allow correctly (within MacDonald's formulation) for the interaction of the Sun with the lunar tide and the Moon with the solar tide, we shall restrict consideration to the lunar tide alone. The solar tide will be treated in the development that follows Kaula's formulation.

MacDonald assumes that the lunar tidal potential produces a second harmonic distortion of the Earth. The effects of friction are modeled by delaying the time of high tide and having the Earth's rotation carry the tidal bulge forward. High tide occurs not directly under the Moon but at an angle δ ahead of the Moon in the direction of the relative angular velocity of the Earth and Moon. Several points should be mentioned in connection with this model of frictionally retarded tides.

First, although the relative angular velocity between the Earth and Moon varies with the Moon's position in orbit (even for a circular orbit if it is inclined to the equator), MacDonald keeps the geometric lag angle δ constant. Thus, the time lag of the tide varies with the Moon's position in its orbit. Second, MacDonald calculates

the perturbing tidal forces by differentiating the instantaneous tidal potential (due to the frictionally retarded tide) along the direction of the instantaneous relative motion of the Moon and the Earth's surface just below it. This gives the correct direction for the tidal bulge only in the limit of small δ. The restriction to small δ is not critical at the present time, however, since from the known rate of tidal energy dissipation we know that δ would be about 7 degrees if the tidal torque acted between the Moon and the oceans, and only about 2.16 degrees if the torque acted between the solid Earth and the Moon.

Following MacDonald, we define several new quantities:

$$\alpha = n/\Omega \; ; \tag{56}$$

$$q^2 = \frac{1 - z^2}{(1 + \alpha^2 - 2\alpha z)} \; ; \tag{57}$$

$$q'^2 = 1 - q^2 \; , \tag{58}$$

where sign q' = sign $(z - \alpha)$;

$$F(q) = \int_0^{\pi/2} \frac{d\varphi}{(1 - q^2 \sin^2 \varphi)^{1/2}} \; ; \tag{59}$$

$$E(q) = \int_0^{\pi/2} (1 - q^2 \sin^2 \varphi)^{1/2} \, d\varphi \; ; \tag{60}$$

$$B(q) = \frac{1}{q^2} [E(q) - q'^2 F(q)] \; ; \tag{61}$$

and

$$A = \frac{3}{2} G m R_\oplus^5 k_2 \; , \tag{62}$$

where k_2 is the dimensionless tidal Love number (Munk and MacDonald, 1960).

In terms of this notation, MacDonald's expressions for the tidal torques, when simplified to a circular orbit, become

$$T_{1\odot} = -T_{1\oplus} = 0 \; ,$$

$$T_{2\odot} = -T_{2\oplus} = \frac{2nmA}{\pi a^6} q B(q) \sin 2\delta \; , \tag{63}$$

and

$$T_{3\odot} = -T_{3\oplus} = \frac{2nmA}{\pi a^6} q' F(q) \sin 2\delta \; .$$

In the absence of solar tides they are the only nonvanishing components of the tidal torque. The component torques listed in Equations 63 have already been averaged on the short time scale; that is, they have been averaged over a lunar orbital period.

In his development of the tidal disturbing function, Kaula follows the procedure first applied by Darwin (1879). The lunar (or solar) tidal potential felt by the Earth is expanded in a Fourier time series. The effects of tidal dissipation are modeled by assigning a phase lag to each component tide, causing it to lag behind the potential that raises it. The magnitude and frequency dependence of these phase lags should be set by observation, but present tidal observations do not permit this. Algebraically, MacDonald's procedure is much simpler, although recent workers have considerably reduced the complexity of Darwin's original development. Both this method and MacDonald's method are subject to the criticism that severe local dissipation (in either the oceans or crust) may make the approximation of a slightly distorted (by friction) static elastic tide a very poor fit to reality. The answer to this uncertainty must await further observational and theoretical investigation.

Following Kaula, we define some new symbols and notations. The asterisk (*) will denote a function of the tide-raising body. The component tides are designated by two subscripts, m and p (both of which assume the values 0, 1, and 2). We will denote the mean anomaly by the customary symbol M, whereas the phase lags of the individual tides will be denoted by ϵ_{mp}. The tidal potential per unit mass on the disturbed body is

$$u = \frac{k_2 R_\oplus^5 B_m m^*}{a^3 a^{*3}} F_{mp}(i^*) F_{mh}(i) \cos\left\{\nu^*_{mp} - \epsilon_{mp} - m\theta^* - \nu_{mh} + m\theta\right\}, \qquad (64)$$

where the summation convention applies to the indices m, p, and h, all of which run through 0, 1, and 2:

$$B_m = G\frac{(2-m)!}{(2+m)!}(2 - \delta_{0m}) \qquad (65)$$

and

$$\nu_{mp} = 2(1-p)(\omega + M) + m\Omega\,. \qquad (66)$$

The $F_{mp}(i)$ are trigonometric polynomials in the angle i and are listed in Table 3. We measure i from the Earth's equator and Ω from an inertially fixed line in the equator. Torques may be derived from the potential u by differentiation using the formulas (Kaula, 1964)

$$T_2 = \frac{1}{\sin i}\frac{\partial u}{\partial \Omega} - \cot i \frac{\partial u}{\partial \omega} \qquad (67)$$

Table 3.–Inclination polynomials $F_{mp}(i)$ of the disturbing function (Kaula, 1964).

m	p	$F_{mp}(i)$
0	0	$-3/8 \sin^2 i$
0	1	$3/4 \sin^2 i - 1/2$
0	2	$-3/8 \sin^2 i$
1	0	$3/4 \sin i (1 + \cos i)$
1	1	$-3/2 \sin i \cos i$
1	2	$3/4 \sin i (\cos i - 1)$
2	0	$3/4 (1 + \cos i)^2$
2	1	$3/2 \sin^2 i$
2	2	$3/4 (1 - \cos i)^2$

and

$$T_3 = \partial u/\partial M \tag{68}$$

These torques act on the disturbed body (Sun or Moon). The torques acting on the Earth are the negative values of these. In this manner, we obtain

$$T_{\oplus 2} = \frac{m^2 k_2 R_\oplus^5 B_m}{a^6} F_{mp}^2(\epsilon)\left[\frac{m - 2(1-p)z}{(1-z^2)^{1/2}}\right] \sin \epsilon_{\odot mp}$$

$$- \frac{M_\odot m R_\oplus^5 B_m}{a^3 a_\odot^3} F_{m1}(\epsilon) F_{m1}(\gamma)\left[\frac{m}{(1-z^2)^{1/2}}\right] \sin [m(\Omega_\odot - \Omega) - \epsilon_{\odot m1}], \tag{69}$$

$$T_{\oplus 3} = \frac{2m^2 k_2 R_\oplus^5 B_m}{a^6} F_{mp}^2(\epsilon)(1-p) \sin \epsilon_{\odot mp}, \tag{70}$$

$$T'_{\oplus 2} = \frac{M_\odot^2 k_2 R_\oplus^5 B_m}{a_\odot^6} F_{mp}^2(\gamma)\left[\frac{m - 2(1-p)x}{(1-x^2)^{1/2}}\right] \sin \epsilon_{\odot mp}$$

$$- \frac{M_\odot m R_\oplus^5 B_m}{a^3 a_\odot^3} F_{m1}(\epsilon) F_{m1}(\gamma)\left[\frac{m}{(1-x^2)^{1/2}}\right] \sin [m(\Omega - \Omega_\odot) - \epsilon_{\odot m1}], \tag{71}$$

$$T'_{\oplus 3} = \frac{2M_\odot^2 k_2 R_\oplus^5 B_m}{a_\odot^6} F_{mp}^2(\gamma)(1-p)\sin\epsilon_{\odot mp}\,, \tag{72}$$

$$T_{\odot 2} = -T_{\oplus 2}\,, \tag{73}$$

and

$$T_{\odot 3} = -T_{\oplus 3}\,. \tag{74}$$

In Equations 69 through 74, Ω and $\Omega_\odot$ are the ascending nodes of the lunar and solar orbits in the equator plane. The first subscript on the phase lags ϵ denotes the tide-raising body. The terms in $T_{\oplus 2}$ and $T_{\oplus 2}'$ containing $M_\odot m$ as a factor arise from the attractions between the Moon and the solar tide and between the Sun and the lunar tide. These terms were not considered by either MacDonald or Kaula; nevertheless, we shall discover that they are of great importance in determining the rate of change of the Earth's obliquity. To express $\sin m(\Omega_\odot - \Omega)$ and $\cos m(\Omega_\odot - \Omega)$ in terms of x, y, z, and w, we use the relations

$$\cos(\Omega - \Omega_\odot) = \frac{(\mathbf{a}\times\mathbf{b})\cdot(\mathbf{a}\times\mathbf{c})}{|\mathbf{a}\times\mathbf{b}||\mathbf{a}\times\mathbf{c}|}\,, \tag{75a}$$

or, equivalently,

$$\cos(\Omega - \Omega_\odot) = (y - zx)/(1-z^2)^{1/2}(1-x^2)^{1/2}\,, \tag{75b}$$

and

$$\sin(\Omega - \Omega_\odot) = \frac{\mathbf{a}\cdot[(\mathbf{a}\times\mathbf{b})\times(\mathbf{a}\times\mathbf{c})]}{|\mathbf{a}\times\mathbf{b}||\mathbf{a}\times\mathbf{c}|}\,, \tag{76a}$$

or, equivalently,

$$\sin(\Omega - \Omega_\odot) = w/(1-z^2)^{1/2}(1-x^2)^{1/2}\,. \tag{76b}$$

From Equations 75 and 76 we see that both $\sin(\Omega - \Omega_\odot)$ and $\sin 2(\Omega - \Omega_\odot)$ are proportional to w. We shall be concerned with averaging Equations 44, 45, 49, and 52 on the intermediate time scale. We note from Equations 55 and 69 through 74 that the expressions to be averaged are functions of x, y, and z multiplied by $\sin\epsilon$ and $\sin[m(\Omega_\odot - \Omega) - \epsilon]$. The latter term may be expanded as $\sin m(\Omega_\odot - \Omega)\cos\epsilon - \cos m(\Omega_\odot - \Omega)\sin\epsilon$. Since $\sin m(\Omega_\odot - \Omega)$ is proportional to w, the terms multiplied by $\sin m(\Omega_\odot - \Omega)\cos\epsilon$ vanish on averaging (because w has odd parity relative to x, y, and z). Thus, all terms yielding nonvanishing averages are proportional to $\sin\epsilon$.

V. INTEGRATION OF THE EQUATIONS OF MOTION

As described previously, our plan of solution calls for successive averaging of the equations of motion on the short and intermediate time scales. The averaging on the short time scale was performed analytically in the derivation of the precessional equations (see Section II). The precessional equations, however, must be numerically integrated. For reasons described in Section II, we chose to integrate simultaneously the three first-order differential equations for y, z, and w (Equations 35, 36, and 38) using the conservation law for Λ (Equation 32) to eliminate x. Equations 33 and 37, for χ and w, may then be used as checks on the accuracy of the integration. We have shown that the precessional motions are periodic, with the half-period corresponding to the time interval between consecutive zeros of w. Furthermore, a change in sign of w produces a sign change in the time derivatives of x, y, and z. Thus, the variations of x, y, and z are symmetric with respect to the zeros of w.

Before integrating the precessional equations, we must determine a set of initial data for x, y, z, and w. We choose the initial value of w to be zero. Then, using the equations for χ, Λ, and w (Equations 32, 33, and 37), we may eliminate x and y and so obtain a single sixth-order polynomial in z. This polynomial has two real roots corresponding to the two possible coplanar configurations of a, b, and c. One of these real roots for z is chosen as the starting value, and the corresponding initial values of x and y are computed from it. The large differences in magnitude of K_1, K_2, and L make it essential that the polynomial be solved using double precision arithmetic. Otherwise, the loss of significance during solution would make the computed roots inaccurate by up to several percent. Since the coefficients of the polynomial, and hence its roots, change only very slightly during one step on the long (tidal) time scale, we may solve for the roots at each stage using Newton's method to correct the previous values of the roots. This procedure is much faster and simpler than solving the polynomial from scratch at each stage. The original values of the roots are determined from the initial data on the Moon's orbit.

Once the initial data have been determined, we integrate the precessional equations using a Runge-Kutta method for simultaneous first-order equations. We wish to integrate these equations over one-half the precessional period. The precessional period is not, however, a known quantity; therefore, we integrate, keeping track of the sign of w, until w has changed sign. We then interpolate back to find the exact location of the zero of w and hence the value of the half-period. The values of x, y, z, and w as functions of time are stored in the computer. These values are then recalled and used in averaging the equations of tidal friction.

A complete set of tidal equations is given by the expressions for dH/dt, $d\Lambda/dt$, da/dt, and $d\chi/dt$ (Equations 44, 49, 50, and 52). Together with these equations, we use the expressions for $\mathbf{T}\cdot\mathbf{a}$, $\mathbf{T}\cdot\mathbf{b}$, and $\mathbf{T}\cdot\mathbf{c}$ (Equations 59) and either MacDonald's or Kaula's expressions for the tidal torques. The equations of tidal friction are averaged on the intermediate time scale, using the stored values of x, y, z, and w obtained by integration of the precessional equations. The numerical averaging is performed by the use of Simpson's rule.

The integration of the tidal friction equations (for H, Λ, a, and χ) is performed by means of a four-step Runge-Kutta method suitable for simultaneous first-order equations. Each complete step on the tidal time scale involves four intermediate steps. For each intermediate step, the precessional equations must be initialized and then integrated, and the tidal friction equations must be averaged. Thus each step on the long time scale involves four solutions of the sixth-order polynomial, four integrations of the precessional equations, and four averages of the equations expressing the tidal rates of change. The size of the time step on the long time scale is variable and is determined by controlling the proportional error. In this procedure, the increments of H, Λ, a, and χ during a single time step are compared with the corresponding increments calculated by dividing the time step into two equal half-steps. Let us denote by Δ and Δ' the increments calculated in these two ways. Then, the proportional error during a single time step is defined as

$$ER = \left|\frac{\Delta H - \Delta' H}{H}\right| + \left|\frac{\Delta\Lambda - \Delta'\Lambda}{\Lambda}\right| + \left|\frac{\Delta a - \Delta' a}{a}\right| + \left|\frac{\Delta\chi - \Delta'\chi}{\chi}\right|. \tag{77}$$

In our integrations, the step size was varied so that ER remained within specified bounds. Typically, the upper and lower bounds on ER were set at 10^{-6} and 10^{-7}. The computations involved in a single time step take about 20 seconds on an IBM 7094 computer, with two-thirds of this time being spent in computing the two half-steps, which are only used to calculate ER. During the 20 seconds, the precessional equations are initialized and integrated, and the tidal equations are averaged 12 separate times.

The results of several numerical integrations are displayed in Figures 3 through 10. In all cases, the lunar semimajor axis has been plotted along the abscissa. We have deliberately chosen not to plot against time, since the time scale merely reflects the input to the calculation and may be accurately determined analytically once this input is known. More precisely, the time scale is determined by the present value of Q (which is derived from observations of the Moon's secular acceleration) and by some assumed extrapolation used to obtain a Q-value at earlier epochs. Thus, the

puzzle presented by the Moon's apparently short time scale (MacDonald, 1964) is not answered by these calculations.

In general, the numerical results are not very significant when carried back to $a < 3R_\oplus$, owing to our neglect of the Moon's orbital eccentricity. As was shown by MacDonald (1964), the eccentricity increases rapidly if the integrations are carried back to within this distance. The problem of handling eccentric orbits within the framework of the present calculational method is discussed in Section VI. Several of the integrations were also run in the forward direction (i.e., into the future) starting at $a = 60.2R_\oplus$. In the most realistic case, where solar tides were included, it was found that, when a reached $75R_\oplus$, the length of the day became equal to 1 month. The calculations are not shown beyond this point, but several alternative paths of evolution may be distinguished.

The most likely possibility is that, once the day and month become equal, the synchronous rotation of the Earth (with the Moon's revolution) would be stabilized for all future times by any residual nonaxisymmetric shape of the Earth's figure. In this way, the Earth's axis of least inertia would always point toward the Moon, just as the Moon's axis of minimal inertia points toward the Earth now. This relation would be maintained as the solar tides continued to attempt to brake the Earth's rotation. As the Sun took angular momentum from the Earth-Moon system, however, the Earth would actually spin up, and the Moon would approach the Earth, since the Moon's moment of inertia (about the center of the Earth) is larger than the Earth's. Two possible alternatives to this evolutionary path should be mentioned.

First, it is possible that the Earth's spin might become locked into a higher-order resonance with the Moon's orbital angular velocity before reaching the synchronous state. These resonances have been discussed in connection with the rotations of Mercury and Venus (Colombo and Shapiro, 1965; Goldreich and Peale, 1966). In this case, the resonance would be maintained, and as in the synchronous case, the Earth would commence to spin up, while the Moon would spiral in toward the Earth.

Second, if the Earth failed to lock into a resonance, the solar tidal torque would cause the day to lengthen beyond 1 month. At this stage, the lunar tides would reverse their effects and begin to transfer angular momentum to the Earth's spin from the Moon's orbital motion. Since even at the maximum distance of $a = 75R_\oplus$, the lunar tidal torque will exceed the solar one, and since the Moon's orbital moment of inertia exceeds the Earth's, the Earth will begin to spin up and the Moon will spiral in, maintaining the day at just slightly longer than 1 month.

Unlike the case in which the day is locked at 1 month by the axial asymmetry of the Earth, the day must be slightly longer than 1 month in order to enable the lunar tides to transfer angular momentum from the Moon's motion into the Earth's spin.

As before, we begin with the calculations based on MacDonald's expressions for the tidal torques. In these calculations, the solar tides are neglected because of the previously mentioned difficulty in properly treating the dual solar-lunar tidal interactions within MacDonald's formulation. In Figure 3, we have plotted the obliquity of the Earth's equator to the ecliptic. The two branches of the curve represent the maximum and minimum values that the obliquity attains during each precession period. From the figure, it is evident that when the Moon is at its present distance of $60R_{\oplus}$, the obliquity barely varies as the Earth precesses. If, however, the Moon were ever much closer to the Earth, the stonger interaction between the Earth's figure and the Moon would have produced a large variation in the obliquity.

In Figure 4, we illustrate the inclination of the Moon's orbit to the ecliptic. The variation between the maximum and minimum values of the inclination is very small

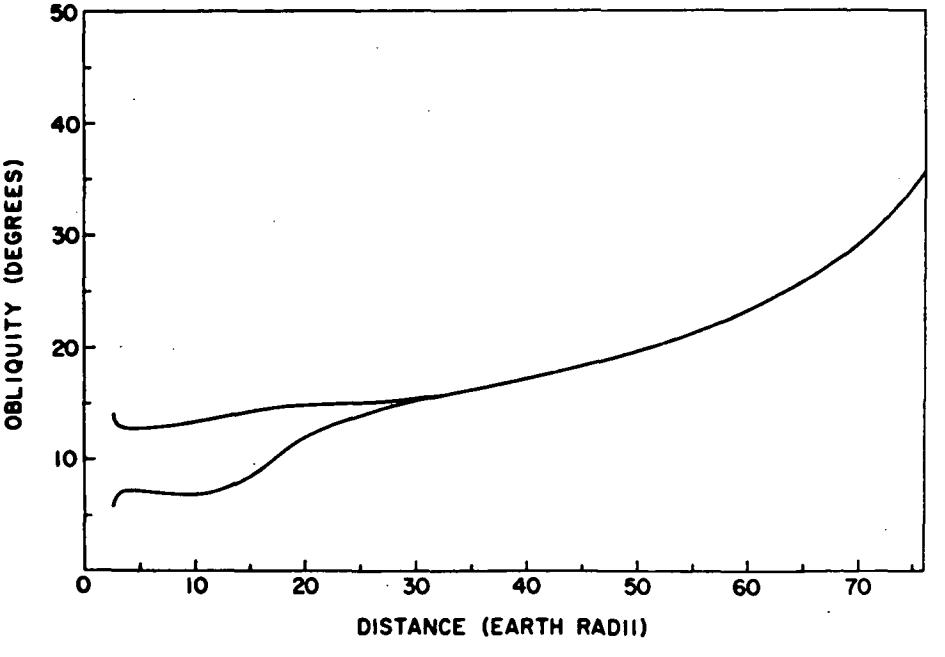


Figure 3.–The obliquity of the Earth's equator to the ecliptic. MacDonald's torques. No solar tides.

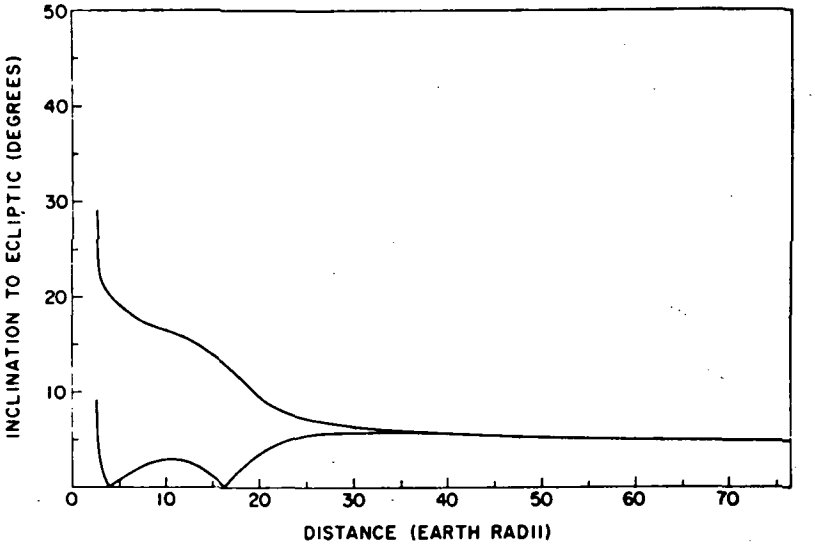


Figure 4.–The inclination of the Moon's orbit to the ecliptic. MacDonald's torques. No solar tides.

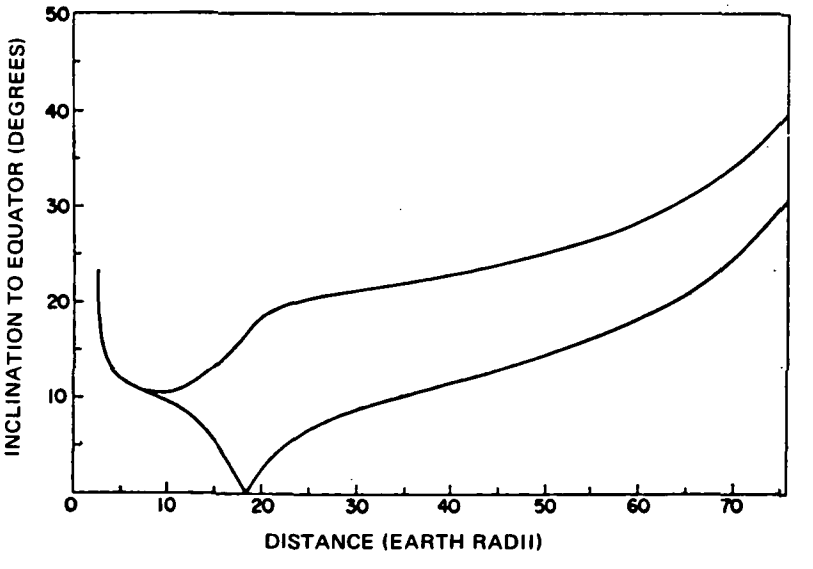


Figure 5.–The inclination of the Moon's orbit to the Earth's equator. MacDonald's *torques. No solar tides.*

at present, but it would have been much more pronounced if and when the Moon was much closer to the Earth.

The most important results are displayed in Figure 5. As was discussed in Section III, we see that, for $a \leqslant 10R_{\oplus}$, the lunar inclination to the equator plane varies by less than 1 degree. The inclination, however, never drops below 10 degrees (when $a \leqslant 10R_{\oplus}$), thus appearing to rule out theories that postulate the separation of the Moon from the Earth. Furthermore, the calculations are also inconsistent with the idea that the Moon accreted from an equatorial disk of particles orbiting the Earth, as was suggested by the discussion in Section III.

A similar calculation has been performed using the torques developed by Darwin. For the purpose of comparison with the previous results, solar tides were neglected. The results of this computation are so similar to those just described that the two sets of graphs could hardly be distinguished when viewed side by side. Thus, they are not reproduced here.

The next set of calculations that are described were made using Darwin's torques, and they included the solar tides (Figures 6 through 8). The tidal phase lags were all set equal in these calculations. The principal difference between the results of this trial and the one in which the solar tides were neglected is seen in the graph of the obliquity. Owing almost entirely to the interactions between the Sun and the lunar tide and between the Moon and the solar tide, the Earth's obliquity decreases less rapidly as we tract the Earth-Moon system back into the past. Otherwise, the graphs are quite similar to the previous ones. Of particular interest, we note that the limiting inclination of the Moon's orbit to the equator never drops below 10 degrees when $a \leqslant 10R_{\oplus}$. Additional graphs showing the number of hours per day and the precession period (between every second coplanar configuration of the unit vectors **a**, **b**, and **c**) are shown in Figures 9 and 10. It is seen that, when $a \leqslant 5R_{\oplus}$, the precession period becomes less than 1 year. At these close Earth-Moon separations, however, the solar interactions are entirely negligible, and our method of averaging remains valid.

Many additional computer runs have been made with varying assumptions about the tidal torques. The major conclusions can be briefly summarized. It was found that setting the phase lags (ϵ_i) proportional to the tidal frequency led to only slight changes from the results obtained with ϵ_i equal to a constant. Also, the introduction of arbitrary changes of obliquity (on the tidal time scale), such as those which might arise from a change in the Earth's moment of inertia or from internal energy dissipation due to core-mantle slippage, does not affect our principal conclusion. That is, no matter how the obliquity is altered, the inclination of the

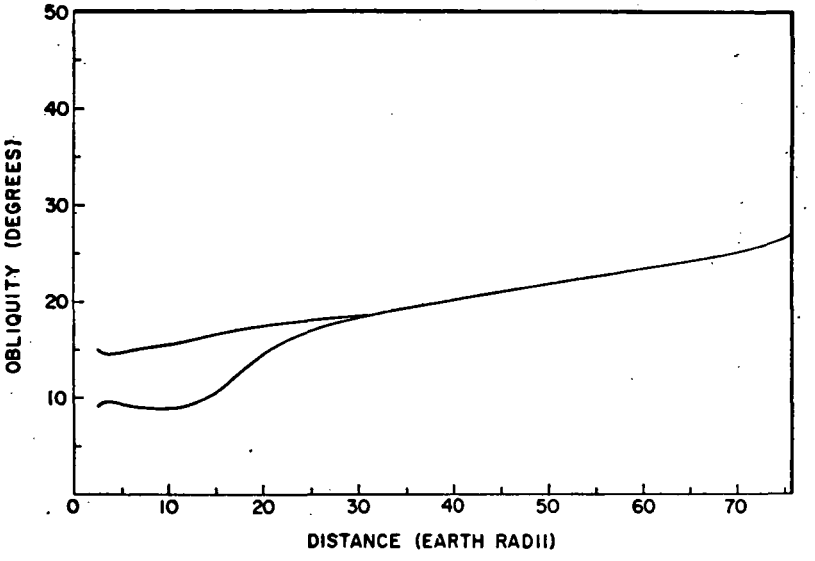


Figure 6.–The obliquity of the Earth's equator to the ecliptic. Darwin's torques; solar tides included.

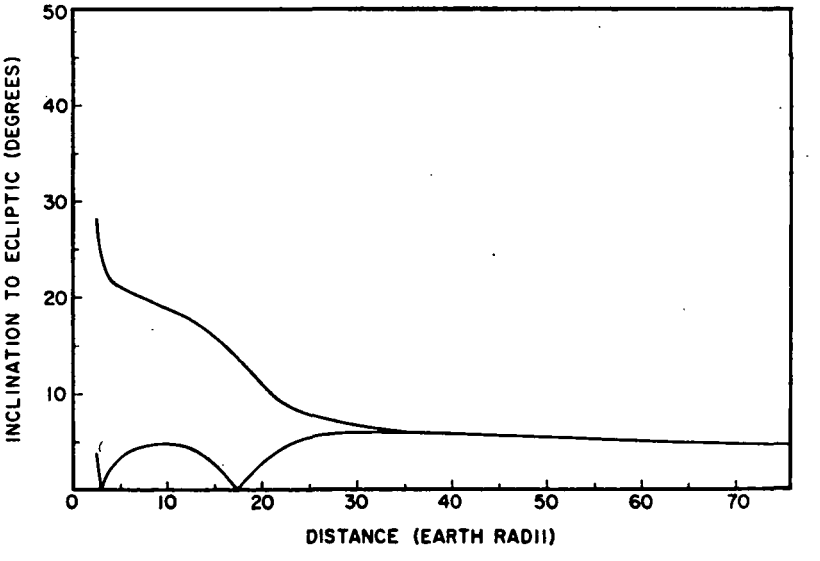


Figure 7.–The inclination of the Moon's orbit to the ecliptic. Darwin's torques; solar tides included.

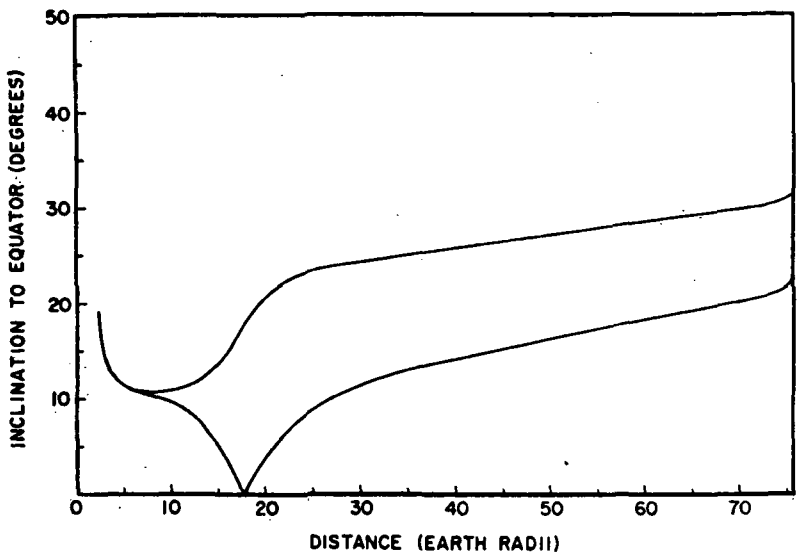


Figure 8.—The inclination of the Moon's orbit to the Earth's equator. Darwin's torques; solar tides included.

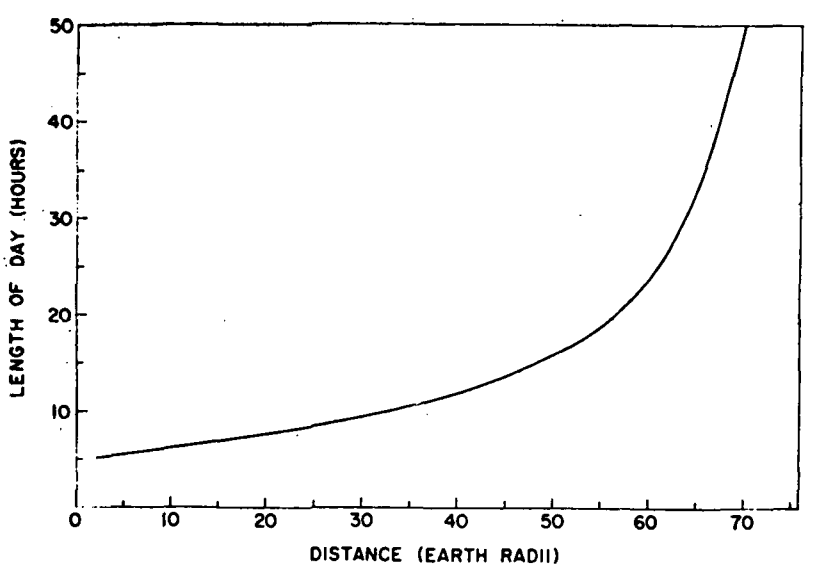


Figure 9.—The length of day. Darwin's torques; solar tides included.

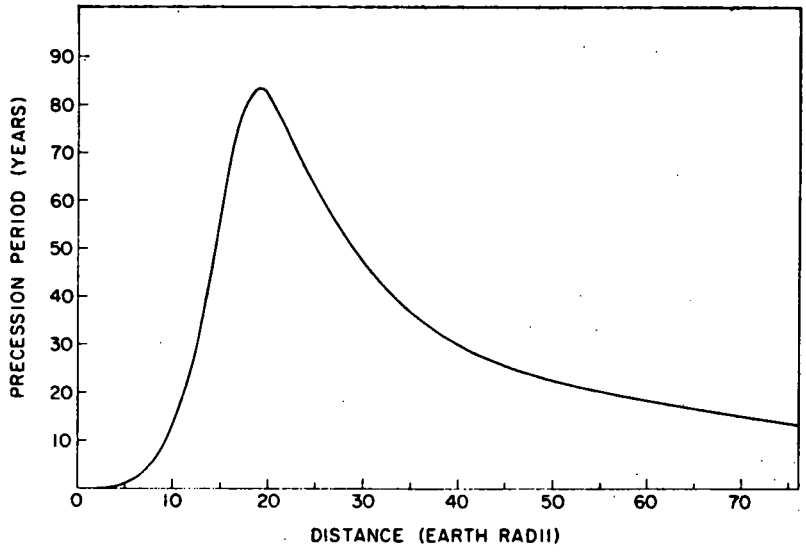


Figure 10.—The precession period. Darwin's torques; solar tides included.

lunar orbit to the Earth's equator plane for $a \leqslant 10R_{\oplus}$ is almost unaffected. In all cases, this inclination dropped to about 10 degrees when $a = 10R_{\oplus}$ and then increased as a decreased.

In an attempt to gain a better understanding of the relation between the present lunar inclination to the ecliptic and the inclination to the equator when $a = 10R_{\oplus}$, several additional computer trials were performed. For each run, the present parameters of the Earth-Moon system served as initial data, except for the inclination of the orbit to the ecliptic plane, which was varied between 1 and 10 degrees in different runs. The general (and surprisingly accurate) rule of thumb that emerged from these numerical experiments was that for every degree the Moon was placed out of the ecliptic at $a = 60R_{\oplus}$, it was 2 degrees out of the equator plane at $a = 10R_{\oplus}$. Thus, the present 5-degree ecliptic inclination implies an equatorial inclination of about 10 degrees at $a = 10R_{\oplus}$. This conclusion is maintained even when rather wild arbitrary variations in the Earth's obliquity are introduced into the computations. Furthermore, the general nature of this conclusion does not appear to depend sensitively on the amplitude and frequency of the planet's Q.

For the reasons just cited, it appears that the origin of the Moon by separation from the Earth or by formation in an equatorial orbit would have led to an ecliptic

orbit at the present time. In view of the present 5-degree inclination of the lunar orbit to the ecliptic, it is tempting to assert that these theories of formation must now be discarded. We shall resist this temptation, however, at least until the present form of the tidal torque is more accurately determined.

VI. DISCUSSION AND CONCLUSIONS

In this section we shall devote some time to the discussion of possible improvements that might be introduced into our calculations. We shall also attempt to indicate some of the major uncertainties that plague our results and how they may eventually be solved. Whereas it is inevitable that some evaluation of the current theories of lunar origin will be presented here, it is not our intention to retread the ground covered by MacDonald (1964); rather, we shall concentrate on those points for which our calculations have special relevance.

A. Lunar Orbital Eccentricity

We have chosen to ignore the lunar orbital eccentricity throughout our calculations. The difficulty in rigorously extending our calculational method to include the eccentricity arises in the treatment of the precessional equations. If a noncircular lunar orbit is considered, after averaging over the short time scale we are left with a spinning elliptical mass distribution instead of a circular one. The precessional torques are now dependent on the orientation of the Moon's perigee with respect to the ecliptic and equator planes. Thus, the precessional equations no longer admit the simple periodic solutions that held for the case of zero eccentricity. Unfortunately, there is apparently no good approximate solution to these equations when the eccentricity is not zero, since the motion of the perigee is comparable to that of the node. This difficulty has been faced but not resolved (Darwin, 1879 and 1880; MacDonald, 1964): MacDonald assumes that the perigee precesses uniformly with respect to the Moon's node on the ecliptic, although this is certainly not the case; Darwin proceeds, as we do, neglecting the influence of the eccentricity on the evolution of the other orbital parameters.

Athough we have chosen to ignore the lunar orbital eccentricity, it might still be included more or less rigorously in calculations of the sort we have performed. As mentioned in the introduction, it appears likely that the lunar orbital eccentricity is

increasing at the present time and, hence, in the past was even smaller than the current value of $e = 0.055$. In this case, we would be justified in retaining only terms linear in e in our equations. Within this approximation, the precessional equations would be independent of eccentricity, and our method of averaging would still be valid. When our integrations brought the Moon back to its closest approach and the eccentricity began to increase, we would have to include higher-order terms in the eccentricity. Near close approach, however, the solar precessional torques may be ignored, and the precessional motion for a finite eccentricity orbit could probably be handled. We have not attempted to carry out this procedure, because we doubt that our results would differ greatly from those already obtained by MacDonald. The only differences we think we might find are those which arise from the differences between Darwin's and MacDonald's expressions for the tidal torque.

B. Tidal Torque

In our investigation, we have used theoretically derived relations for the tidal torque. Any confidence we place in the results obtained is based less on our confidence in the validity of these torques than on the apparent insensitivity of our results to changes in them. Needless to say, we would feel much more secure if the present tidal torque could be observationally determined. In particular, we cannot exclude the possibility that strong local dissipation in a few places in the Earth's oceans or crust may lead to unanticipated deviations from our results. Thus, a future determination of the actual tidal torque will be awaited with great interest.

C. Theories of Lunar Origin

1. Fission Theories

Fission theories suffer from three substantial difficulties. First, special conditions appear to be necessary to have initiated the fission. Specifically, considering the angular momentum currently in the Earth-Moon system and adding to it the angular momentum which the solar tides have taken out during geologic time, we find that total is insufficient to have made the Earth rotationally unstable. Second is the difficulty of a short time scale for the subsequent evolution of the

Earth-Moon system. Third, such theories would inevitably lead to an initial equatorial orbit for the Moon. As we have seen, the present lunar orbit appears to be irreconcilable with an equatorial orbit in the past.

2. *Formation in Orbit About the Earth*

Our discussion of the origin of natural satellites (Section III) together with the results of the numerical integrations implies that if the Moon had accreted from particles orbiting within $10R_{\oplus}$, its initial orbit would have been in the equatorial plane. Furthermore, our calculations would then indicate that the Moon should now lie in the ecliptic plane. Thus, the current 5-degree inclination of the Moon's orbit to the ecliptic argues against this mode of formation for the Moon. If the Moon had formed by accretion from particles lying outside $30R_{\oplus}$, we should expect these particles to have formed a disk in the ecliptic plane. This mode of formation should also have led to an ecliptic lunar orbit at the present time. Thus, the possibility that the Moon formed by accretion appears to require that most of the accreted mass originally lay between $10R_{\oplus}$ and $30R_{\oplus}$. Accretion in this region would have been hampered, however, by the inability of the particles to concentrate into a thin disk. Finally, all accretion theories suffer from the aforementioned problem of the short time scale for tidal evolution of the Earth-Moon system.

3. *Capture Theories*

It does not seem possible to exclude the idea that the Moon was captured in a highly inclined (≈90-degree), highly eccentric orbit of the sort discussed by MacDonald (1964). As emphasized there, however, this form of the capture theory requires the Earth to have possessed an initial angular momentum density that was inordinately high. The case against capture in a slightly inclined direct orbit is quite different. Now, angular momentum is no longer a problem. We would not, however, have expected the Moon to have been captured in an orbit as nearly circular as the present one ($e \approx 0.055$). Thus, we arrive at a contradiction, since tidal friction would almost certainly have acted to increase the eccentricity above the value it had at capture. As discussed in Section I, in order for the lunar orbital eccentricity to be decreasing at the present time, a Q is required for the Moon which is smaller than

that of the Earth. Indeed, from Goldreich (1963), Kaula (1964), or MacDonald (1964), we see that Q for the Moon must not exceed about two-thirds the value of the Earth's Q. It seems highly unlikely that the lunar Q could satisfy this restriction, especially since the low value of the Earth's tidal effective Q [$Q \approx 13$ (MacDonald, 1964)] appears to be at least partially due to tidal dissipation in the shallow seas (Miller, 1966).

In conclusion, we have found that our calculations do not strongly support any of the current theories of the origin of the Moon. If the time-scale difficulty is ignored, the most attractive possibility appears to be formation by accretion between $10R_\oplus$ and $30R_\oplus$. This proposal has the advantage of satisfying current estimates of the Earth's initial angular momentum density (at least for accretion at the outer range, near $30R_\oplus$); also, it does not clash with the eccentricity and inclination of the present lunar orbit. On the negative side, since it is impossible to form a thin disk of particles in this region, accretion is made less plausible.

ACKNOWLEDGMENTS

The calculations presented here are the answer to an examination question posed to me by Thomas Gold in June 1963. I am indebted to him for the remarkable patience he showed in awaiting my reply. I have also benefited from frequent discussions with Gordon J. F. MacDonald and from expert assistance in programming by David Ross.

REFERENCES

Allen, C. W., *Astrophysical Quantities*, Athlone Press, London, 1963.

Brouwer, D., and Clemence, C. M., *Methods of Celestial Mechanics*, Academic Press, New York, 1961.

Colombo, G., and Shapiro, I. I., "The Rotation of the Planet Mercury", Smithsonian Astrophysical Observatory Special Report 188R, 1965.

Darwin, G. H., "On the Precession of a Viscous Spheroid and on the Remote History of the Earth", *Phil. Trans. Roy. Soc. London* 170:447-530, 1879.

Darwin, G. H., "On the Secular Change in the Elements of the Orbit of a Satellite Revolving About a Tidally Distorted Planet", *Phil. Trans. Roy. Soc. London* 171:713-891, 1880.

Gerstenkorn, H., "Über Gezeitenreibung beim Zweikörpenproblem", *Z. Astrophys.* 26:245-274, 1955.

Goldreich, P., "On the Eccentricity of Satellite Orbits in the Solar System", *Mon. Notic. Roy. Astron. Soc.* 126:257-268, 1963.

Goldreich, P., "Inclination of Satellite Orbits About an Oblate Precessing Planet", *Astron. J.* 70:5-9, 1965.

Goldreich, P., and Peale, S. J., "Resonant Spin States in the Solar System", *Nature* 209:1078-1079, 1966.

Goldreich, P., and Soter, S., "*Q* in the Solar System", *Icarus* 5:375, 1966.

Kaula, W. M., "Tidal Dissipation by Solid Friction and the Resulting Orbital Evolution", *Rev. Geophys.* 2:661-685, 1964.

Kozai, Y., "The Motion of a Close Earth Satellite", *Astron. J.* 64:367-377, 1959.

Lyttleton, R. A., "On the Possible Results of an Encounter of Pluto With the Neptunian System", *Mon. Notic. Roy. Astron. Soc.* 97:108-115, 1936.

MacDonald, G. J. F., "Tidal Friction", *Rev. Geophys.* 2:467-541, 1964.

Miller, G., "The Flux of Tidal Energy Out of Deep Oceans", *J. Geophys. Res.* 70:2485-2489, 1966.

Munk, W. H., and MacDonald, G. J. F., *The Rotation of the Earth*, Cambridge University Press, New York, 1960.

☆ U. S. GOVERNMENT PRINTING OFFICE : 1972 721-813/436