Cometary Physics

W.-H. Ip

Institutes of Astronomy and Space Science, National Central University, Taiwan

GLOSSARY

a Semimajor axis of an orbit.

AU Astronomical unit $= 1.496 \times 10^{13}$ cm, which is the semimajor axis of the earth's orbit around the sun.

Blackbody An idealized perfectly absorbing body that absorbs radiation of all wavelengths incident on it.

Bow shock A discontinuity defining the interface of the transition of a supersonic flow to a subsonic flow as it encounters an obstacle.

Carbonaceous chondrites A special class of meteorites containing up of a few percent carbon.

Contact surface A surface separating two flows of different chemical and/or flow properties.

Comet showers The infrequent injection of a large number of new comets into the inner solar system by passing stars at close encounters with the solar system.

CONTOUR An American space mission to perform flyby observations of comets Encke, Schwassmann-Wachmann 3, and d'Arrest with the spacecraft to be launched in 2002.

DE (disconnection event) The major disruption of a cometary ion tail by the apparent ejection of large plasma condensation or separation of the main ion tail from the comet head.

Deep Impact An American space mission to comet 9P/Tempel 1 for active Impact experiments. The scheduled launch date is in 2004.

Deep Space 1 An American Spacecraft with ion propulsion to flyby comet Borrelly in 2002.

eV Electron volt $= 1.6 \times 10^{-12}$ ergs.

Giotto A spacecraft from the European Space Agency for flyby observations of comet Halley in March 1986.

HST Hubble Space Telescope, an optical telescope launched into geocentric orbit by NASA in 1990.

ICE International Cometary Explorer, an interplanetary spacecraft (ISEE 3), which was redirected to interception of comet Giacobini/Zinner in September 1985.

IHW International Halley Watch, an international program to coordinate the ground-based observations of comet Halley.

IMS Ion mass spectrometer that can make mass-separating measurements of ions.

ISO Infrared Space Observatory, an infrared telescope launched into geocentric orbit by ESA in 1993.

Jet force The repulsive force on a cometary nucleus due to anisotropic gas emission. The jet force was invoked to explain the nongravitational effects observed in cometary orbital evolutions.

keV Kilo electron volt $= 10^3$ eV.

KT boundary The Cretaceous–Tertiary boundary separating two distinct geological structures about 65 million years ago. This boundary is characterized by a thin layer of enhanced iridium deposition that might have come from an impact event of a comet or an asteroid of kilometer size.

Mass extinction Episodes of large-scale disappearances of life forms on earth. Statistical correlations suggest a periodicity of about 26 million years.

Meteoroid A small solid particle orbiting around the sun in the vicinity of the Earth.

NMS Neutral mass spectrometer, which can make mass-separating measurements of neutral gas.

nT Nanotesla (10^{-9} T) units of magnetic field strength.

Oort cloud The reservoir of new comets at large distances (10^4–3×10^4 AU) from the sun.

PAH Polycyclic aromatic hydrocarbons, which might be representative of the very small grains in interstellar space.

POM Formaldehyde polymers of the form of $(H_2CO)_n$, with $n = 10$–100. The chainlike polymers could be terminated and stabilized by the addition of monovalent atoms or ions (i.e., HO–CH_2–O–..CH_2–CN).

pc Parsec: 1 pc $= 3.086 \times 10^{18}$ cm.

Rosetta A European mission designed to collect surface material from the short-period comet P/Wirtanen with a lander to be launched from a mother spaceship orbiting around the cometary nucleus. The spacecraft will be launched in 2003.

Sakigake A Japanese spacecraft for remote-sensing measurements at comet Halley in 1986.

Solar nebula The primordial disk of gas and condensed matter befor the formation of the planetary system.

Solar wind A radial outflow of ionized gas (mostly protons) from the solar corona. The mean number density of solar wind at 1 AU solar distance is 5 cm^{-3}, average speed at earth is 400 km s^{-1}, and the mean electron temperature is 20,000 K.

Stardust An American sample return mission to flyby comet Wild 2 for dust collection. The comet encounter will be in 2004 and the sample will be parachuted back to the earth in 2006.

Suisei A Japanese spacecraft for *in-situ* plasma measurements at comet Halley in 1986.

VEGA Two Soviet spacecraft dedicated to close flyby observations of comet Halley in 1986 after performing a series of measurements at Venus swingbys.

COMETS are known to be the Rosetta Stone to decipher the origin of the solar system. The chemical composition of the volatile ice, the mineralogical properties and isotopic abundances of the dust grains, and finally the physical structure of the cometary nuclei all carry fundamental information on the condensation and agglomeration of small icy planetesimals in the solar nebula. The process of comet–solar-wind interaction is characterized by very complex and interesting plasma effects of special importance to solar system plasma physics; also, the different aspects of momentum exchange, energy transfer, and mass addition in a plasma flow are closely linked to large-scale outflows in interstellar space. For example, the generation of massive outflows from the T Tauri stars and other young stellar objects and their interaction with the surrounding medium are reminiscent of the comet–solar-wind interaction. The formation of small structures called cometary knots in the Helix planetary nebula is reminiscent of the comet–solar wind interaction (Fig. 1). The cometary plasma environment therefore could be considered a laboratory simulating many basic astronomical processes.

I. INTRODUCTION

The historical context of cometary research is an interesting document of progress in science. In Oriental records, comets were usually called guest stars or visiting stars to describe their transient appearances in the celestial sphere. One of the earliest systematic records of comets can be found in the Chinese astronomical history of 467 B.C. The historian Kou King-ting noted in 635 B.C. that cometary tails point away from the sun. In Europe, study can be traced back to Aristotle's postulation that comets should be associated with transient phenomena in the earth's atmosphere. As astronomy started to blossom in the Middle Ages, Tycho Brahe, Kepler, and Newton all paid special attention to the dynamical nature and origin of comets. The study of comets has been closely coupled with scientific and technical progress in the past few centuries. The most important example is comet Halley, which was deduced to be in elliptic orbit around the sun with a period of 75.5 years by Edmond Halley in 1682. His famous

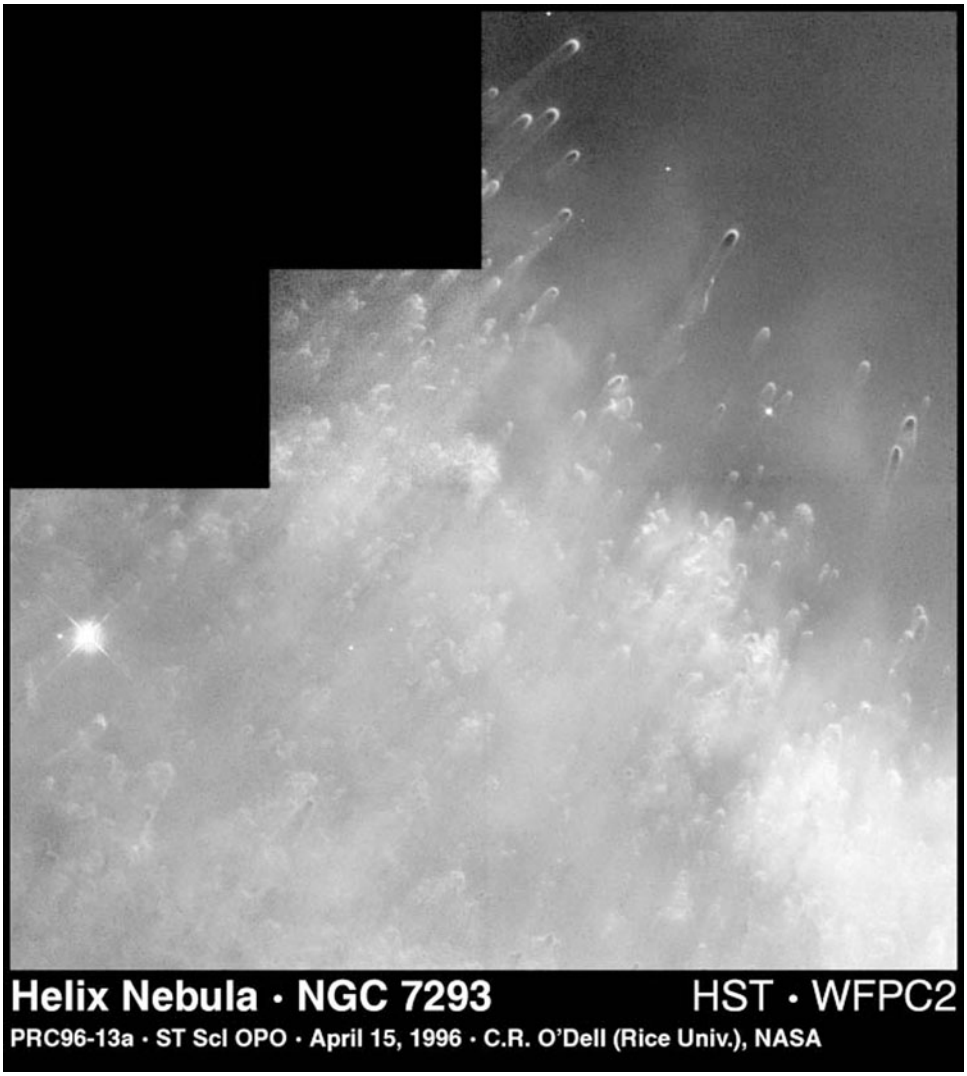


FIGURE 1 Structures of the so-called cometary knots observed at the Helix Nebula (NGC 7293). They are formed by interstellar cloudlets being blown off by intense ultraviolet radiation from the central massive stars. Photo origin: NASA.

prediction that this comet should return in 1758 has become a milestone in astronomy. Three returns later, in March 1986, comet Halley was visited by a flotilla of spacecraft taking invaluable data revealing the true nature of the most primitive bodies in the solar system. In total, six spacecraft from the former Soviet Union, Japan, Europe, and the United States participated in this international effort (see Fig. 2). In addition, an extensive network for ground-based comet observations was organized by NASA to coordinate the study of comet Halley. This program, called International Halley Watch (IHW), proved to be successful in maintaining a high level of cometary research in all areas. The long-time coverages provided by ground-based observations are complementary to the snapshots produced by spacecraft measurements. In fact, investigations of the nucleus rotation, coma activities, and dynamics of the plasma tail all require comparisons of the spacecraft measurements with the observational results gathered by the IHW and other similar programs.

Before the Halley encounters, a first look at the cometary gas environment was obtained by the NASA spacecraft International Cometary Explorer (ICE), at comet Giacobini–Zinner on September 11, 1985. Even though it did not have a scientific payload as comprehensive as the payloads carried by the Giotto and Vega probes, many exciting new observations pertinent to comet–solar-wind interaction were obtained. Because ICE went through the ion tail of comet Giacobini–Zinner, the experimental data are particularly useful in addressing questions about the structures of cometary plasma tails.

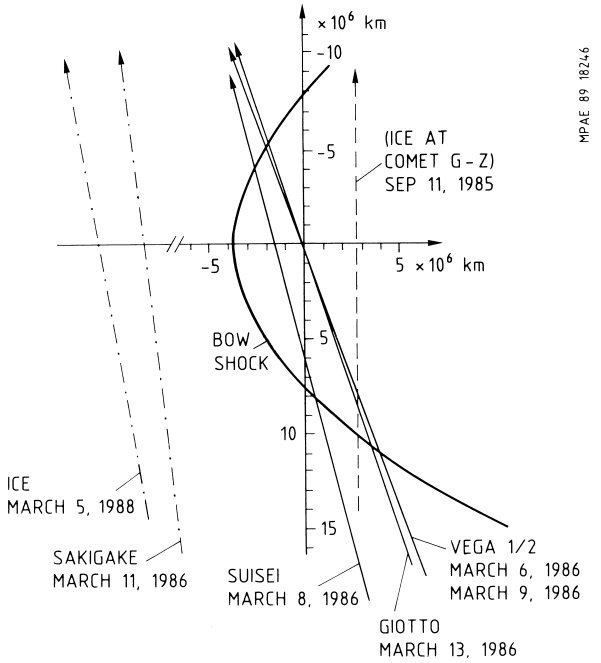


FIGURE 2 A schematic view of the encounter geometries of several spacecraft at comet Halley in March 1986 and the ICE spacecraft at comet Giacobini–Zinner in September 1985.

Because of the retrograde orbital motion of comet Halley, the encounter speeds during flyby observations were very high (68 km s^{-1} for Giotto and 79 km s^{-1} for Vega 1). High-speed dust impact at close approaches to the comet nucleus became a hazard for the spacecraft. This problem caused a temporary loss of radio link of Giotto to the earth when the probe was at a distance of about 1000 km from the cometary nucleus. As a result, no images were obtained beyond that point even though the spacecraft was targeted to approach the comet as close as 600 km on the sunward side. Several other instruments were also damaged. In spite of these calculated risks, the comet missions as a whole achieved major scientific successes beyond expectations.

In other arenas of cometary research, there has been very interesting progress as well. The new concept of an inner Oort cloud between 10^3 AU and 10^4 AU is one of them. In the present work, we shall make use of these many new results to attempt to build a concise picture of modern cometary physics. Because of the page limit, it is impossible to go into all details. For further information, the reader should consult the appropriate references as listed.

II. ORBITAL DYNAMICS

The Keplerian orbit of a comet can be characterized by several orbital elements, namely, the semimajor axis (a), the inclination (i), the eccentricity (e), the argument of perihelion (ω) and the longitude of the ascending node (Ω) with respect to the vernal equinox (γ). Their mutual relations are shown in Fig. 3. The orbital period is given by

$$p = a^{3/2} \text{ years} \tag{1}$$

where a is in units of AU. Thus, for comet Halley, $a = 17.9$ AU and $P = 76$ years. The closest distance to the sun, the perihelion, is given by

$$q = a(1 - e) \tag{2}$$

and the largest distance from the sun, the aphelion, is given by

$$Q = a(1 + e) \tag{3}$$

The total specific angular momentum is expressed as

$$L = \sqrt{\mu a(1 - e^2)} \tag{4}$$

whereas the component perpendicular to the ecliptic plane is

$$L_z = \sqrt{\mu a(1 - e^2)} \cdot \cos i \tag{5}$$

In the foregoing equation, $\mu = GM_\odot$, where G is the gravitational constant and $M_\odot$ the solar mass. The orbit of a comet is prograde (or direct) if its inclination $(i) < 90°$, and retrograde (or indirect) if $i > 90°$.

Because of perturbations from the passing stars, planetary gravitational scattering, and/or jet forces from surface outgassing, the orbital elements can change as a function of time. The orbital elements determined at a certain time (t_0) are called osculating elements for this particular epoch. To infer the future or past orbital elements of a comet, orbital integrations taking into account all relevant perturbation effects must be carried out from $t = t_0$

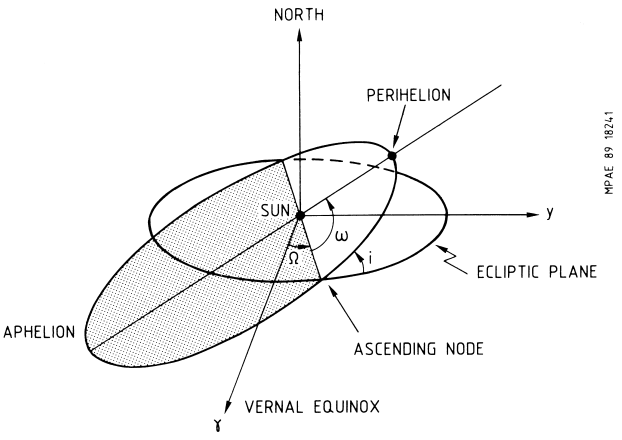


FIGURE 3 Description of a cometary orbit with i denoting the inclination; ω is the argument of perihelion; Ω is the longitude of the ascending node; and γ is the equinox.

to the time interval of interest. Because of limitations of computing machine time, computational accuracies, and observational uncertainties, such calculations usually do not cover a time period much more than a few thousand years if very accurate orbital positions are required. To investigate long-term evolutions, statistical methods are often used such that the orbital behavior of a sample of comets over a time span of millions to billions of years can be described.

Certain invariants in celestial mechanics are useful in identifying the orbital characteristics of small bodies (comets and asteroids) in the solar system. In the restricted three-body problem with the perturbing planet (Jupiter in this case) moving in a circular orbit, the Tisserand invariant is defined as

$$T_J = \frac{1}{a} + 2\left[2q\left(1 - \frac{q}{2a}\right)\right]^{1/2} \cos i \qquad (6)$$

where a and q are in units of Jupiter's semimajor axis, a_J. Most of the short period comets have $T_J < 3$.

Comets are generally classified into three orbital types according to their periods: namely, the long-period comets with $p > 200$ years; the intermediate-period comets with P between 20 and 200 years; and finally, the short-period comets with $P < 20$ years. A compilation of the orbital data of the observed comets shows that there are in total 644 long-period comets 25 intermediate-period comets, and 88 short-period comets. Although such classification is somewhat arbitrary in the divisions of the orbital periods, it makes very clear distinctions in the inclination distribution. The inclinations of the short-period comets are mostly less than 30°, and those of the long-period comets have a relatively isotropic distribution. As discussed below, one active research topic at the present moment is whether these different inclination distributions are also indicative of different dynamic origins of these comet populations.

Note that among the long-period comets, there is a group of sun-grazers, called the Kreutz family, that might have originated from the breakup of a single large comet. A number of them have perihelion distances inside the sun. The SOLWIND and Solar Maximum Mission satellites detected 13 such comets, while the SOHO space solar observatory detected more than 200 of the Kreutz family comets (see Fig. 4).

Figure 5 shows the $1/a$ distribution for the new, long-period comets in units of 10^{-4} AU^{-1}. As first discussed by Oort (1950), the peak at $1/a < 10^{-4}$ AU^{-1} suggests that the new comets mostly come from an interstellar reservoir in the form of a spherical shell surrounding the sun. The inner radius of this so-called Oort cloud is at about 10^4 AU and that of the outer radius at about $3–5 \times 10^4$ AU. Because of the perturbing effects of passing stars, interstellar molecular clouds, and the galactic tidal forces (see Section IX), a continuous influx of new comets from the Oort cloud to the inner solar system can be maintained.

In addition to the classical Oort cloud, made detectable via the injection of a constant flux of new comets, the presence of a massive inner Oort cloud between 10^3 AU and 5×10^3 AU has recently been postulated. Only the very infrequent passages of stars would scatter comets in this region into the inner solar system. Such events could give rise to the so-called comet showers lasting about 2–3 million years at irregular intervals of 20–30 million years.

For a comet in the Oort reservoir, with radial distance $r > 10^4$ AU from the sun, its angular momentum can be expressed in terms of the orbital velocity (V_c) and the angle (θ) between the comet–sun radius vector and the velocity vector (see Fig. 7), that is,

$$r V_c \sin\theta = [\mu a(1 - e^2)]^{1/2} \qquad (7)$$

which yields [cf. Eq. (4)]

$$V_c^2 \theta^2 \approx \frac{2\mu q}{r^2} \qquad (8)$$

for small θ. Assuming that the velocity vectors of comets in the Oort region are sufficiently randomized by stellar perturbations, we can express the distribution function

$$f_\theta(\theta)d\theta = \frac{\sin\theta}{2} d\theta \approx \frac{\theta d\theta}{2} \qquad (9)$$

In combination with Eq. (8) we find that the perihelion distance of new comets should be

$$f_q(q)dq = \frac{\mu dq}{2V_c^2 r^2} \qquad (10)$$

In other words, f_q is independent of q. The population of long-period comets with $1/a \approx 10^{-4}$ AU, however, is a mixture of new comets and the evolved ones with one or more perihelion passages through the solar system. Because of planetary perturbations, the perihelion distribution of the evolved comets would be significantly modified. Numerical simulations show that the frequency distribution of perihelion distance of long-period comets should be highly depleted inside the orbit of Jupiter. It becomes constant only for $q \gtrsim 30$ AU where the planetary perturbation effects are no longer important.

Note that the velocity of a long-period comet near its aphelion in the Oort region is of the order of 200 m s^{-1}. On the other hand, the average speed (V_*) of a passing star relative to the solar system is about 30 km s^{-1}. The effect of stellar perturbation can therefore be treated using the impulse approximation as follows. With closest approach distances given as D_c and $D_\odot$, respectively, the velocity increments received by the comet and the sun can be given as

$$\Delta V_c = \frac{2GM_*}{V_*} \frac{D_c}{D^2} \qquad (11)$$

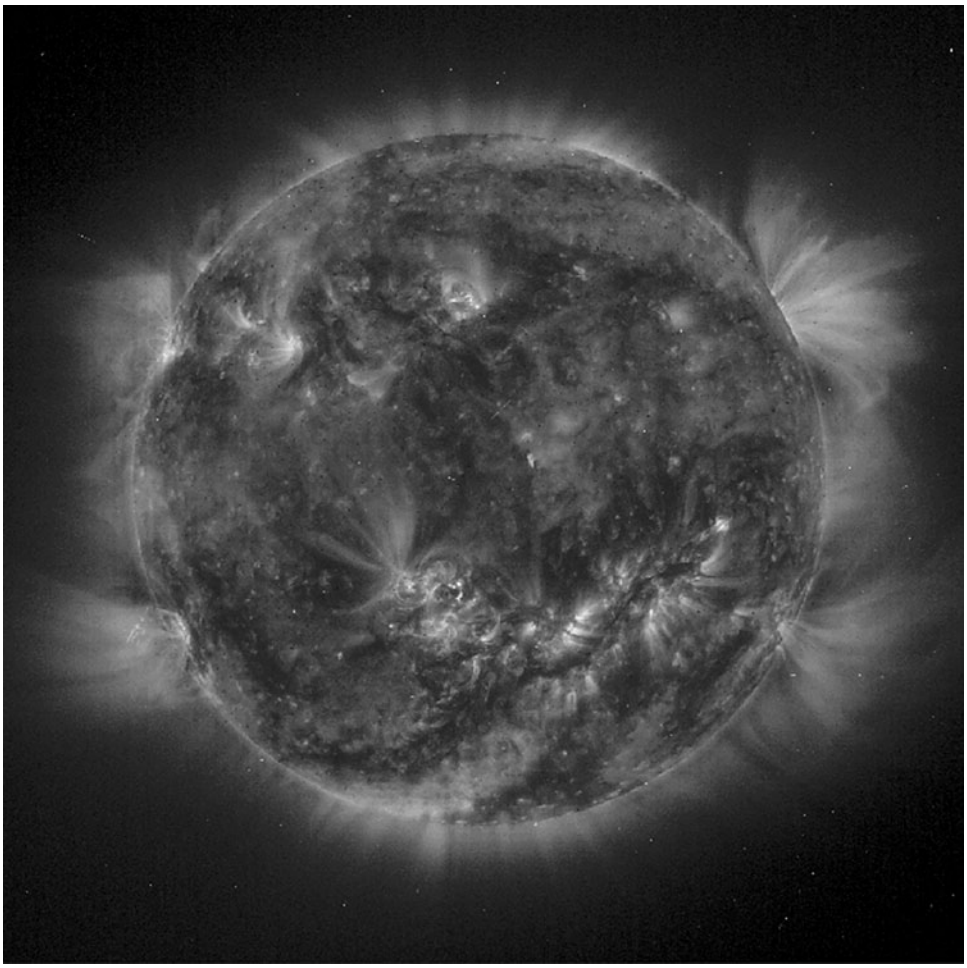


FIGURE 4 The appearance of a sun-grazing comet observed by the SOHO coronograph instrument. Photo origin: ESA.

and

$$\Delta \underline{V}_{\odot} = \frac{2GM_*}{V_*} \frac{\underline{D}_{\odot}}{D_*^2}. \tag{12}$$

The overall effect in the velocity change of the comet relative to the sun is then

$$\delta \underline{V}_{\mathrm{c}} = \Delta V_{\mathrm{c}} - \Delta \underline{V}_{\odot} \tag{13}$$

Over the age of the solar system (4.5×10^9 years), the root-mean-squared velocity increment of a comet as a result of stellar perturbations can be estimated to be $V_{\mathrm{RMS}} \approx 100$ m s^{-1}. As this value is comparable to the orbital velocity of a comet in circular motion at 9×10^4 AU, the outer boundary of the Oort cloud may be set at this distance as a result of stellar perturbations.

As mentioned before, the binding energy of a new comet is of the order of 10^{-4} AU^{-1}. During its passages through the inner solar system, a comet will be subject to gravitational perturbations by the planets. The most important perturbing planet is Jupiter, followed by Saturn. According to numerical calculations, the typical energy change (ΔE) per perihelion passage is a function of the perihelion distance and orbital inclination. For a comet with i between 0° and 30°, $E > 10^{-4}$ AU^{-1} if $q < 15$ AU, and for i between 150° and 180°, $E > 10^{-4}$ AU^{-1} only if $q < 5$ AU. This difference is due to the larger relative velocities of retrograde comets during planetary encounters and resulting smaller orbital perturbations. The excess of retrograde orbits in the frequency distribution of the inclinations of long-period comets may be a consequence of such a selective effect in planetary encounters.

The orbital transformation of long-period comets into short-period comets is one of the possible outcomes of a sequence of random walk processes. The capture efficiency depends on the inclinations and other orbital elements of the comets in question. For example, a "capture" zone exists for long-period comets with 4 AU $< q <$ 6 AU and $i < 9°$. Another possible scenario is that the majority of short-period comets are not supplied by long-period comets with isotropic inclination distribution

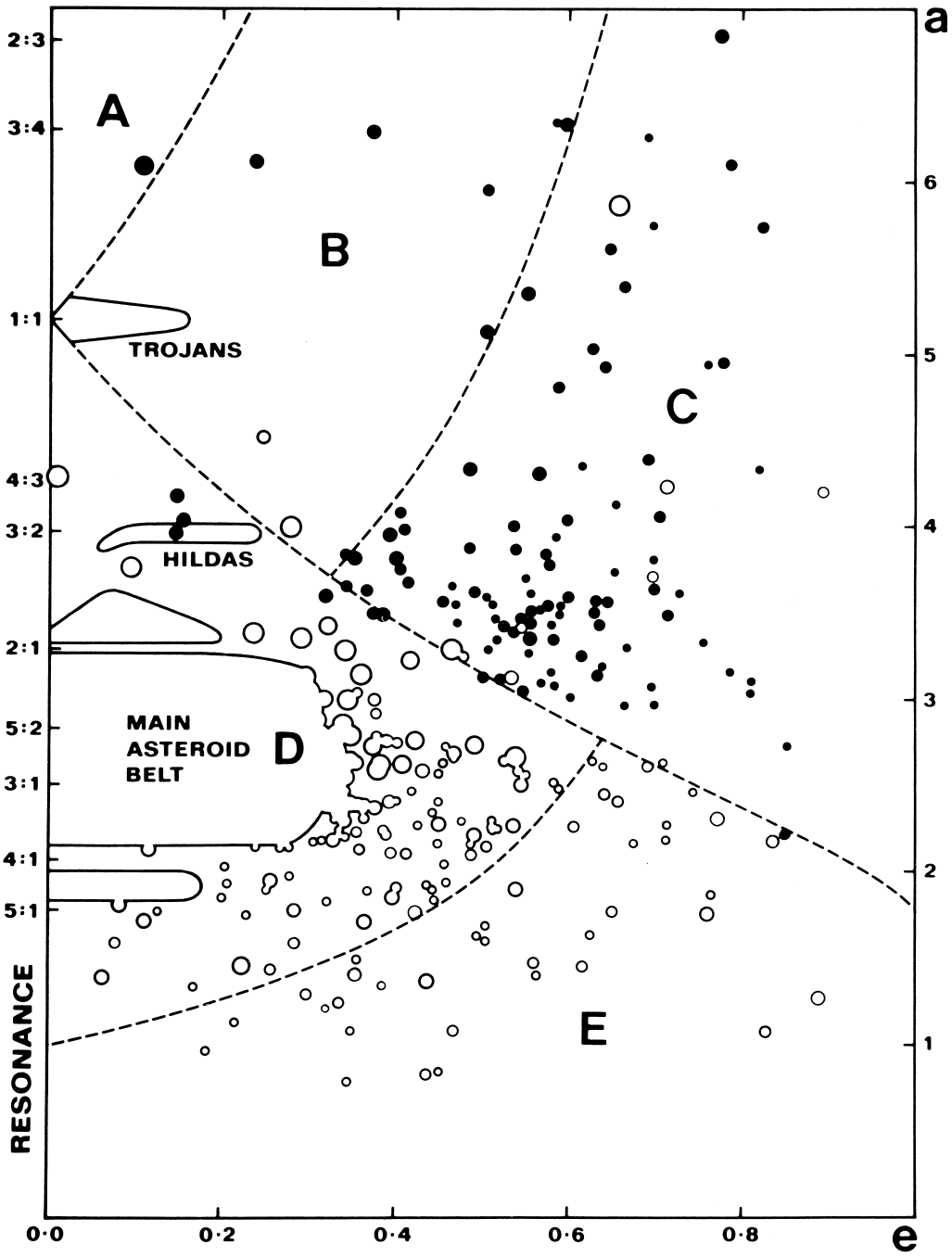


FIGURE 5 Distributions of short-period comets (solid circles) and asteroids (open circles) plotted in a diagram of semimajor axis (a) vs eccentricity (e). A indicates the transjovian region, B is the Jupiter family of weak cometary activity, C is the Jupiter family of strong cometary activity, D is the minor planet region, and E is the apollo region. The thick dashed curve denotes the critical value of $T_J = 3$ (with $\cos i = 1$) separating the cometary region (B + C) with $T_J < 3$ from the asteroidal region (D and E) with $T_J > 3$. [From Kresak, L. (1985). *In* "Dynamics of Comets: Their Origin and Evolution" (Carusi, A., and Valsecchi, G. B., eds.), IAU coll. 83, Reidel, Dordrecht, pp. 279–302.]

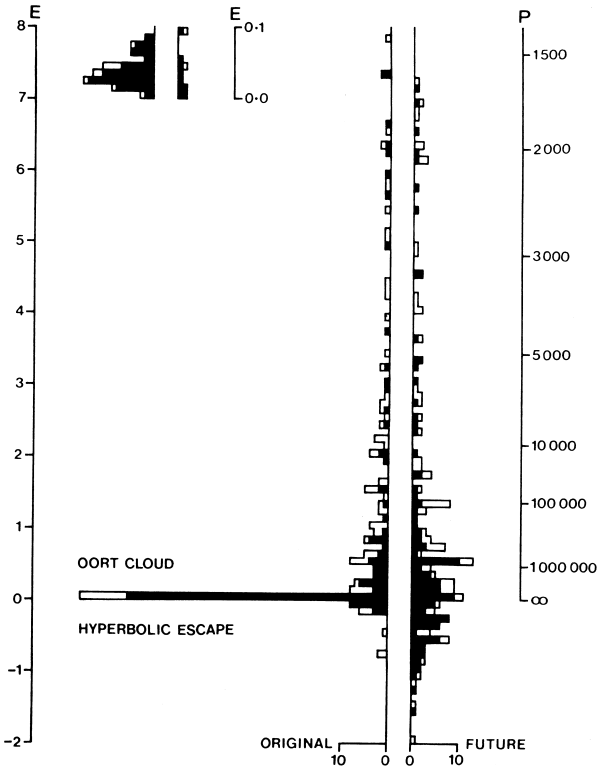


FIGURE 6 Frequency distribution of the original reciprocal semimajor axes of long-period comets with $(1/a)_{\text{orig}} < 10^{-3}$ AU^{-1}. [From Kresak, L. (1987), in "The Evolution of the Small Bodies of the Solar System," (Fulchignoni, M., and Kresak, L., eds.) *Proc. of the Intern. School of Physics* "Enrico Fermi," Course 98, Societa Italiana di Fisica, Bologna-North-Holland publ. co., pp. 10–32.]

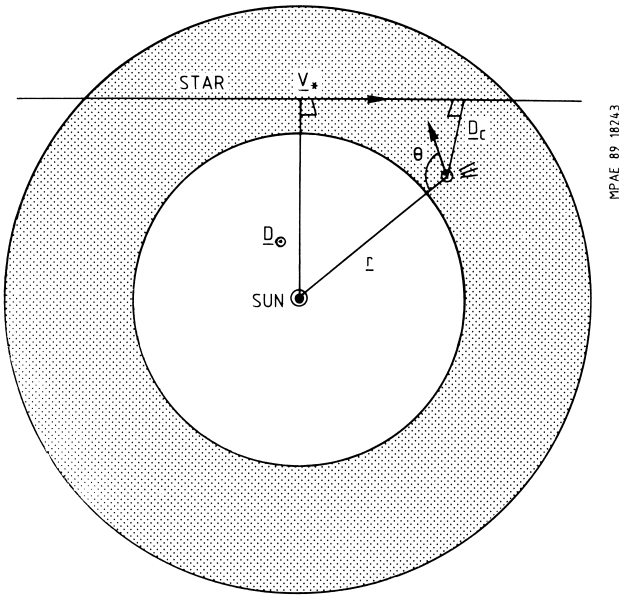


FIGURE 7 The encounter geometry of a passing star with the solar system.

[i.e., $f_i(i)\,\alpha \sin i$], but rather by a comet belt of low inclinations, located just outside the orbit of Neptune.

The dynamical influence of Jupiter can be recognized in the aphelion distribution of short-period comets that tend to cluster near the semimajor axis of Jupiter. Furthermore, the distribution of the longitudes of perihelion (ω) for such a Jupiter family has a mimimum near the perihelion of longitude (ω_J) of Jupiter.

The final fate of the orbital evolution of short-period comets would be determined by perturbation into escape orbit via close encounter with Jupiter or other planets, direct collision with a planet, or catastrophic fragmentation by hypervelocity impact with interplanetary stray bodies or crashing into the sun (see Fig. 4). The fragmentation process would lead to the formation of a meteor stream composed of small dust particles. Meteor streams could also be produced by partial fragmentation, surface cratering, and, of course, outgassing activities. For example, the short-period comet P/Encke is associated with the meteor stream S. Taurids, and P/Giacobini–Zinner is associated with the October Draconids.

The Geminids stream is connected with the Apollo object 3200 Phaethon. Since Apollo objects are basically defined as Small bodies in Earth-crossing orbits (and Amor objects are bodies in Mars-crossing orbits), suggestions have been made that 3200 Phaethon may in fact be a defunct cometary nucleus. There are several possible clues supporting the hypothesis that short-period comets could be a source for the Apollo–Amor objects. First, several of these Apollo–Amor objects are in relatively high inclinations ($>30^\circ$), which are very unusual for the asteroidal population. Second, three Apollo–Amor objects have aphelia beyond Jupiter's orbit. Finally, the Apollo object 2201 Oljato has a surface UV reflectance very different from that of the asteroids. Short-period comets like P/Arend–Regaux and P/Neujmin II are almost inactive; therefore, they might have reached the turning point of becoming Apollo–Amor objects. Recent dynamical calculations show that, in addition to the injection from the main-belt asteroidal population via chaotic motion near the 3:1 Jovian commensurability and the ν_6 secular resonance in the inner boundary of the asteroid belt, a significant fraction of the Apollo–Amor objects (≈2400 in total) could indeed come from short-period comets.

III. GENERAL MORPHOLOGY

Despite its brilliance in the night sky, a comet has a solid nucleus of only a few kilometers in diameter. The brightness comes from the dust, gas, and ions emitted from the nucleus. For example, during its 1986 passage near the earth's orbit, comet Halley lost on the order of 3.1×10^4 kg

of volatile ice per second and about an equal amount in small nonvolatile dust particles. The expansion of the neutral gas (mostly water and carbon monoxide) from the central nucleus would permit the formation of a large coma visible in optical, ultraviolet, and infrared wave-lengths. The optical emission in a cometary coma is largely from the excitation of the minor constituents, such as CN and C_2, by the solar radiation. These radicals are the daughter products from photodissociation of parent molecules such as HCN, C_2H_4, C_2H_6, and other more complicated molecules. The dirty snowball model of the cometary nucleus first proposed by Whipple (1950) describes the nucleus as a mixture of frozen ice and nonvolatile grains. The main components of the volatile ice are H_2O, CO_2, CO, H_2CO, CH_4, and NH_3, followed by minor species such as HCN, C_2H_4, C_2H_6, CS_2, and others.

A. Surface Sublimation

An icy nucleus will begin to evaporate significantly at perihelion approach to the sun once its surface temperature exceeds a certain value. Since the gas pressure in the cometary coma is much smaller than the critical pressure at the triple point, direct sublimation from the solid phase into vapor will occur. Under steady-state conditions, the energy equation in its simplest form can be written as

$$F_\odot e^{-\tau}(1-A_v)r^{-2}\cos\theta\cos\phi = \varepsilon_{IR}\sigma T^4 + \frac{L(T)}{N_A}\dot{Z} - K(T)\nabla T|_s \qquad (14)$$

where $F_\odot$ is the solar flux at 1 AU, τ is the optical depth of the dust coma, A_v is the surface albedo, $_r$ is the heliocentric distance in AU, ϕ and θ are the local hour angle and latitude, ε_{IR} is the infrared emissivity, σ is the Stefan–Boltzmann constant, T is the surface temperature, $L(T)$ is the latent heat of sublimation, N_A is the Avogadro number, $\dot{Z}$ (in units of molecules $cm^{-2}s^{-1}$ str^{-1}) is the gas production rate, and $K(T)$ is the thermal conductivity at the surface.

Another important equation is the Clapeyron–Clausius equation. For water we have

$$\log p(\text{mm Hg}) = \frac{-2445.5646}{T} + 8.2312\log T - 0.01677006\,T + 1.20514 \times 10^{-5}T^2 - 6.757169 \qquad (15)$$

with the equilibrium vapor pressure p in units of millimeters of mercury. And for CO_2 we have

$$\log p(\text{mm Hg}) = \frac{-1367.3}{T} + 9.9082 \qquad \text{for } T > 138\text{ K} \qquad (16)$$

and

$$\log p(\text{mm Hg}) = -\frac{1275.6}{T} + 0.00683T + 8.307 \qquad \text{for } T < 138\text{ K} \qquad (17)$$

The latent heat of vaporization of water ice is

$$L(T) = 12420 - 4.8T \qquad (18)$$

with $L(T)$ in cal $mole^{-1}$. For CO_2 ice we have

$$L(T) = 12160 + 0.5\,T - 0.033\,T^2 \qquad (19)$$

Finally, the sublimation rate is related to the equilibrium vapor pressure by the following equation:

$$m\dot{Z}(T) = p(T)\left(\frac{m}{2\pi KT}\right)^{1/2} \qquad (20)$$

Figure 8 shows the variations of $\dot{Z}(T)$ as a function of the solar distance r with $\langle\cos\theta\cos\phi\rangle = \frac{1}{4}$, $A_v = 0.02$, and $\varepsilon_{IR} = 0.4$. For CO_2 ice, a significant level of surface sublimation starts at $r \approx 10$ AU; strong outgassing would occur at $r \approx 2$–3 AU for H_2O ice.

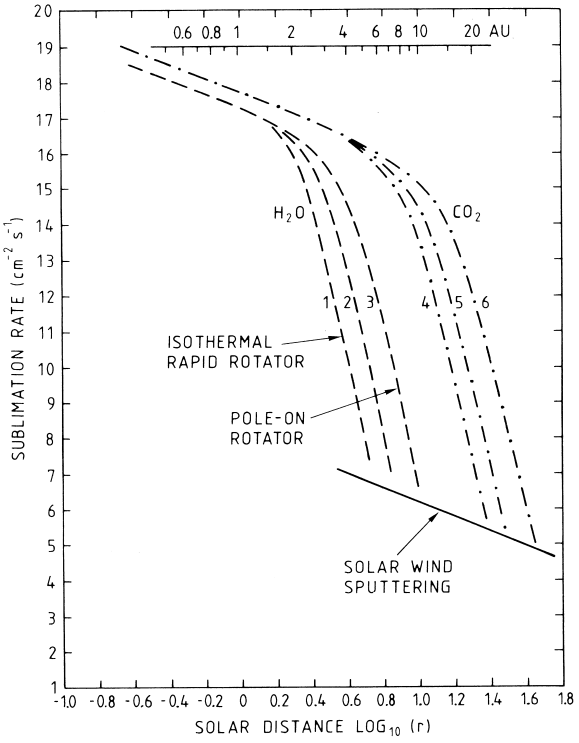


FIGURE 8 The variations of the sublimation rates of H_2O and CO_2 ices as a function of the solar distance. The surface albedo A_v is taken to be 0.02 and the infrared emissivity ε_{IR} is assumed to be 0.4. Two extreme cases of the orientation of the spin axis (isothermal rapid rotator and pole-on rotator) are shown. The surface sputtering rate via solar-wind protons is also given.

B. Gas Comas and Ion Tails

As the parent molecules (H_2O, CO, CO_2, CH_4, H_2CO, etc.) move away from the nucleus, they will be either dissociated or ionized by solar ultraviolet radiation. The photodissociation life-time for water molecules is about 10^5 s at 1 AU solar distance. With an expansion speed (V_n) of around 1 km s^{-1}, most of the water molecules will be dissociated into hydrogen atoms and hydroxide (OH) at a cometocentric distance of a few 10^5 km. The OH will also be dissociated subsequently. A similar process occurs for other parent molecules. As a result, at a cometocentric distance of a few 10^6 km, the extended neutral coma will be made up only of atomic species (H, C, O, N, S, . . .). In Lyman alpha emission at 1216 Å, the atomic hydrogen cloud of a comet can become the most prominent structure in the solar system (Fig. 9). The ionization of the cometary neutrals by solar radiation and charge exchange with solar wind protons will lead to the formation of an ion tail flowing in the antisolar direction. In optical wavelengths, CO^+ and H_2O^+ are the most visible among the ions. The general features of the interaction of the cometary coma with the solar radiation and interplanetary plasma are sketched in Fig. 10.

C. Dust Coma and Mantle

The gaseous drag from the expanding outflow is able to carry dust particles away from the nuclear surface. There is, however, an upper limit on the particle size to be ejected in this manner. The critical particle size is reached when the gravitational attraction of the nucleus outweighs the gaseous drag. The stronger the surface gas sublimation rate, the larger will be the critical size (see Section VI). For comet Halley, its out-gassing rate at 1 AU solar distance would enable the emission of solid particles up to a

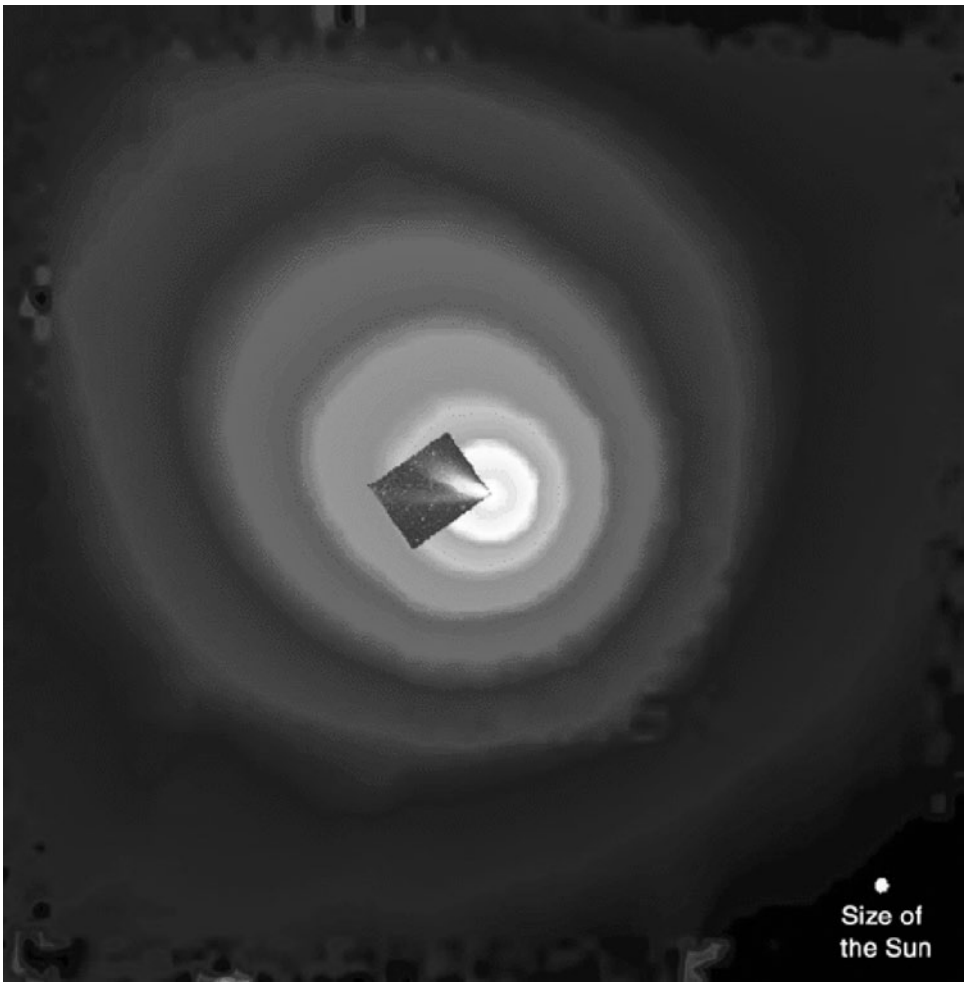


FIGURE 9 The SWAN experiment on the SOHO spacecraft observed a huge atomic hydrogen cloud surrounding Comet Hale–Bopp. The hydrogen cloud is 70 times the size of the Sun. Photo origin: ESA/NASA.

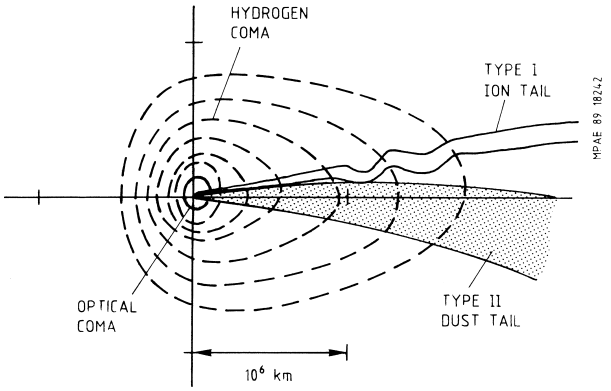


FIGURE 10 A schematic view of the interaction of the cometary coma with the solar wind. The Type I ion tail points along the radial direction; the Type II dust tail is usually more diffuse and curved.

diameter of a few tens of centimeters. A halo of centimeter-size particles in its vicinity was indeed detected by radar observations.

In the above scenario, solid particles with diameters larger than the critical value will remain on the nuclear surface, forming a sort of dust mantle. When a comet moves along a Keplerian orbit, its outgassing rate will change as a function of solar distance (see Fig. 8). The formation of a dust mantle is therefore a time-dependent process. Also, irregular structures of the nuclear surface might lead to rather patchy coverage by such a dust layer. Since a dust mantle a few centimeters thick would be sufficient to insulate the heat flow such that the effective gas sublimation rate is reduced by a factor of 100, the surface area covered by a thin dust mantle would become much less active.

As a comet goes through many perihelion returns near the sun, its gas sublimation rate will begin to subside. One consequence is that the critical size of dust grain ejection will become smaller and smaller. As a result, more and more dust particles will be accumulated on the nucleus surface; that is, the dust mantle will become thicker and more widespread as the comet ages. It is perhaps for this reason that "old" comets (such as comet Encke and comet Halley) are generally observed to have anisotropic gas emission from a few localized active regions. The rest of the surface, presumably, is covered by a thickening dust layer. On the other hand, comet Wilson (1986), which is a new comet in the Oort sense, did not display any anisotropic outgassing effect. This may be simply because its "fresh" surface still lacks an insulating dust mantle of significant size.

D. Dust Tails

The elongated dust tails are formed of micron-size dust particles being accelerated away by the solar radiation pressure force. The ratio of the radial component of the solar radiation force (F_r) to the solar gravitation (F_g) can be expressed as

$$1-\mu=\frac{F_r}{|F_g|}=\frac{1.2\times10^4}{\rho d}Q_{pr} \qquad (21)$$

where Q_{pr} is the Mie efficiency factor for radiation pressure, ρ is the density of the dust grain, and d is its size. The value of Q_{pr} depends on both the size of the grain and on its optical properties. For a dielectric particle with $d\approx1\ \mu$m, $Q_{pr}<0.5$, while $Q_{pr}>1$ for conductors. In general, the $1\ \mu$ value is largest in the size range between 0.1 and 1 μm. Figure 11 shows the efficiency factor for radiation pressure, Q_{pr}, integrated over the solar spectrum for different grain properties as a function of grain radius. It can be seen that Q_{pr}, for dielectric grains (e.g., olivine), decreases sharply for $a<0.3\ \mu$m. Submicron grains are therefore relatively invisible.

In the cometocentric frame, the instantaneous locus in the dust tail formed of solid particles with the same value of $1-\mu$ is called syndyname. Another type of locus of interest is the so-called synchrone formed of particles of different values of $1-\mu$ but emitted from the nucleus

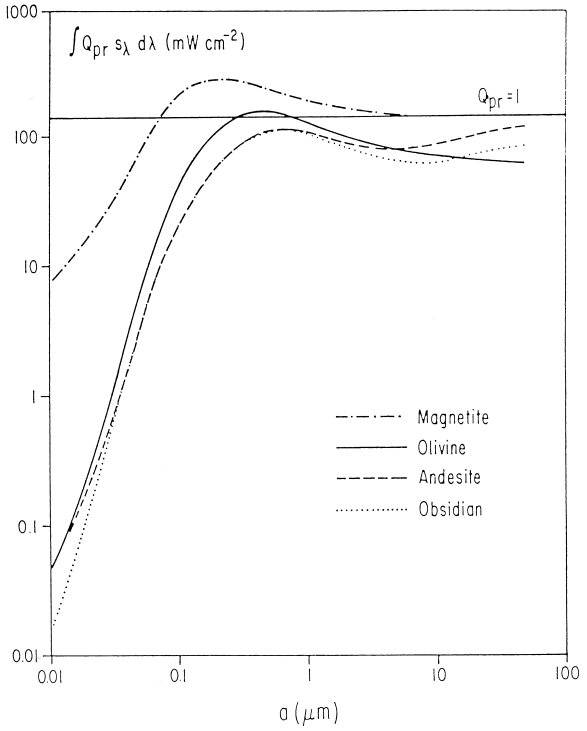


FIGURE 11 The Mie efficiency factor for radiation pressure Q_{pr} integrated over the solar spectrum for different grain materials as a function of grain radius. [From Hellmich, R. and Schwehm, G. (1982), *in* "Proc. Intern Conf. on Cometary Exploration" Gombosi, T., ed., Vol. I, p. 175.]

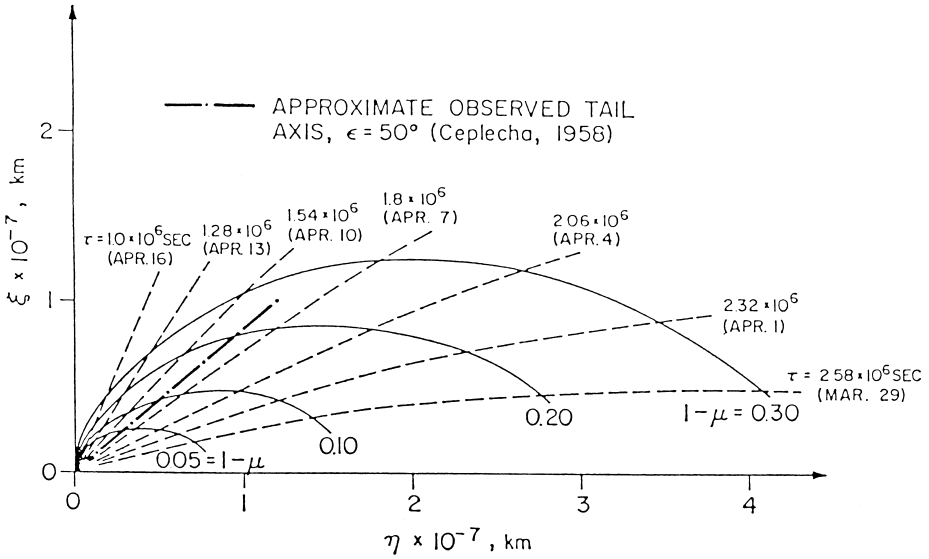


FIGURE 12 A description of the formation of the syndynames (——) and synchrones (– –) in the dust tail of comet Arend–Roland (1957 III) on April 27.8. 1957. [According to Finson, M. L., and Probstein, R. F. (1968). *Astrophys. J.* **154,** 353.]

at the same time. A classical description by Finson and Probstein (1968) of these two types of loci in the dust tail structure of comet Arend–Roland 1957 is given in Fig. 12. The brightness distribution and orientation of a cometary dust tail can be analyzed by taking into consideration (a) the size distribution of the dust particles, (b) the total production rate of dust particles as a function of time, and (c) the initial velocity of the emitted dust particles as a function of time and size. In this way, a synthetic dust tail model can be constructed and compared with observations. From parameter studies, a best fit could be obtained with a certain dust particle size distribution. Such a treatment has been used to derive useful information on the size distributions of dust particles in the dust tails of many comets (see Section VI).

IV. COMETARY SPECTRA

Figure 13 is a composite spectrum of comet Tuttle 1990 XIII between the ultraviolet and infrared wavelength range. The ultraviolet data up to 0.28 μm came from the International Ultraviolet Explorer (IUE) and the rest from ground-based observations. The figure shows emission features from various atomic and molecular species (HI, OI, CI, SI, CS, OH, NH, NH_2, CH, CN, C_2, C_3). Before spacecraft missions to comet Giacobini–Zinner in 1985 and to comet Halley in 1986, such spectra were the only means to infer the composition of the gas comas. The information on the spatial distributions of the brightness of individual species can be used to infer the corresponding abundances and life-times against photodissociation.

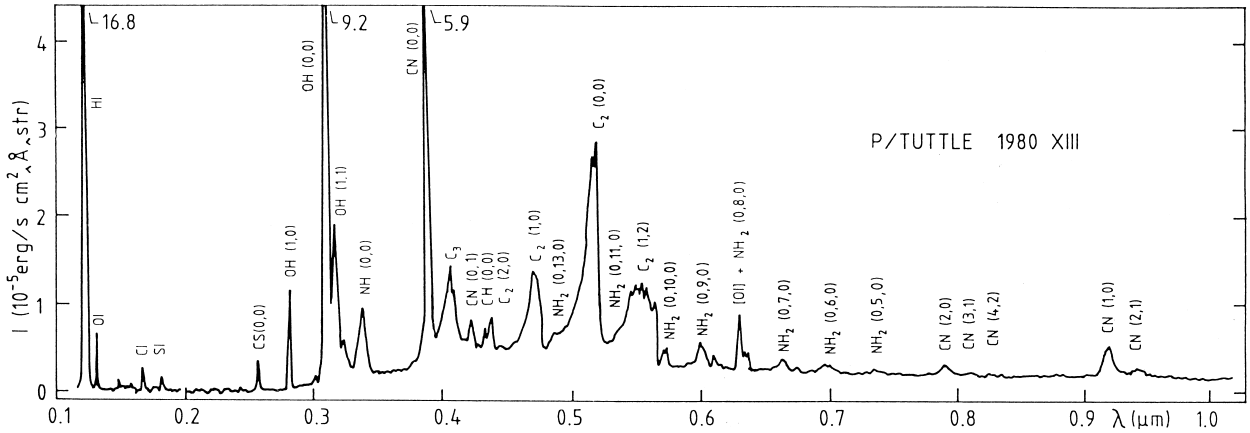


FIGURE 13 A composite spectrum of the coma emissions from comet Tuttle 1980 XIII. [From Larson, S., and Johnson, J. R. (1985).]

For example, the Lyman alpha emission of HI and the very strong emission of OH in many comets have shown that water is the main constituent of cometary ice. The CO emission at 1510 Å in several comets suggests that carbon monoxide can be a very important component as well, but its abundance relative to H_2O varies from comet to comet. Rocketborne UV observations and neutral mass spectrometer measurements at comet Halley have further shown that nearly half of the CO molecules were released in a distributed region (of radius ≈25,000 km) surrounding the nucleus. Dust grains and formaldehyde polymers have been suggested as the possible parents.

A fraction of the CN radicals could be released from small dust grains. Ground-based optical observations of comet Halley detected jet-like features of CN emission in the coma. Sublimation from organic dust grains and emission from local inhomogeneities on the nucleus surface are two possible explanations. Despite their brightness, the CN, C_2, and C_3 radicals are all minor constituents with relative abundances on the order of 0.1%.

Table I lists most of the species identified in cometary spectra from observations at different wavelengths. Spacecraft measurements at comet Halley showed that the chemical composition of the gas coma is much more complex than indicated here. Remote-sensing observations, however, will continue to be a powerful tool in gathering basic chemical information about a large sample of comets. The radio observations of comet Hyakutake and comet Hale–Bopp have shown the presence of hydrocarbon molecules such as C_2H_2, C_2H_6, and HNC.

Continuum emission from dust particles could be very strong for some comets. There is a trend indicating that the dust production rate is proportional to the production rate of the CN radicals. This effect, if confirmed by statistical correlations of more comets, would indicate that a part of the CN radicals could indeed originate from the dust particles in the coma. The dust comas of several comets were observed to be highly anisotropic. As illustrated in Fig. 14, a system of dust jets could be seen in the coma of comet Halley near the time of the Giotto encounter. The CN– (and C_2–) jets detected at the same time, however, did not align well with these dust jets. It is possible that "invisible" submicron dust grains of organic composition could release these gaseous molecules or their parent molecules via sublimation.

TABLE I Species in Cometary Optical, UV, IR and Radio Spectra

Location	Species
Coma (gas)	H, C, C_2, $^{12}C^{13}C$, C_3, O, OH,
	H_2O, S, S_2
	CH, CH_4, CN, CO, CO_2, CS, C_2H_2, C_2H_6,
	HCN, HNC, HCO, H_2CO, CH_3CN, NH, NH_2, OCS
	Na, K, Ca, Cr, Mn, Fe, CO, Ni, Cu, V, Ti(?)
Tails (ions)	C^+, CH^+, CN^+, CO^+, CO_2^+, N_2^+, OH^+, H_2O^+,
	H_3O^+, Ca^+, H_2S^+

Infrared observations of comet Halley from the ground discovered an emission feature near 3.4 μm. The same feature was seen in the spectra taken by the infrared spectrometer experiment (IKS) on Vega 1 (see Fig. 15). Several mechanisms might be responsible for such emission, such as heating of very small organic grains of large polycyclic aromatic hydrocarbon molecules (PAH) by UV photons, thermal emission by small cometary grains with organic mantles, and resonant fluorescence of small gaseous molecules. Emissions from the OCS, CO, CO_2, H_2CO, and H_2O molecules were also detected in the IKS spectra (see Section VII).

Infrared emissions from water molecules were observed from the Kuiper Airborne Observatory (KAO) with very high spectral resolution. In addition to deducing water production rates at different times, the KAO observations are important in finding strong anisotropic outflow of the water vapor from the central nucleus and in determining the ortho–para ratio (OPR) of comet Halley. The latter can be used as a probe to the temperature at which the water was last equilibrated, that is, when it condensed into water ice. Infrared observations of water molecules of this type therefore hold the promise of mapping the temperature variations of the solar nebula during the condensation of the icy planetesimals.

V. COMETARY NUCLEI

A. Albedo

Before the spacecraft flyby observations of comet Halley, ground-based broadband photometric observations of several comets had shown that the V-J and J-K colors of their nuclei are similar to those of the Trojan D-type asteroids. This in turn suggests that the cometary nuclei must be very dark, with the albedo $p_v < 0.1$. At the same time, the mean albedo of cometary dust grains measured at 1.25 μm for a number of comets was found to be on the order of 0.02–0.04. The close-up views from the Vega and Giotto spacecraft have shown that the surface albedo ($p_v \approx 0.02$–0.05) of comet Halley is indeed among the darkest of solar system objects. One possible reason for such a dark color might be related to early irradiation effects on the hydrocarbon material mixed in the cometary ice.

B. Size

When coma activity is negligible, the sunlight reflected by the bare nucleus accounts for all of the optical emission

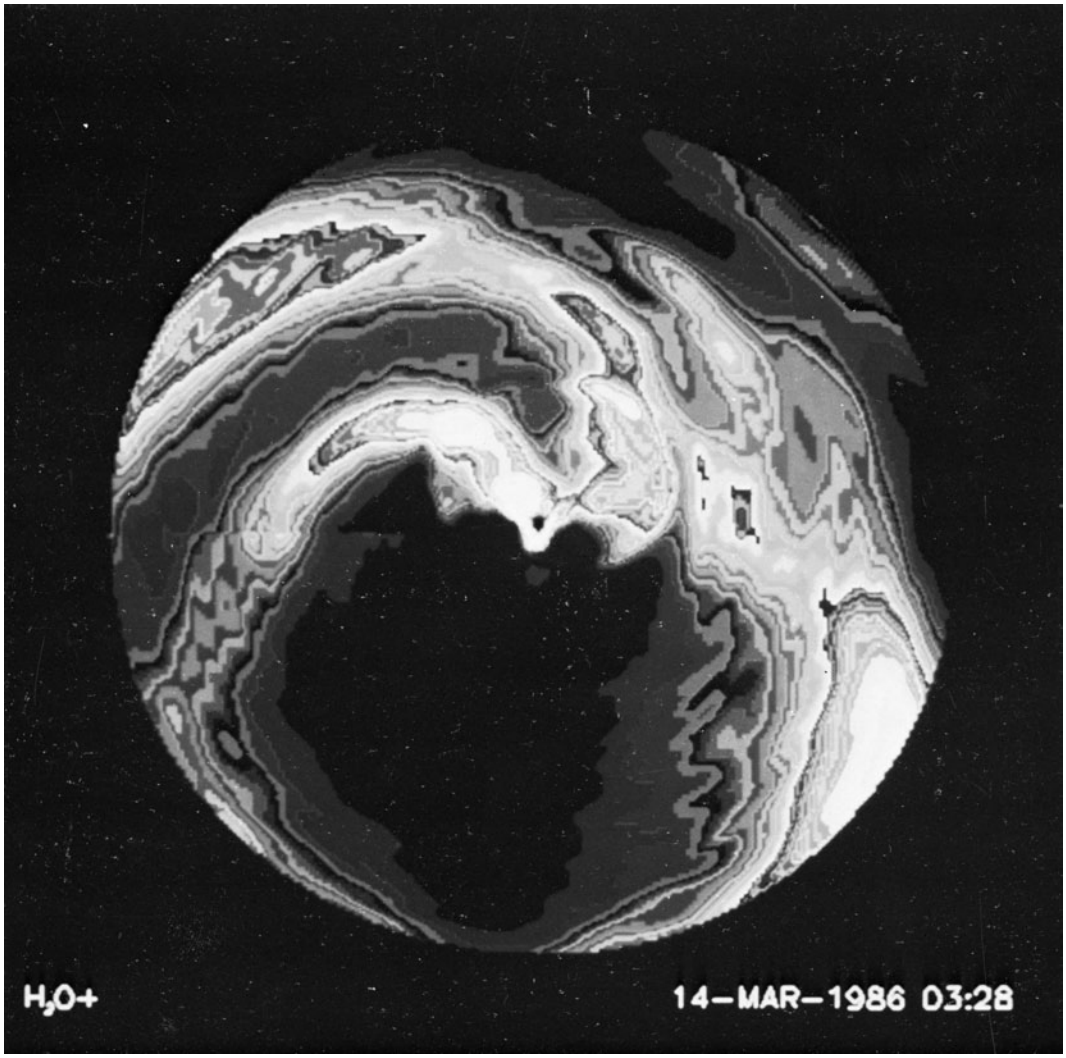


FIGURE 14 A false color image-enhanced picture of the dust jet system in the coma of comet Halley taken on March 14, 1986, at the South Africa Astronomical Observatory by C. Cosmovici and P. Mack. Image processing by G. Schwarz, DLR. Wessling, FRG.

observed. Under this circumstance, the radius (a) of the cometary nucleus can be given as

$$a^2 = [p_v \phi(\theta)]^{-1} 10^{0.4(M_\odot - H + 5\log(r_\Delta))} \qquad (22)$$

where p_v is the geometrical albedo, $\phi(\theta)$ is the phase function, $M_\odot$. is the absolute magnitude of the sun, H is the apparent nuclear magnitude, r is the solar distance of the comet, and Δ is the geocentric distance. Figure 16 shows the distributions of the nucleus radii for a number of periodic comets and near-parabolic comets. In the present compilation of the data from Roemer (1966), p_v is assumed to be 0.02. As a rule, most comets have a radius of a few kilometers, except for some giant comets such as comet Humason 1962 VIII and comet Wirtanen 1957 VI, which showed signs of being new comets entering the inner solar system for the first time. Near parabolic comets appear to be larger in general than the short-period ones if they all have the same albedo of 0.02.

Radar observations using large antennas have been successful in making independent determination of the radii of a few comets. For example, the periodic comet P/Encke was found to have a radius of $1.5^{+2.3}_{-1.0}$ km from the Arecibo radar observations. The observations of comet IRAS–Araki–Alcock 1983d showed that the radius ranges from 2.5 km for a solid-ice surface to 8 km for a surface of loosely packed snow. The nucleus surface seems to be very rough or porous on scales of a few meters or more.

According to the imaging experiments on the Giotto and Vega spacecraft, the physical size of comet Halley, which

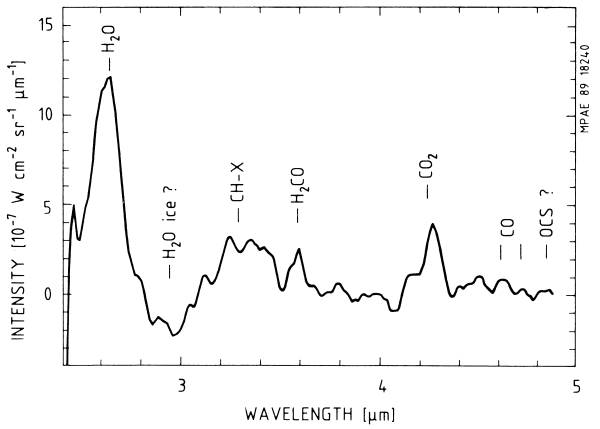


FIGURE 15 A spectrum of comet Halley between 2.5 and 5 μm, obtained from the average of five individual spectra taken at a cometocentric distance of 42,000 km. [From Combes, M, *et al.* (1989). *Icarus*, **76,** 404.]

has a 2:1 ellipsoidal shape in the first approximation, can be characterized as follows: (a) maximum length $=$ 16 ± 1 km, (b) intermediate length $= 8.2 \pm 0.8$ km, and (c) minimum length $= 7.5 \pm 0.8$ km. The total volume is thus of the order of 500 km^3.

C. Mass and Density

On the basis of detailed calculations of the nongravitational effects on comet Halley's orbital motion, the mass of the nucleus has been estimated to be between 5×10^{16} g and 1.3×10^{17} g. This would mean an average density of 0.08–0.24 g cm^{-3} for the nucleus. The uncertainties involved in the computations do allow a higher density, however. If comet Halley has already experienced 3000 orbital revolutions in similar orbits in the inner solar system, the original mass before its commencement of active mass loss could be estimated to be about 5 to 6 times its present mass.

D. Surface Activity

As indicated by the dust jet features on the sunward side of the nucleus, the surface outgassing process of comet Halley was highly an isotropic (see Fig. 17). The dust jets were confined within cones of about 120° for Vega and 70° for Giotto. The active region covered an area of no more than 10% of the total surface area. Interestingly, in spite of the strong anisotropic emission, a significant dust background was seen on the nightside with a ratio of 3:1 for the sunward–antisunward brightness variation. It is not yet clear what mechanism was responsible for this nightside dust coma. Nonradial transport of the dust particles from the jet source to the nightside would require very special lateral flow condition in the inner coma.

E. Surface Features

Radar observations have indicated that the nuclear surfaces of comets could be rough on scales of a few meters or more. The Giotto HMC showed that comet Halley's

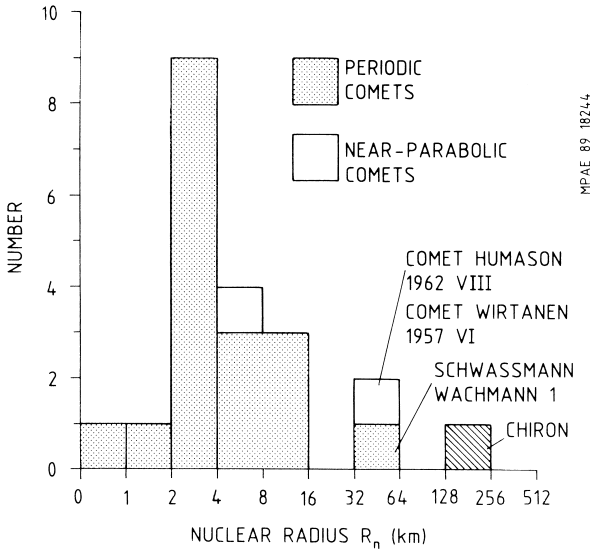


FIGURE 16 Frequency distributions of the radii of periodic comets and new comets. All comets are assumed to have a surface albedo of 0.02 [Data from Romer, E. (1966). *In* "Nature et Origine des cometes." *Pub. Institut d'Astrophysique*, Liege, Belgium, p. 15.]

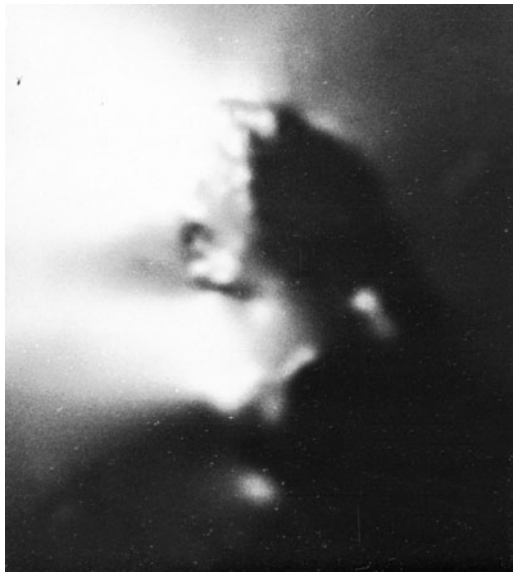


FIGURE 17 An image of the nucleus of comet Halley taken by the Halley multicolor camera onboard Giotto. [Photo courtesy of the Max-Planck-Institut für Aeronomie.]

surface morphology has a roughness of a few hundred meters. Furthermore, a number of "craters" can be identified. One of them has a depth of 200 m and a diameter of about 2 km. Some other topological structures such as "mountains" and "hills" could be delineated as well.

F. Internal Structure

Even though the subsurface structure of comet Halley's nucleus is hidden from view, there are several clues to its possible physical nature. First, the very low bulk density (<0.3 g cm^{-3}) suggests that the internal structure must be very fluffy at different scales. One concept is that, as a result of random accretion of icy planetesimals in the solar nebula, a fractal model might be a good approximation. In this way, the internal configuration of a cometary nucleus of a few kilometers in size could be very similar to the irregular share of porous interplanetary dust grains (see Fig. 18). Besides the nonvolatile building blocks, the nucleus would also consist of icy aggregates sintering or glueing the matrix together. In other words, a comet nucleus could be visualized as an assemblage of many cometesimals of different sizes (1–100 m, say).

During the long-term storage in the Oort cloud for a period of 4.5×10^9 years, comets would be constantly bombarded by galactic cosmic rays. Subsurface material down to a depth of 1–10 m would be processed by energy deposition of such energetic charged particles. Such surface irradiation might permit the buildup of a layer of very volatile cover on the surface of comets in the Oort cloud. This effect could be related to the fact that new comets such as comet Kohoutek 1973 XII, during their first inbound entry into the solar system, tend to brighten up anomalously at large heliocentric distances (>4 AU) but fizzle away at perihelion and outbound passage.

For an evolved comet, its subsurface structure may be approximated by a number of layers with different chemical compositions. The innermost core, containing pristine ice without suffering from any substantial sublimation loss, is covered by a mantle of core material with the most volatile component (i.e., CO or CO_2) already purged. In this mantle, water ice and the refractory dust grains are still intact. At the nuclear surface, a crust of dust particles with much-reduced water-ice content could exist. The exchange of mass and hence the chemical differentiation effect as a result of time variation of the temperature profile

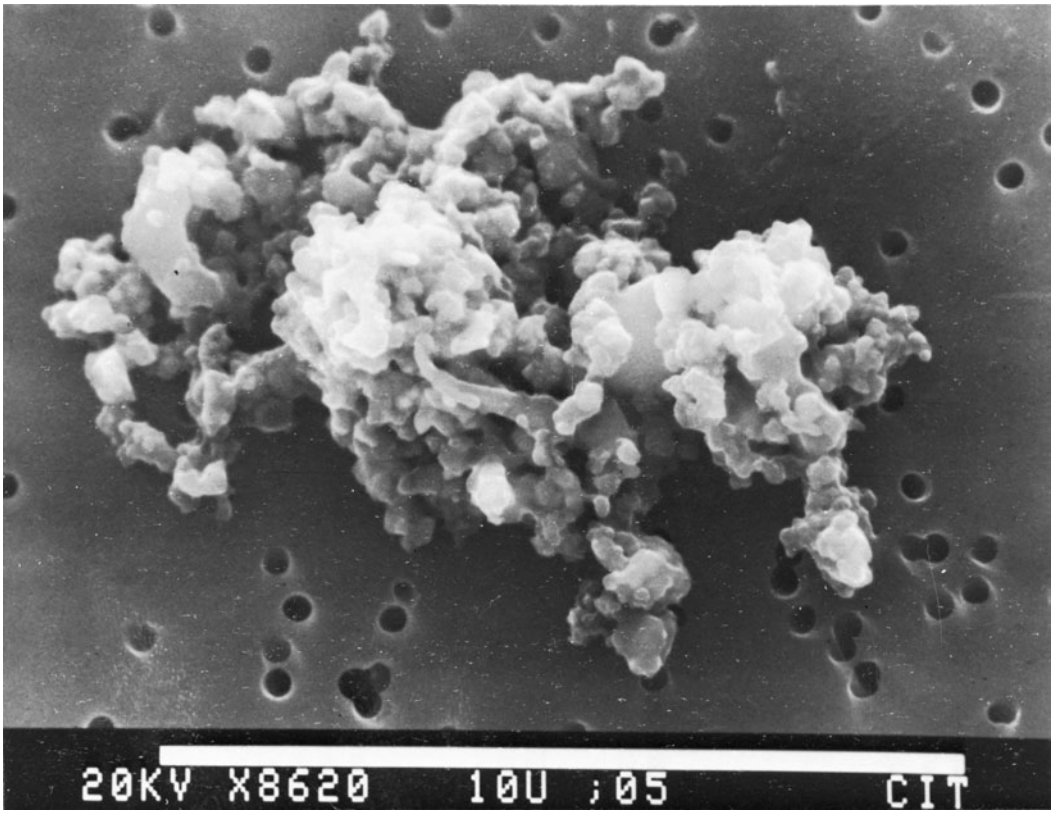


FIGURE 18 An electron microscopic picture of a highly porous chondritic interplanetary dust particle (U2-18A3B). [From Fraundorf, P., Brownlee, D. E., and Walker, R. M. (1982). *In* "Comets" (Wilkening, L. L., ed.), Univ. of Arizona Press, p. 383.]

could be determined by not just the orbital motion of the comet but also its rotational movement (i.e., day-and-night variations).

The sudden flareups of the coma activities of some comets have been suggested to be the result of phase transition of the cometary ice. If the interior of a cometary nucleus is made up of amorphous ice from low-temperature condensation in the solar nebula, a transformation into cubic ice will take place at about 153 K. The latent heat release would be an additional heat source to the cometary nucleus. Furthermore, the heat conduction coefficient of crystalline ice is about a factor of 10 higher than that for amorphous ice at 153 K; the heat balance of the nucleus consequently would be strongly influenced by the physical nature of the ice.

G. Rotation

From a number of observations such as the time variations of the light curves and the production of CN-shells in the coma of comet Halley, two periodicities of the temporal changes were found. A period of 2.2 days is superimposed on a longer period of 7.3 days. This means the nucleus of comet Halley could be spinning as well as precessing. If the ellipsoidal shape of the nucleus is approximated as a symmetric top with a major axis c and minor axes $(a = b) < c$, the rotational motion of the nucleus could be in one of two possible states, namely, one with the ratio of the spin period (P1) to the precession period (P2) being characterized by P1/P2 = 3.37 and the other with P1/P2 = 0.30. For example, arguments have been presented to support the scenario that the nucleus is rotating with a spin period of 7.4 days and a nutation period of 2.2 days. The cone angle between these two rotational vectors is about 76°. Other possible configurations have been suggested as well. Because of existing uncertainties in the ground-based and spacecraft data, no definite answer can be derived yet.

VI. DUST

A. Composition

The compositional variation of cometary dust grains had been inferred from collected samples of the interplanetary dust particles (see Fig. 18). A porous aggregate of CI carbonaceous chondrite composition, containing elements from C to Fe in solar abundance, is generally considered to be typical of cometary grains. While the dust-particle experiments on Vega and Giotto have not provided complete information on the mineralogical makeup of different kinds of grains in the coma of comet Halley, the preliminary results show that a large portion of the solid particles

TABLE II Average Atomic Abundances of the Elements in Halley's Dust Grains and in the Whole Comet, Dust and Ice

	Halley			
Element	**Dust**	**Dust and ice**	**Solar system**	**CI-chondrites**
H	2025	3430	2,600,000	492
C	814	≡940	940	70.5
N	42	76	291	5.6
O	890	1627	2,216	712
Na	10	10	5.34	5.34
Mg	≡100	100	≡100	≡100
Al	6.8	6.8	7.91	7.91
Si	185	185	93.1	93.1
S	72	72	46.9	47.9
K	0.2	0.2	0.35	0.35
Ca	6.3	6.3	5.69	5.69
Ti	0.4	0.4	0.223	0.223
Cr	0.9	0.9	1.26	1.26
Mn	0.5	0.5	0.89	0.89
Fe	52	52	83.8	83.8
Co	0.3	0.3	0.21	0.21
Ni	4.1	4.1	4.59	4.59

[From Jessberger, E. K., and Kissel, J., (1990). *In* "Comets in the Post-Halley Era" Newburn, R., Neugebauer, M., and Rahe, J., (eds.).]

indeed have compositions similar to that of the CI carbonaceous chondrites (see Table II). Furthermore, several types of dust grains can be identified. The first one is characterized by the CI chondritic composition with enrichment in the elements carbon, magnesium, and nitrogen. A small population of particles of pure silicate composition (Si, O, Mg, and Fe) constitute the second group. Organic (CHON) grains composed solely of light elements, H, C, N, and O constitute the third group. The CHON particles might be partly of interstellar grain origin with their outer mantles covered by organic polymerized material from cosmic ray and/or UV photon irradiation.

According to ground-based infrared observations, the composition of dust particles in the coma of comet Halley should be about 56% olivine, 36% pyroxene, and 8% layer lattice silicates. These estimates are not in disagreement with the *in-situ* measurements at comet Halley. Thus the PIA and PUMA dust experiments indicated that the dust of comet Halley is composed of two main components: a refractory organic phase (CHON) and a siliceous, Mg-rich phase. The CHON organic material further serves as coating of the silicate cores (see Fig. 19). The CHON-dominated grains have an average density of about 1 g cm^{-3} and the silicate-dominated grains of 2.5 g cm^{-3}. Also noteworthy is the finding that the isotopic ratio of $^{12}C/^{13}C$ in different grains varies from about 1 to 5000. Ground-based spectroscopic observations of

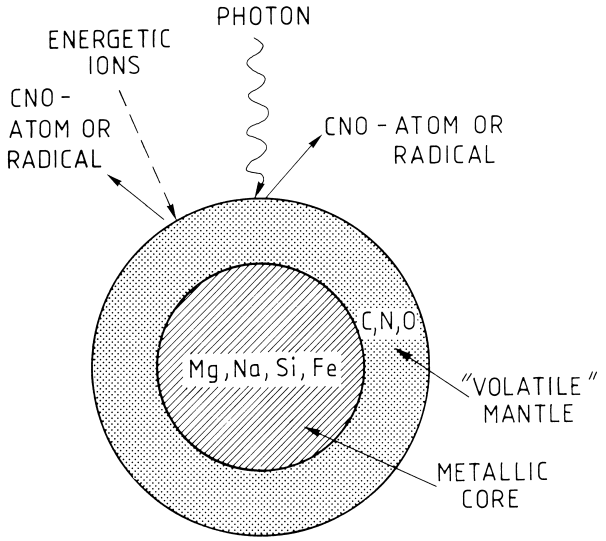


FIGURE 19 A schematic view of a possible inhomogeneous structure of cometary dust grains.

comet Halley showed that $^{12}C/^{13}C = 65$ in CN. There is therefore evidence that the isotopic ratio of carbon in comet Halley (and perhaps other comets also) is below the nominal terrestrial value of $^{12}C/^{13}C = 90$. Such isotopic anomaly might be the result of nucleosynthetic environment of the solar nebula or ion–molecule reactions in interstellar clouds.

B. Size Distribution

Before the *in-situ* measurements at comet Halley, a standard dust-size distribution had been constructed by Devine *et al.* (1986). This was given as

$$n(a)\,da \begin{cases} C_0[1 - a_0/a]^M \left[\dfrac{a_0}{a}\right]^N da & a \geq a_0 \\ O & a < a_0 \end{cases} \qquad (23)$$

where $M = 12$, $N = 4$, C_0 is a normalization constant, a_0 ($= 0.1\ \mu$m) is the radius of the smallest dust grain, and a is the grain radius. In Fig. 20 this model distribution is compared to the size distribution obtained by the Vega and Giotto measurements. It is clear that while the Devine *et al.* size distribution predicted the absence of submicron grains, the spacecraft data indicated a large number of such small particles. The possible existence of a very high flux of 10^{-17} g particles (i.e., $a = 0.01\ \mu$m) was also reported. This component of very small grains might be related to the polycyclic aromatic hydrogen carbons (PAHs) in interstellar space.

The total mass production rate of the dust particles detected by the Giotto DIDSY dust particle experiment, if extrapolated to $m_{max} = 1$ g, amounts to a gas-to-dust ratio of about 7:1. But if the observed mass distribution is extrapolated to a larger mass, say $m_{max} = 1$ kg, the corresponding gas-to-dust ratio would be about 1:1. It is interesting to note that radar observations of comet Halley in 1985 detected a weak echo that might have come from a halo of large (>2-cm radius) grains ejected from the cometary nucleus. Figure 21 shows the Doppler spectra of the full radar echo (nucleus and coma components) for comets (a) IRAS–Araki–Alcock, and (b) Hyakutake. The data show the presence of extended structures of large dust grains with sizes of a few centimeters surrounding the nuclei of comet IAA and Hyakutake.

C. Dynamics

After emission from the nuclear surface, dust particles will be accelerated by the drag force of the expanding gas. At the same time, the gravitational attraction of the nucleus itself tends to decelerate the motion of the dust particles is relatively small, such that they do not affect the expansion of the gas, we can write the equation of momentum transfer as

$$m\frac{du}{dt} = \pi a^2 \rho_{gas} C_D (u_{gas} - u)^2 - \frac{GmM_N}{r_c^2} \qquad (24)$$

where $m = \frac{4}{3}\pi a^3 \delta$ is the mass of the grain with radius a and density δ, u_{gas} is the flow velocity of the gas, ρ_{gas} is the gas density, M_N is the mass of the cometary nucleus, G is the universal constant of gravitation, C_D is the drag coefficient, and r_c is the cometocentric distance. From Eq. (24) it can be seen immediately that there exists an

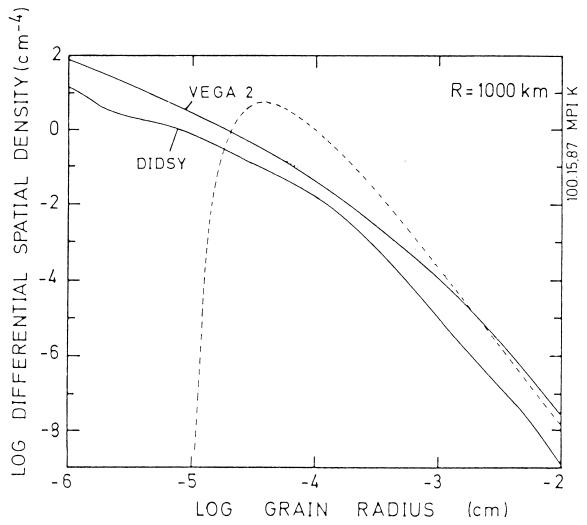


FIGURE 20 A comparison of the observed differential spatial density of cometary dust particles as a function of radius of the grains with theoretical curve determined by Devine *et al.* (1986). The experimental curve was deduced from the cumulative densities measured by the DIDSY impact detectors on Giotto. [From Lamy, Ph., *et al.* (1987). *Astron. Astrophys.* **187,** 767.]

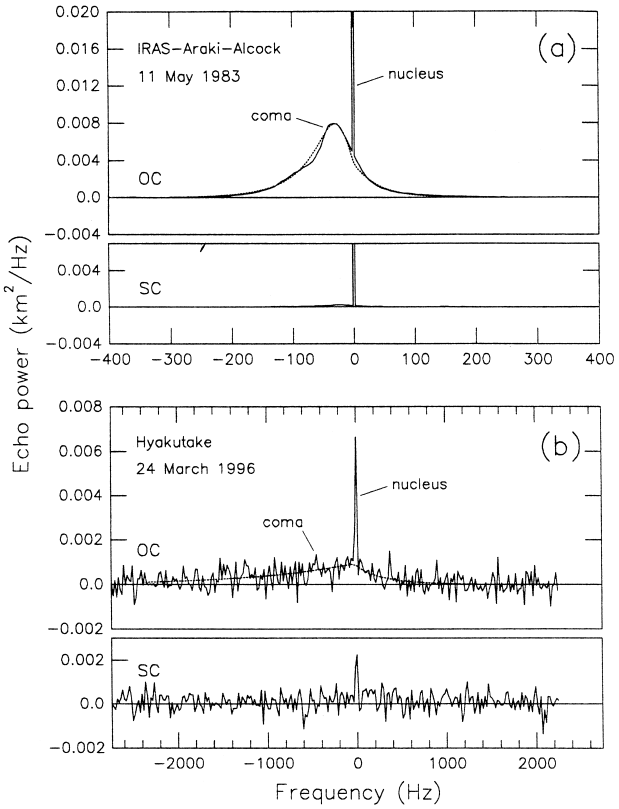


FIGURE 21 The Doppler radar spectra of comets (a) IRAS-Araki-Alcock, and (b) Hyakutake. [From J. K. Harmon et al. (1999). *Planetary and Space Science* **47,** 1409.]

upper limit of the grain radius (a_c) above which $du/dt < 0$, and hence no cometary grains of the corresponding sizes could be lifted off the nuclear surface. This critical radius can be approximated as

$$a_c \approx 3\frac{C_D Q_m u_0}{4\pi\delta G M_N} \tag{25}$$

where u_0 is the initial velocity of the gas at the nuclear surface and Q_m is the total gas mass production rate. Thus, in the case of comet Halley at 1 AU solar distance, with $Q_m \approx 3.1 \times 10^7$ g s^{-1} and $u_0 \approx 0.3$ km s^{-1}, $C_D \approx 0.5$, $M_n \approx 10^{17}$ g, and $\delta \approx 1$ g cm^{-3}, we have $a_c \approx 50$ cm.

Dust grains with $a < a_c$ will be lifted over the surface and will move radially outward. After a distance of a few cometary radii, the gas drag effect will become insignificant and the dust particles will begin to move radially with a certain terminal velocity. For comet Halley at 1 AU solar distance, the velocity of μm-size particles is on the order of 0.3–0.5 km s^{-1}. The solar radiation pressure will continue to act on the dust particles such that, for particles emitted initially in the sunward direction, they will be eventually stopped at a parabolic envelope (see Fig. 22).

D. Thermal Emission

The solar radiation energy absorbed by cometary dust grains is mostly reemitted in the infrared wavelengths. For a spherical grain of radius a, the surface temperature of a dust particle will be determined by the following energy balance equation:

$$\int_0^\infty F_\odot(\lambda) Q_{abs}(\lambda, a)\pi a^2 \, d\lambda = \int_0^\infty \pi B(\lambda, T) Q_{abs}(\lambda, a) 4\pi\alpha^2 \, \delta\lambda \tag{26}$$

where $F_\odot(\lambda)$ is the solar radiation at wavelength λ, $Q_{abs}(\lambda, a)$ is the absorption efficiency, and $B(\lambda, T)$ is the Planck function. With $Q_{abs} = 1$ (i.e., for a blackbody) or $Q_{abs} =$ constant (i.e., for a gray body), we have

$$T = \frac{280}{\sqrt{r}} \text{ K} \tag{27}$$

Now, for a certain size distribution $n(a)$, the thermal flux is given as

$$F_\lambda = \int_{a_{min}}^{a_{max}} \pi B(\lambda, T) Q_{abs}(a, \lambda) a^2 n(a) \, da \tag{28}$$

where a_{min} and a_{max} are the minimum and maximum radii of the dust grains. A set of infrared spectra for the central coma of comet Halley taken in March 1986 are shown in Fig. 23. Superimposed on the thermal emission of heated dust particles with a temperature of the order

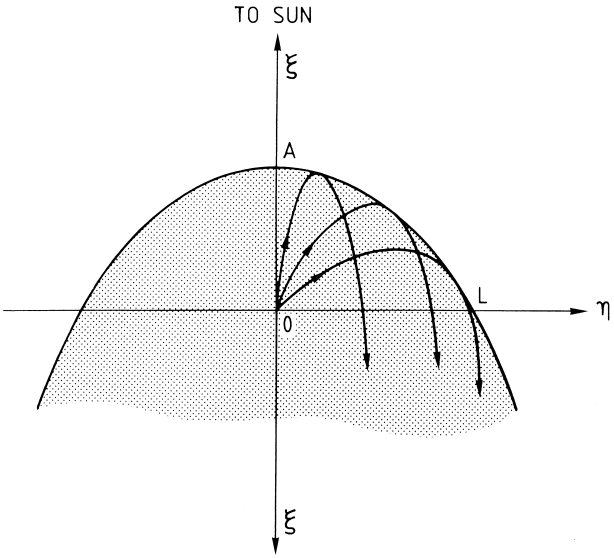


FIGURE 22 The fountain model of the dust trajectories in the sunward hemisphere near the nucleus together with their enveloping parabola. [Adapted from Mendis, D. A., Houpis, H. L. F., and Marconi, M. (1985). *Fundamental Cosmic Physics* **10,** 1.]

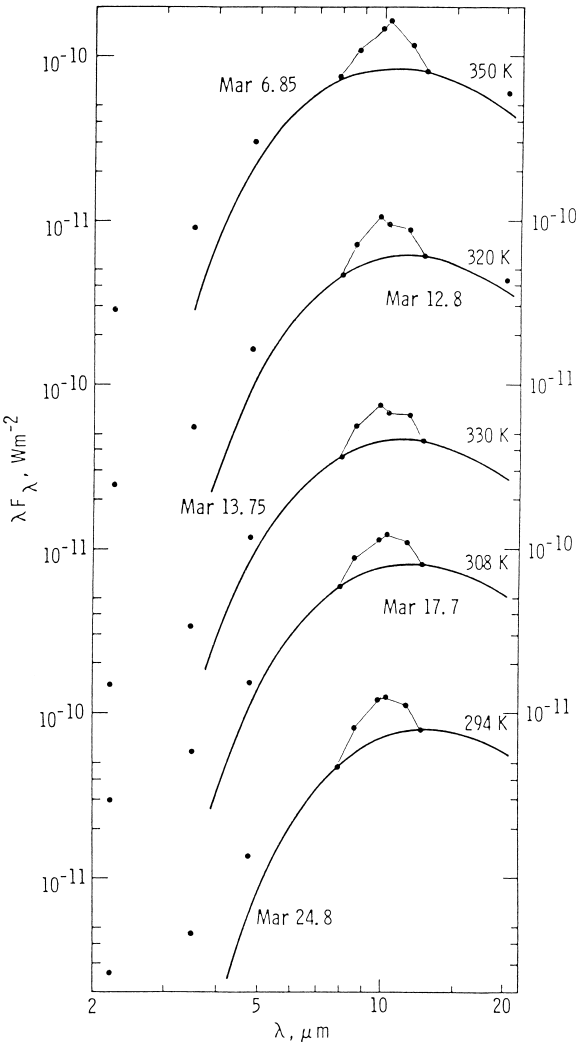


FIGURE 23 Thermal emissions from the central coma of comet Halley in March 1986, arranged chronologically. Each data set is offset by factor 10; the scales are shown on alternate sides of the figure. Blackbody temperature curves have been fit through the 7.8 μm and 12.5 μm data points, and the corresponding color temperatures are given. [From Hanner, M., *et al.* (1987). *Astron. Astrophys*. **187,** 653.]

of 300–350 K, a broad emission at 10 μm can be seen. This emission, together with another emission signature at 18 μm, was generally attributed to silicate grains; suggestions have also been made that these features could have come from polymerized formaldehyde grains in the coma.

E. Dust Jet Features

A combination of the presence of active regions on the cometary nuclear surface and rotational motion could lead to the formation of spiral dust jets in a cometary coma. Figures 14 and 17 show the closeup views of a system of dust jets in the coma of Comet Halley. Prominent jet structures had also been observed in the coma of Comet C/1995 O1 Hale–Bopp (Fig. 24). From the dynamical behavior of the jet morphology, the rotation period of this comet has been deduced to be about 11 h.

VII. ATMOSPHERE

A. Composition

1. Comet Halley

A compilation of the molecular abundances obtained from the observations of comet Halley is given in Table III. After H_2O, CO is the second most dominant species, with an abundance of about 10–15% relative to water. According to the neutral mass spectrometer (NMS) experiment on Giotto and rocketborne ultraviolet observations, a significant fraction (≈50%) of the CO was produced in a distributed region within a radial distance of about 25,000 km. The organic CHON dust grains or formaldehyde polymers could be the sources of these molecules. On the other hand, the third most abundant molecules, CO_2, appeared to have been emitted solely from the central nucleus. The heavy ion detector (PICCA) experiment on Giotto detected cometary ions with mass up to 213 AMU in the inner coma (see Fig. 25). The quasi-periodic spacing of the mass peaks has led to the suggestion that fragmentation of complex formaldehyde polymers $(H_2CO)_n$ or POMs could be responsible for the appearance of these heavy ions. But the presence of other kinds of hydrocarbons and organic molecules is also possible.

As the most primitive small bodies in the solar system, comets were generally thought to have formed at large

TABLE III Main Molecular Abundances of the Ice of Comet Halley

Molecules	Relative abundance	Remark
H_2O	1.0	
CO	0.10–0.15	Up to 50% from distributed source
CO_2	0.04	
CH_4	0.01–0.05	
H_2CO	0.04	
NH_3	0.003–0.01	Lower limit more likely
N_2	<0.001–0.1	Lower limit more likely
HCN	0.001	
OCS	<0.07	
CS_2	0.001	
S_2	<0.001	Inferred from observations of comet IRAS–Araki–Alcock (1983d)

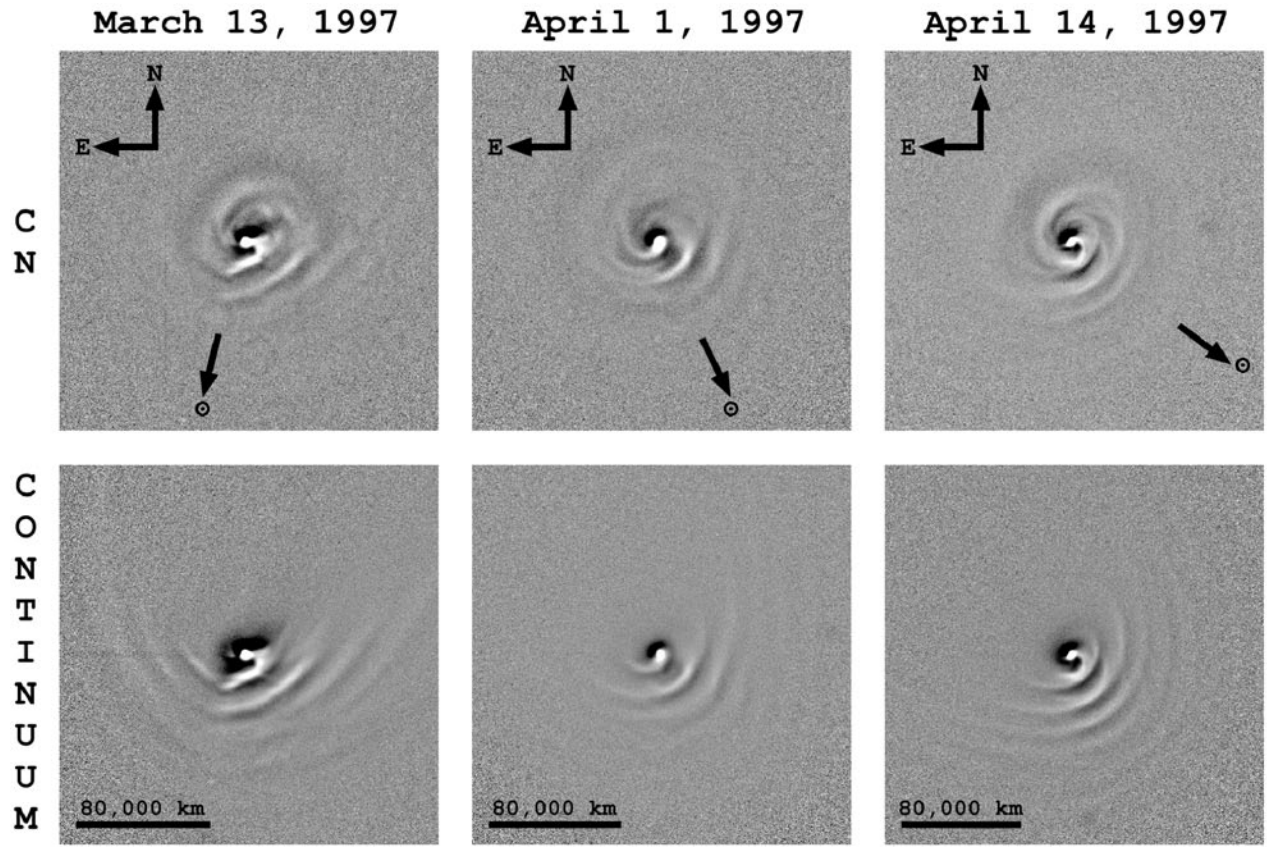


FIGURE 24 The enhanced images showing the coma morphology of comet Hale-Bopp in March/April, 1997. The direction to the Sun is shown with black arrows. [From B. E. A. Mueller, N. H. Samarasinha, and M. J. S. Belton, "Earth, Moon, Planets, . . .]

distances from the sun, where the condensation temperature is relatively low ($T_c \lesssim 50$ K). The bulk composition is therefore expected to be very volatile in comparison with other solar system objects, such as the asteroids. The compositional measurements of the gas and dust components at comet Halley indeed support this assessment. Figure 26 is a summary of the abundances of several major elements (C, N, O) relative to silicon. It is obvious that comet Halley's bulk composition is most similar to that of the sun.

The large amount of CO and CO_2 in comparison with the small abundance of CH_4 means that the chemistry in the solar nebula might be dominated by the oxidation of carbons, as may be found in interstellar clouds. The very small amount of N_2 may be understood in terms of the inefficiency of trapping N_2 during the low-temperature condensation of water ice. The details are still to be clarified by laboratory simulation of condensation processes.

According to the Giotto NMS experiment, the deuterium to hydrogen (D/H) ratio was determined to be $0.6 \times 10^{-4} < \mathrm{D/H} < 4.8 \times 10^{-4}$, which brackets the corresponding value in the ocean water of the earth ($\mathrm{D/H} = 1.6 \times 10^{-4}$). Both values represent a deuterium enrichment compared with interstellar space, where the D/H ratio $\approx 10^{-5}$. On the other hand, the cometary and planetary values are considerably smaller than the D/H ratios observed in several molecules (HCN, H_2O, HCO^+, etc.) found in interstellar dark clouds and in the organic inclusions of a few meteorite samples (the Chainpur meteorite shows a deuterium enrichment with D/H ratio up to 9×10^{-4}). Although ion chemistry may play a role in the deuterium enrichment of the molecules mentioned above, the D/H ratio of comet Halley could be more closely related to the ill-understood condensation process of icy planetesimals.

Two other isotope ratios deduced from the Giotto NMS experiment are $^{18}O/^{16}O = 0.0023 \pm 0.0006$ and $^{34}S/^{32}S = 0.045 \pm 0.010$, which, within the experimental errors, are both equal to the corresponding terrestrial values. Thus, the D/H, $^{18}O/^{16}O$, and $^{34}S/^{32}S$ ratios all seem to indicate that comet Halley's isotopic compositions are the same as those of the solar system material. One exception is the

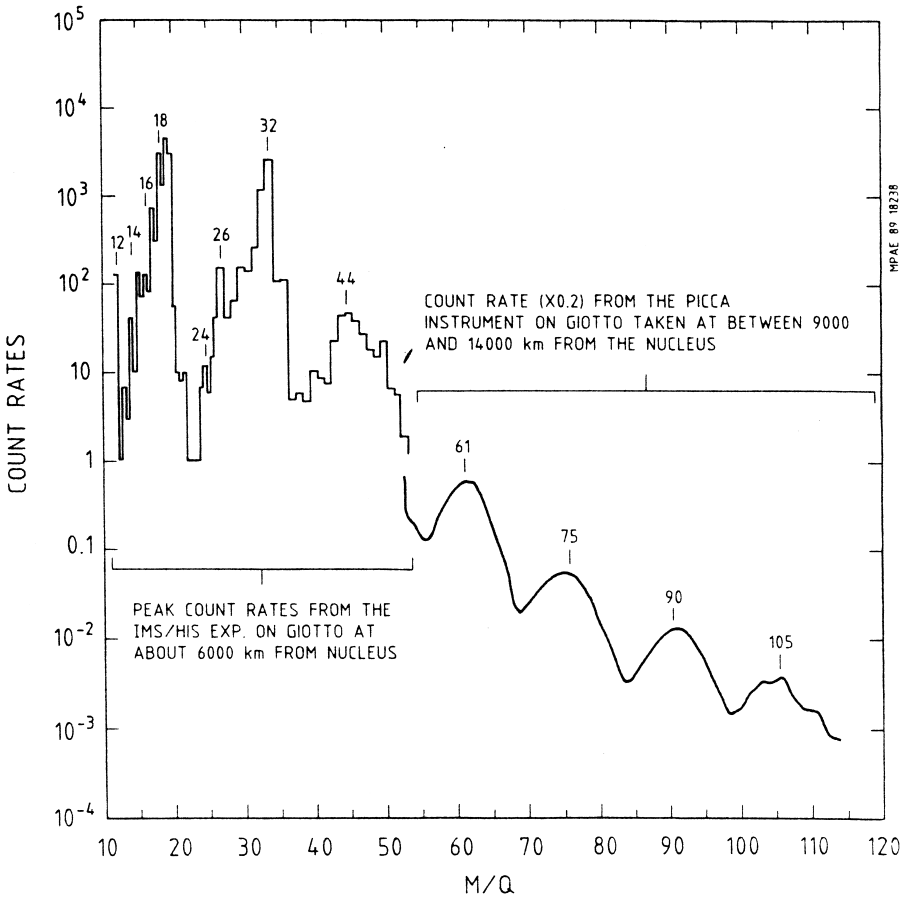


FIGURE 25 A composite ion mass spectrum taken by the IMS and PICCA instruments on Giotto in the coma of comet Halley. [From Balsiger, H., *et al.* (1986). *Nature*, **321,** 326, and A. Korth *et al.* (1986). *Nature*, **321,** 335.]

$^{12}C/^{13}C$ ratio which was determined to be ≈65 by ground-based high spectral resolution measurements of CN.

2. Comets Hyakutake and Hale–Bopp

The apparitions of two bright comets, C/Hyakutake and C/Hale–Bopp in 1995 and 1996, had provided new information on the chemical compositions of long-period comets. This was particularly opportune because of the availability of several advanced radio telescopes and the operation of the Infrared Space Observatory at that time, permitting the tracking of the coma developments of C/Hale–Bopp from large heliocentric distances to perihelion. Table IV shows the molecular abundance of C/Hyakutake and of C/Hale–Bopp. It can be seen that at a heliocentric distance of 1 AU, the molecular composition of C/Hyakutake is not too different from that of Comet Halley. On the other hand, the relative abundances of CO and of CO_2 are much larger for C/Hale–Bopp at 4 AU. This effect is consistent with the expectation that volatile ices such as CO and CO_2 should be the first to sublimate at large heliocentric distance when the surface temperature is still too low for water ice to sublimate at high level of gas production rate. Figure 27 summarizes the gas production rates of molecular species observed in radio wavelengths. The crossover of the CO and H_2O values at about 4 AU during the inbound orbit of C/Hale–Bopp is most conspicuous.

B. Coma Expansion

In the inner coma, the cometary outflow is collision-dominated within a radial distance of r_c given as

$$r_c = \frac{Q\sigma}{4\pi u_{gas}} \tag{29}$$

where Q is the total gas production rate, σ is the collisional cross section among the neutral molecules, and u_{gas} is the expansion speed of the neutral coma. For comet Halley at a solar distance of 1 AU, $Q \approx 10^{30}$ molecules s^{-1}, we obtain $r_c = 25,000$ km if $\sigma = 3 \times 10^{-15}$ cm^2 and $u_{gas} \approx 0.8$ km/s.

At sublimation, the gas molecules have an average radial velocity below the sonic value. But the expansion velocity quickly reaches supersonic value within a short

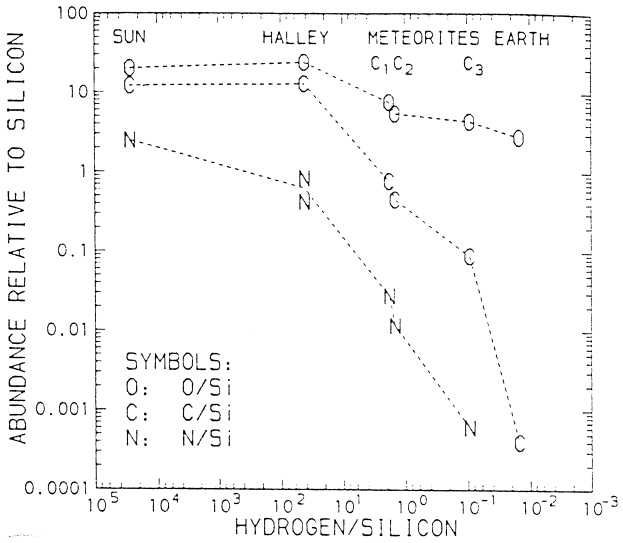


FIGURE 26 Relative abundances of elements in the material released by comet Halley. A gas/dust ration of 2 at the source was assumed. For comparison, the abundances in the solar nebula and in carbonaceous are given. Estimates for the abundance on the earth (crust plus mantle) are also included. [From Geiss, J. (1987). *Astron. Astrophys.* **187,** 859.]

distance ($\lesssim$30 m) from the nuclear surface. Several effects are important in the thermal budget and in further acceleration of the gas. First, adiabatic cooling is essential in reducing the temperature of the coma gas to a value as low as 10–30 K at a cometocentric distance of 300 km. Second, as a result of photolytic processes, the photodissociation fragments could obtain a significant amount of excess kinetic energy ($\approx$ a few eV) at dissociation. For example, hydrogen atoms from H_2O dissociation move with an initial speed of around 20 km s^{-1}. The fast H-atoms created inside the collision-dominated region will be thermalized and slowed down via collision with the background gas. Such an energy-transfer process is of importance in the photolytic heating of the coma. Since infrared emission due to rotational transitions in the highly polar molecules H_2O could be a very effective cooling mechanism, the total energy budget is then determined by a balance between the infrared cooling and photolytic heating. The exact magnitude of the infrared cooling rate, however, is still subject to debate. In Fig. 28 we show the effects of including and neglecting the IR cooling effect on the expansion velocity of the gas. A detailed fit with the gas velocity from the Giotto NMS experiment can be achieved by adjusting different parameters in the momentum-transfer process and energy budget.

The images of the near-nucleus environment of comet Halley indicate that gas and dust production were concentrated mostly on the sunlit side. Further evidence was provided by the TKS spectroscopic experiment on Vega, which could be used to map the spatial distribution of dust and water vapor in the vicinity of the comet nucleus. According to the experimenters, most of the dust emission is confined to a narrow cone with a half-angle of about 25°. They further suggested that the angular distribution of H_2O vapor up to several thousands of kilometers from the nucleus is even narrower. Given such narrow gas jet structure, theoretical model calculations generally show that an anisotropic gas outflow should be maintained to a cometocentric distance of a few 10^4 km. An unexpected result is that the neutral gas density distribution measured by the Giotto NMS experiment can be well fitted by a spherically symmetric model.

Besides the jet structures, expanding shells of CN radicals were observed in the coma of comet Halley. Their formation could be characterized by two quasi-periods of 2.2 days and 7.4 days, respectively (cf. Section V.G). This means that the outgassing process of comet Halley was strongly modulated by the rotational movement of its nucleus. Such halo formation had already been noticed in the 1910 return of comet Halley.

The expansion velocity of the CN shells was found to vary with the heliocentric distance at production. At $R \approx 1.3$ AU, the shell velocity V_s is about 0.8 km s^{-1} and at $R \approx 0.8$ AU, V_s increases to 1.2 km s^{-1}. Such systematic variation can be explained in terms of the dependence of the photolytic heating rate on the solar distance.

TABLE IV Molecular Abundances in Comets

Molecule	T(K)	C/Hyakutake at 1 AU	Others at 1 AU	C/Hale-Bopp at 1 AU
H_2O	152	100.	100.	100.
CO	24	5.–30.	2.–20.	80.
CO_2	72	$\leq$7.	3.–6.	30. at 4.6 *AU*
CH_4	31	0.7	$\leq$0.5–2.	
C_2H_4	54	0.3–0.9		
C_2H_6	44	0.4		
CH_3OH	99	2.	1.–7.	6.
H_2CO	64	0.2–1.	0.05–4.	0.1–0.2
NH_3	78	0.5	0.4–0.9	
N_2	22		0.02	
HCN	95	0.15	0.1–0.2	0.6
HNC		0.01		
CH_3CN	93	0.01		
HC_3N	74		$\leq$0.02	
H_2S	57	0.6	0.3	6.
OCS	57	0.3	$\leq$0.5	
S_2		0.005	0.02–0.2	
SO_2	83		$\leq$0.001	

[From Bockelee-Morvan, D., *in* IAU Symp. No. 178, "Molecules in Astrophysics: Probes and Processes" (van Dishoeck, E. F., ed), Leiden July 1–5, 1996.]

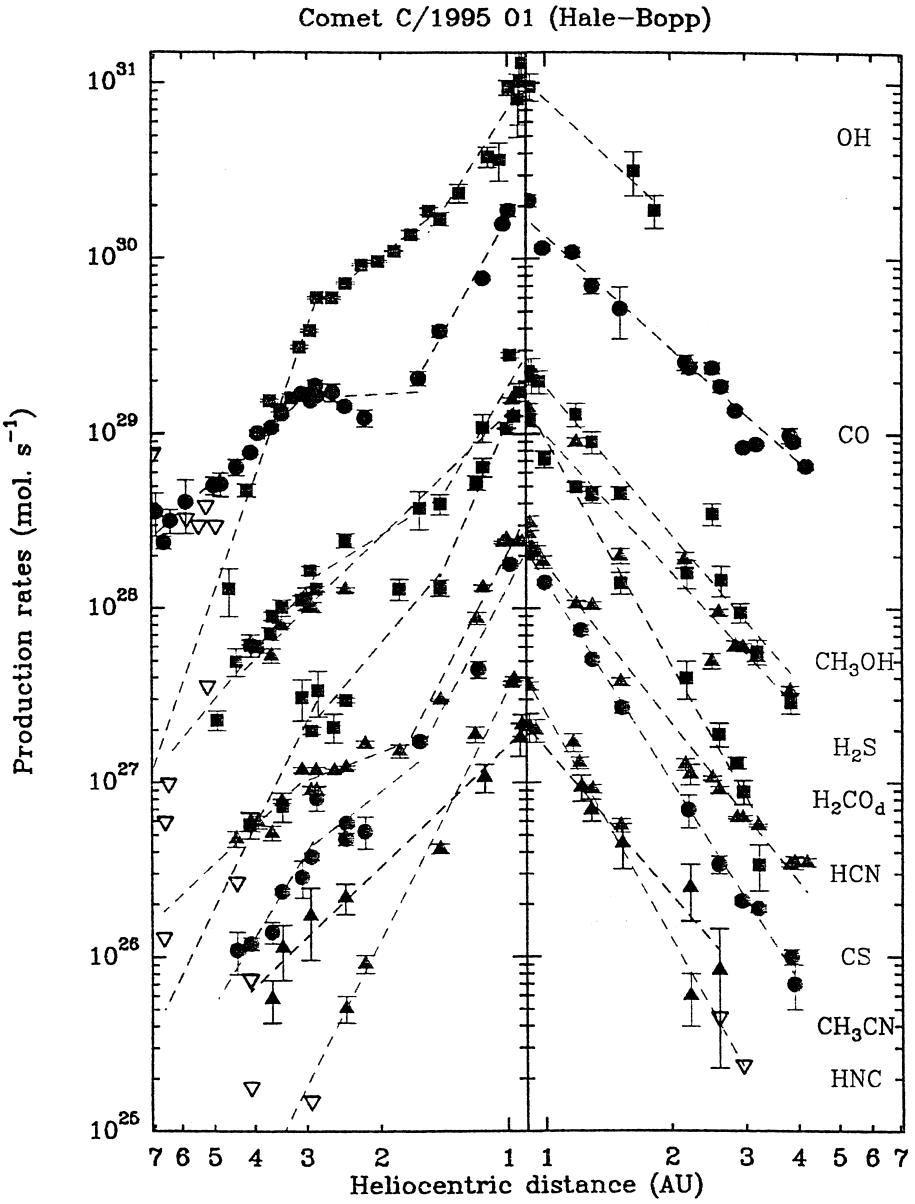


FIGURE 27 Gas production curves from observations at IRAM, JCMT, CSO, SESST and Nancay radio telescopes. [From Bockelee-Morvan, D., and Rickman, H., *Earth, Moon, and Planets*: *An International Journal of Solar System Science* (M. E. Bailey, ed.), Dordrecht, Holland.]

Another topic of interest concerns the possible recondensation of water droplets. During coma expansion, small icy grains may be dragged away from the nuclear surface together with the sublimating gas flow. The detection of an icy grain halo is quite difficult, however, as the lifetimes of icy grains—if they contain dark-colored absorbing material—should be very short, thus limiting the dimension of the icy grain halo to just a few hundred kilometers at 1 AU solar distance. Only for solar distances >3 AU would the low evaporation rate of the icy grains allow the icy grain halo to be stable against sublimation. It is perhaps for this reason that the 3-μm absorption signature of water ice was detected for comets Bowell (1980b) and Cernis (1983f) only at large solar distances. The concept of a distributed source is still a valid one if the recondensation of water molecules into particulate droplets during the adiabatic cooling phase is possible. The icy droplets (or clusters of a few hundred H_2O molecules) so formed would probably have radii of the order of 15 Å. Depending on the fraction of water vapor that might go through the recondensation process, the latent heat released within a few nuclear radii could lead to substantial increase of the gas temperature.

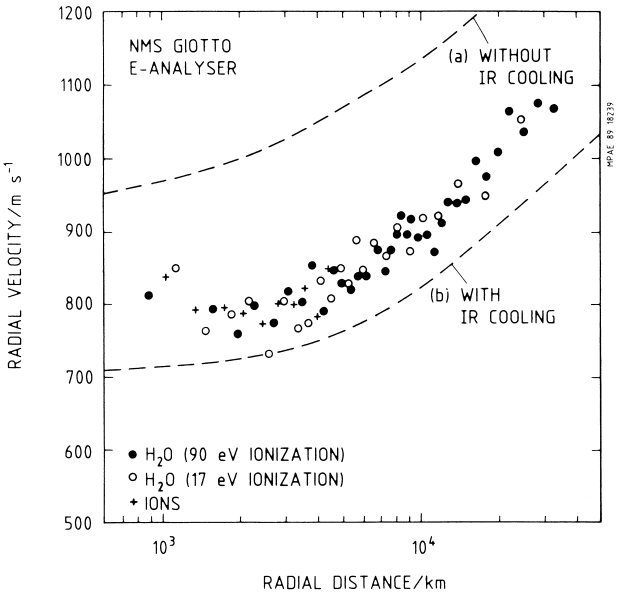


FIGURE 28 A comparison of the neutral gas radial velocity measured by the Giotto NMS experiment with theoretical model calculations assuming the presence and absence of infrared cooling effect of the water molecules in the coma. [Giotto NMS data from Lämmerzahl, P., *et al.* (1987). *Astron. Astrophys.* **187,** 169.]

No clear indication of the existence of icy droplets in the inner coma of comet Halley (or other comets) has been found. A possible indirect method of investigating this potentially important effect is to simulate the surface sublimation of water ice in a vacuum under controlled conditions. Laboratory simulation experiments should be designed to examine *ang* the possible production of water dimers $[(H_2O)_2]$ and other water clusters of larger structures.

VIII. PLASMA

Before the space missions to comets Giacobini–Zinner and Halley, a general picture of solar-wind–comet interaction included the following steps (see Fig. 29):

(a) Ionization and pickup of the cometary ions beginning at large cometocentric distances ($r \gtrsim 10^6$ km)
(b) Heating and slowdown of the solar-wind flow due to assimilation of the heavy cometary ions
(c) Formation of a cometary shock
(d) Amplification of the magnetic field strength on the front side of the coma and draping of the field lines into a magnetic tail
(e) Stagnation of the solar-wind plasma flow by ion-neutral friction in the inner coma
(f) Formation of a contact surface shielding the outward-expanding ionospheric flow from the external plasma flow

It is satisfying that *in-situ* measurements at these two comets have confirmed this global picture. There are, however, many new and interesting effects that had not been previously considered. In this section we shall focus our attention on the major observational results from the spacecraft measurements. The survey will start from the inner region.

A. The Ionosphere

In the inner coma within a few thousand kilometers, where $r < r_c$ [see Eq. (24)], the ion temperature should be similar to the neutral temperature, since these two gases are closely coupled in this region. The ion temperature inside the contact surface of comet Halley at about 4800 km was determined to be ≈200 K by the Giotto IMS and NMS experiments (see Fig. 30). This result is in reasonable agreement with theoretical calculations. Because of the very effective cooling by the H_2O molecules via inelastic collision, the temperature of the electron gas (T_e) is expected to be nearly the same as the temperature of the neutral gas over a radial distance of about 1000 kilometers, beyond which point there is a sharp increase of kT_e to ≈2–3 eV. The exact location and gradient of T_e depends on a combination of the H_2O electron-neutral cooling function, electron transport, and thermal conductivity in the actual

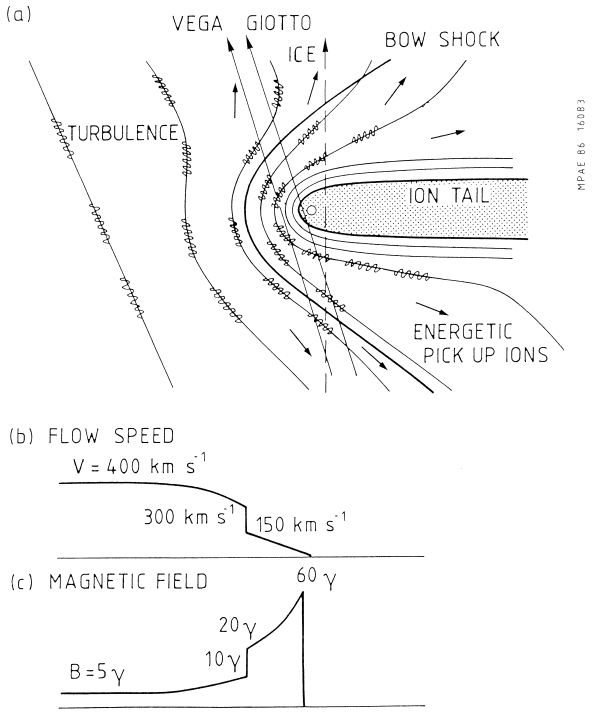


FIGURE 29 A schematic view of the comet–solar-wind interaction process.

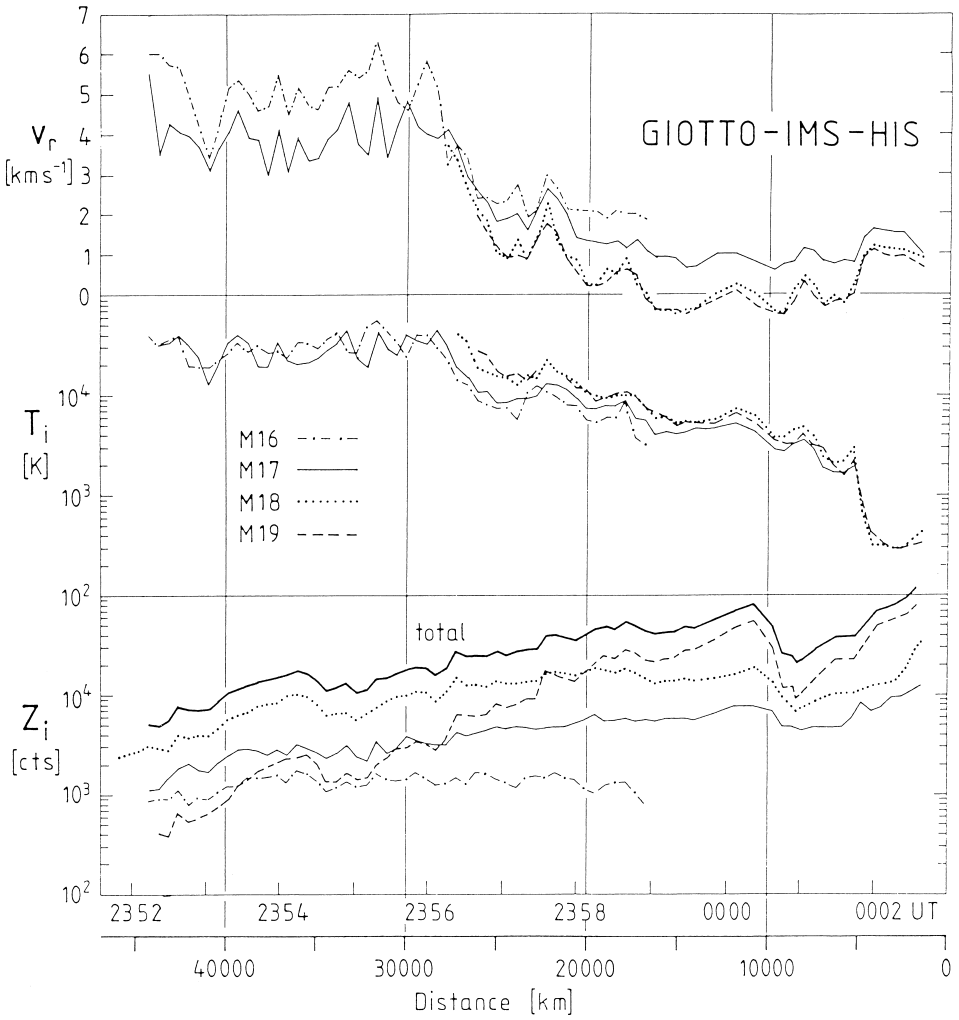


FIGURE 30 Profiles of the ion velocity (V_r) opposite to the spacecraft ram direction in the rest frame of the nucleus, temperatures (T_i), and the relative densities (Z_i) for ion masses M16, M17, M18, and M19. The measurements are from the Giotto IMS experiment. [From Schwenn, R., *et al.* (1987). *Astron. Astrophys.* **187,** 502.]

situation. Because of a lack of plasma instruments capable of measuring low-energy electrons ($kT_e < 1$ eV) the Vega and Giotto missions have not been able to shed light on the ionospheric electron temperature distributions in the coma of comet Halley. Note that the plasma wave experiment on the ICE spacecraft to comet Giacobini–Zinner measured a very dense, cold plasma in the central part of the ion tail. The electron temperature there is as low as 1.3×10^4 K (i.e., $kT_e \approx 1$ eV). Since the point of closest approach to the nucleus was at a distance of 7800 km, the electron temperature near the center of the ionosphere ($r < 1000$ km) is expected to be even lower.

The observational data from the Giotto mission for the ion temperature and the radial velocity of the ions are summarized in Fig. 30. The discontinuities at the contact surface are most notable. The ions created inside the contact surface are very cold, with a temperature of $T_i \approx 200$ K; those outside have $T_i \approx 3000$ K. While the external plasma is nearly stagnant, the cold ionosphere plasma has a radial outward velocity of about 1 km s^{-1}. The magnetic field was found to drop from a value of 50 nT to effectively zero at this interface (see Fig. 31). The general behavior of the magnetic field variation in the vicinity of the contact surface can be understood in terms of a balance between the ion-neutral friction and Lorentz force, that is,

$$J \times B = K_{in} n_i m_i n_n u_{gas} \tag{30}$$

or

$$\frac{1}{4\pi}\left(B\frac{\partial B}{\partial R} + \frac{B^2}{R_{cur}}\right) = k_{in} n_i m_i n_n u_{gas} \tag{31}$$

where k_{in} is the ion–molecule collision rate, n_n is the neutral number density, n_i is the ion number density, u_{gas} is

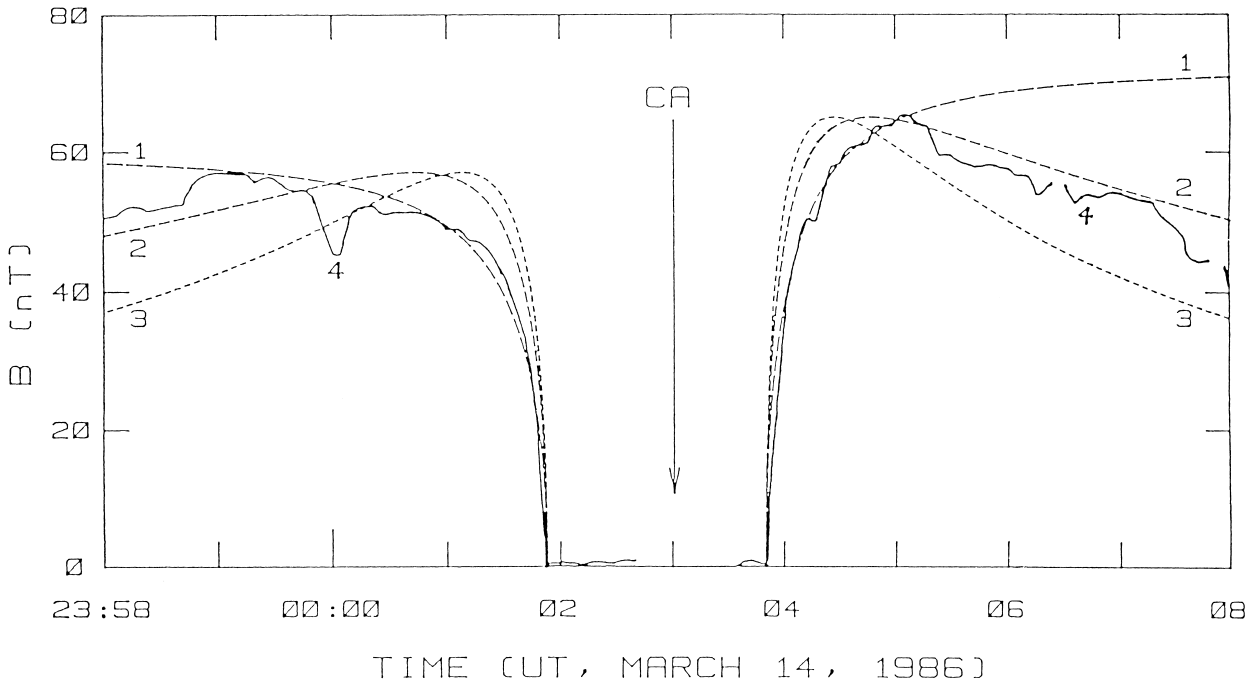


FIGURE 31 Magnetic field measurements made by the Giotto magnetometer experiment showing the inner pile-up region inbound and outbound and the magnetic cavity region. Theoretical curves following the ion-neutral friction theory are also compared. [Experimental curves from Neubauer *et al.* (1986). *Nature* **321,** 352; theoretical curves from Wu, z.-J. (1987). *Ann. Geophys.* **6,** 355.]

the neutral gas speed, and R_{cur} is the radius of curvature of the magnetic field lines.

The plasma can be taken to be stationary as a first approximation, and the cavity boundary can be assumed to be fixed. Furthermore, the ionospheric plasma inside the contact surface can be described in terms of photochemical equilibrium such that the ion number density can be expressed as

$$n_{\text{i}}(R) = \left(\frac{\beta Q}{4\pi\alpha u_{\text{gas}}}\right)^{1/2} \frac{1}{R} \tag{32}$$

where α is the electron dissociative recombination coefficient and β the photoionization rate. In this case Eq. (32) has a very simple solution for the magnetic field profile:

$$B(R) = B_{\text{max}} \frac{[1 - 2\ln(R/R_{\text{max}})]^{1/2}}{(R/R_{\text{max}})} \tag{33}$$

where R_{max}, proportional to $Q^{3/4}$, is the radial distance at which the piled-up magnetic field reaches its maximum value $B_{\text{max.}}$ The essential feature of a rapid decrease of the magnetic field by about 40 to 60 nT over a distance of about 3000 km can be reproduced rather well. According to the Giotto magnetometer experiment, a final decrease of the magnetic field strength by $\Delta B \approx 25$ nT takes place over a radial distance as small as 25 km. This abrupt change cannot be explained in terms of ion-neutral friction; and additional force such as particle pressure might be invoked to account for this effect.

At the boundary of the contact surface, a spike of ion density enhancement should be produced, accompanying the sharp rise of the magnetic field. The structure of this density jump can be described in terms of accumulation of the ionospheric plasma into a stationary thin shell. The total content of ionization is determined by the injection of ionospheric particles and their subsequent loss via electron dissociative recombination in this layer.

Outside the magnetic field-free cavity, in the region $r = 10^4$ to 2.4×10^4 km, the ion temperature $T_{\text{i}} \approx 10^3$–$10^4$ K and the ion velocity $v_{\text{i}} \approx 3$–6 km s^{-1}. The ion temperature was determined by a thermal equilibrium condition in which

$$T_{\text{i}} \approx T_{\text{n}} + \frac{m_{\text{i}}}{3k}(v_{\text{i}} - v_{\text{n}})^2 \tag{34}$$

In this region, another interesting feature is that there was a sharp discontinuity in the ion density at $r \approx 10^4$ km (see Fig. 32). Outside the density maximum at about 3×10^4 km, the cometary ion density profile follows an r^{-2} dependence. Inside the density maximum, the density profile has an r^{-1} dependence, as predicted by photochemical equilibrium models in which the ion production

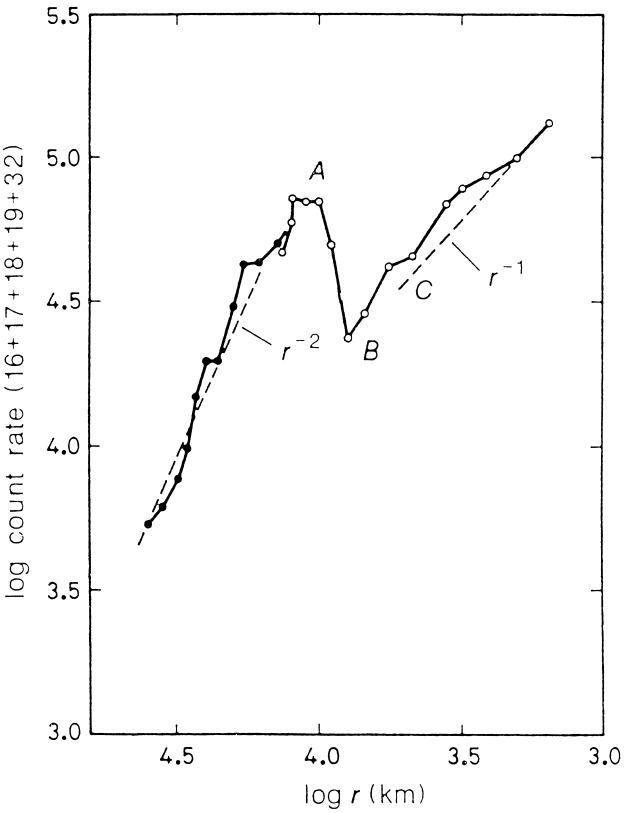


FIGURE 32 The radial profile of the sum of the ion counting rates for masses 16 to 19 and 32, showing that inside the magnetic cavity (C at $r = 4600$ km) the total counting rate tends to follow a $1/r$ dependence and that in the outer part ($r > 16{,}000$ km) a $1/r^2$ dependence fits the data well. The peak in the ion count rate is at A and the minimum is at B. [From Balsiger, H., *et al.* (1986). *Nature* **321,** 326.]

via photoionization is balanced by electron dissociative recombination.

A possible explanation for such a change in the ion density distribution is that there exists a gradient in the electron temperature at 10^4 km with the electron temperature T_e being much higher outside than inside. In this scenario, the plasma loss effect via electron dissociative recombination will mainly occur inside this radius. It should be mentioned that such a feature appeared to be quasi-stationary during the time interval of the spacecraft encounters with comet Halley. In addition to the Giotto measurements, both the plasma measurements on Vega and ground-based spectroscopic observations of the H_2O^+ emission in the coma revealed such a plasma structure.

One possible mechanism responsible for this electron temperature gradient is that the electron temperature could depend on the configuration of the magnetic field. In a magnetically closed region, the electron temperature is largely determined by a balance between photo-electron heating and collisional cooling by the neutral coma. In contrast, in the case of an open-field region, there could be an extra heat input via thermal conduction along field lines from the external interplanetary medium.

B. Cometopause

Another interesting result from *in-situ* measurements by the Vega and Giotto probes at comet Halley concerns the detection of the so-called cometopause separating the plasma regimes dominated by the fast-moving solar-wind protons on one side and by the slow, cold, heavy ions of cometary origin on the other side. The Vega plasma experiment first discovered this sharp boundary of about 10^4 km thickness at a cometocentric distance of about 1.4×10^5 km (see Fig. 33). While the solar-wind proton flux quickly decreased across this boundary, the intensity of the cometary heavy ions rapidly increased. Such a plasma boundary effect was also observed during Giotto inbound at a cometocentric distance of 1.35×10^5 km, although the transition was not as sharply defined as that seen by Vega. The issue was further complicated by the fact that near this location a sudden jump of the magnetic field strength by 20 nT was detected by Giotto's magnetometer experiment, whereas the magnetic field variations measured by the Vega 1 and 2 spacecraft across the cometopause region were rather smooth. Within the framework of standard magnetohydrodynamic process in comet–solar-wind interaction, a model calculation separating the chemical processes from the two-dimensional plasma flow dynamics shows that charge exchange would be effective in eliminating solar-wind protons and hot cometary ions only at cometocentric distance ≤ 6–8×10^4 km. This theoretical result is consistent with the IMS observations on Giotto. Thus it may be concluded that the charge-exchange effect is not the immediate cause of the formation of the cometopause structure. On the other hand, plasma instabilities, such as the nonresonant firehose instability, might be important in leading to such a cometary plasma boundary under appropriate conditions.

C. Cometary Bow Shock and the Upstream Region

The transition region between the supersonic and the subsonic solar-wind flow was found to be very diffuse in the case of comets Giacobini–Zinner and Halley (see Fig. 34). The wavy and oscillatory patterns observed are somewhat similar to the quasi-parallel bow shock of the earth's magnetosphere, in which case the solar-wind flow direction is nearly parallel to the normal of the shock surface. The generation of large-amperiod of 75 to 135 s. These "100-s" waves have been suggested to correspond to ion cyclotron waves for water-group ions (i.e., H_2O^+, O^+, OH^+, etc.) ionized and picked up in the solar wind.

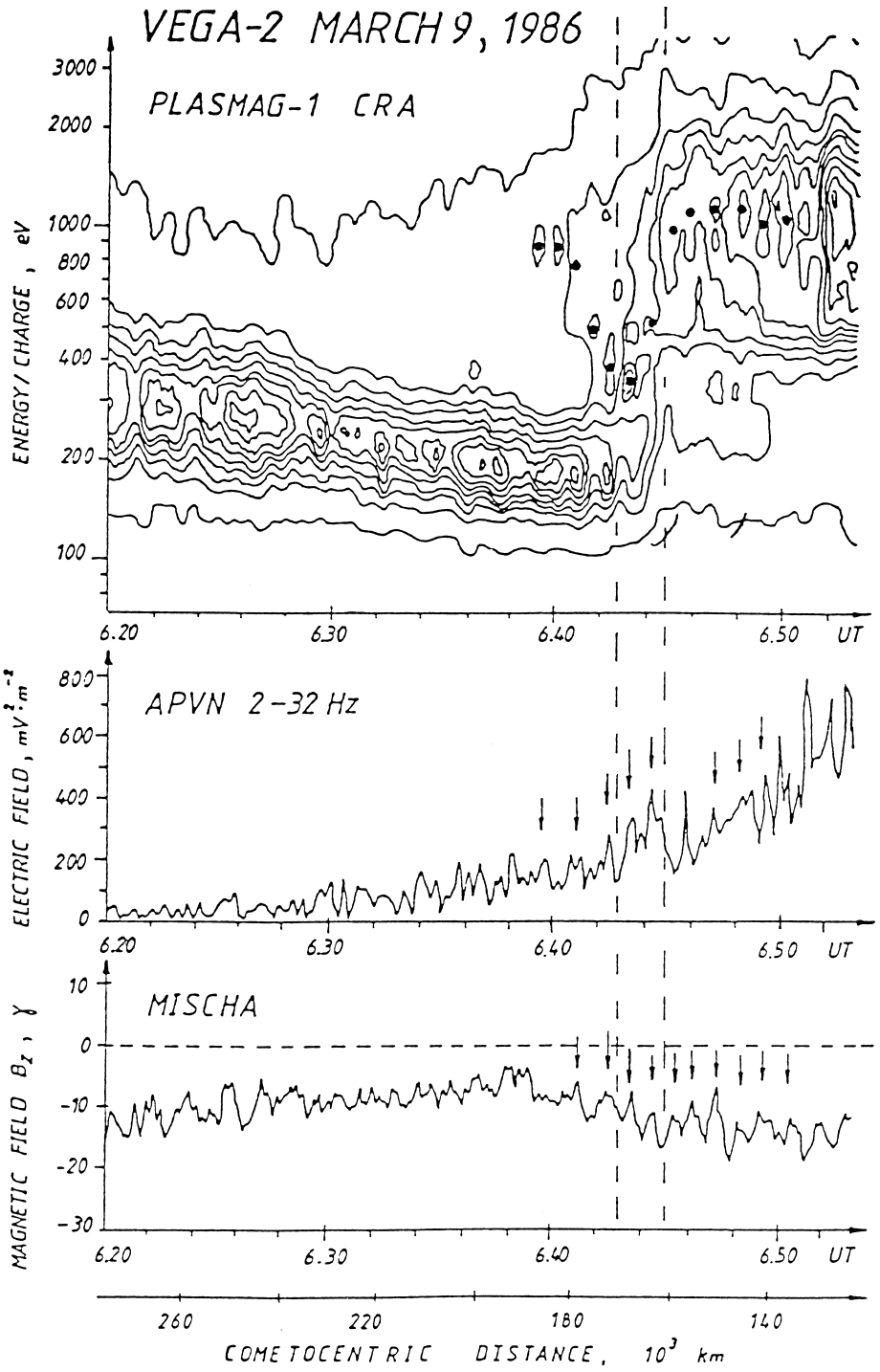


FIGURE 33 The formation of a cometopause structure (indicated by vertical dashed lines) as indicated by the ion measurements on Vega 2. The solar wind protons disappeared abruptly at about 1.5×10^5 km away from the comet nucleus. A cold population of heavy cometary ions showed up afterward with increasing fluxes. Oscillations in the ion flux, electric field, and magnetic field components (B_z pointing toward the north pole of the ecliptic plane) were also observed in the vicinity of the cometopause. Maxima are shown by dots and arrows. [From Galeev, A. A., *et al.* (1988). *J. Geophys. Res.* **93,** 7527.]

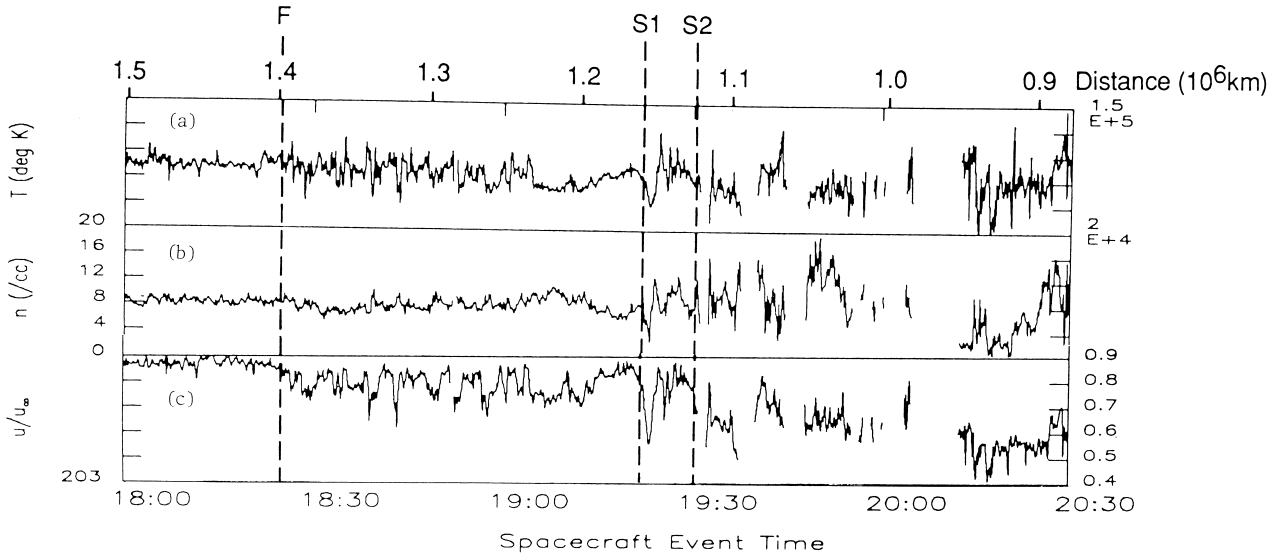


FIGURE 34 Solar wind proton bulk flow parameters in the bow shock region: (a) temperature [T], (b) number density [n], and (c) flow speed (u). The foreshock is marked F, the beginning of the low-speed dip i is marked S1, and the "permanent" speed change is marked S2. [Adapted from Coates, A. J., *et al.* (1987). *Astron. Astrophys.* **187,** 55.]

The basic cause of the plasma turbulence is associated with free energy produced by pickup of cometary ions. Initially, the velocity of these ions relative to the solar wind plasma flow has two components, $V_{\parallel}$ and $V_{\perp}$, as shown in Fig. 35. In the quasi-perpendicular case with $\alpha \approx \pi/2$, the cometary ions are accelerated by the $V_{sw} \times B$ electric field into cycloidal trajectories. In the solar-wind frame, the new ions gyrate around the magnetic-field lines with large pitch angles. Such a velocity distribution is unstable to the growths of left- and right-hand polarized Alfven (L and R) waves. The propagation direction of these waves should be sunward along the average, spiral magnetic-field direction.

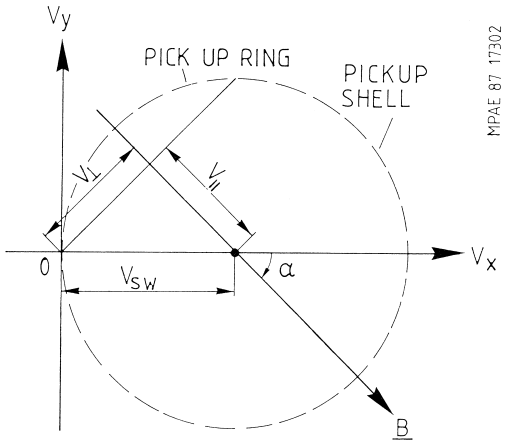


FIGURE 35 The two components, $V_{\parallel}$ and $V_{\perp}$, of the initial velocity of the cometary pickup ions. Pitch angle scattering effect leads to a transformation of the ring distribution into a thin spherical shell.

If the average magnetic field is nearly parallel to the solar-wind flow direction, the new cometary ions are at first stationary with respect to the comet. However, the resultant cold beam of heavy ions having a speed of about V_{sw} relative to the solar wind leads to the nonresonant firehose instability under the following condition:

$$P_{\parallel} > P_{\perp} + \frac{B^2}{8\pi} \tag{35}$$

where $P_{\parallel}$ is the thermal plasma pressure component parallel to the magnetic field and $P_{\perp}$ is the component in the perpendicular direction. This implies that the firehose instability, which excites long-wavelength waves, becomes possible when the cometary ion pressure exceeds the sum of the magnetic-field and solar-wind thermal pressure.

The main consequence for resonating ions of the various plasma instabilities is to rapidly isotropize the ring-beam distribution into a spherical shell. The growth rate is usually on the order of 0.1 Ω_i to Ω_i (ion gyrofrequency). The large level of magnetic-field turbulence will in turn permit stochastic acceleration of the cometary ions at large distances from the cometary bow shock. The detection of energetic heavy ions with energies up to a few hundred keV in large upstream regions where other acceleration processes are weak has been considered generally to be evidence of the second-order Fermi acceleration process. Other effects, such as first-order Fermi acceleration and lower hybrid turbulence, could be operational at localized regions at the same time.

As for high-frequency waves, one somewhat surprising discovery of cometary kilometric radiation (CKR)

was made by the Sakigake spacecraft as it flew by comet Halley. The plasma wave experiment picked up discrete radio emissions in the frequency range of 30 to 195 kHz. These emissions occurring at the local plasma frequency may be the result of conversion of the electrostatic plasma waves to electromagnetic waves in the turbulent plasma environment of comet Halley. The generation of the CKR thus may be similar to the Type II solar radio bursts from coronal shock waves. In fact, the cometary bow shocks were identified as the source region of the CKR. It was further proposed that motion of the cometary bow shock could excite Alfven and Langmuir turbulences that eventually lead to the CKR emissions.

D. The Ion Tail

Combining the plasma measurements at comet Giacobini–Zinner on the tailward side of the comet at a cometocentric distance of 7800 km, we might produce a schematic model for the cometary magnetic field configuration as follows (see Fig. 36):

(a) The field draping model is basically confirmed except for the detection of the formation of a magnetosheath at the ion tail boundary. The magnetic field strength in the lobes of the ion tail is on the order of 60 nT. This relatively high field may be explained in terms of pressure balance at the tail boundary, where the total external pressure of the cometary ions was as large as the solar-wind ram pressure.

(b) A thin plasma sheet with a total thickness of about 2000 km and a width of about 1.6×10^4 km was found at the center of the ion tail. The peak electron density and an electron temperature were determined to be $n_e = 6.5 \times 10^2$ cm^{-2} and $T_e = 1.3 \times 10^4$ K by the plasma wave instrument.

(c) The plasma flow velocity gradually decreased to zero at the ion tail center. A significant amount of electron heating was seen between the ion tail and the bow shock.

We expect similar morphologies to be found in the plasma environment of comet Halley after appropriate spatial scalings. Snapshots by spacecraft flyby observations, however, do not reflect the many time-variable features seen in the ion tails of bright comets. For instance, the structure of the ion tail is often characterized by the appearance of a system of symmetric pairs of ion rays, with diameters ranging between 10^3 and a few 10^4 km, folding toward the central axis. When the ion rays are first formed, with an inclination of about 60° relative to the central axis, the angular speed is high with a linear speed $\approx$50 km s^{-1}. But near the end of the closure, the perpendicular speed is extremely low, no more than a few kilometers per second. Time-dependent MHD simulations have been applied to these phenomena with emphasis on the effect of temporal changes of the interplanetary magnetic field. No satisfactory answers to the formation of the ion rays have yet been found. A similar situation exists for the large-scale ion-tail disconnection events that had been suggested to be result of magnetic field reconnection.

E. Electrostatic Charging

Cometary dust grains, once released from the nucleus surface, will be subject to solar radiation and plasma interaction. The emission of photoelectrons and the impact of positive ions would be balanced by the surface collection

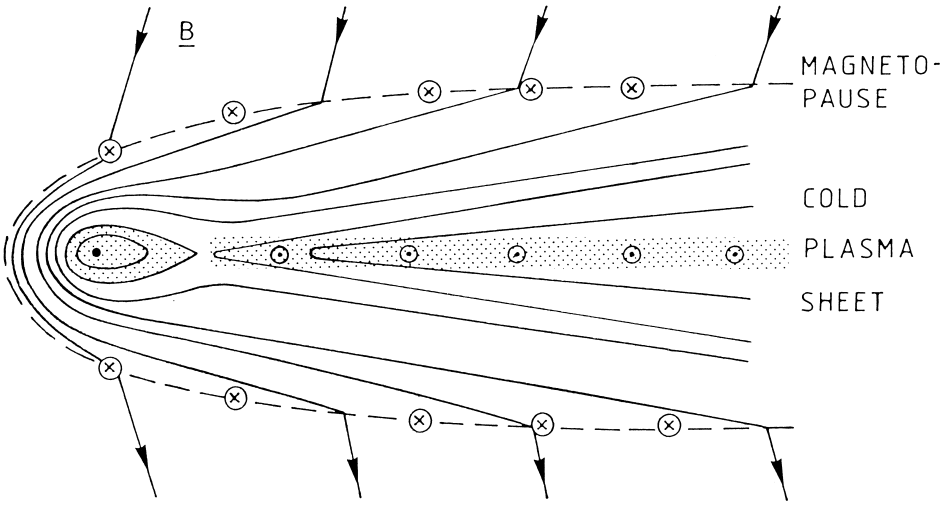


FIGURE 36 The general nature of the magnetic field configuration as observed by the ICE spacecraft at comet Giacobini–Zinner. The kinky structure of the magnetic field lines at the "magnetopause" is generated by a current layer required to couple the slow-moving cometary plasma with the external solar wind.

of electrons. The electrostatic potential of the grain surface is determined by the equilibrium condition that the net current should be zero. The electron charge on the dust grain with surface potential ϕ is

$$q = \frac{a\phi}{300} \text{ e.s.u.} \tag{36}$$

with ϕ in volts and particle radius in centimeters. In the solar wind, the Lorentz force experienced by the charged dust particle can be written as

$$f_L = q(E + V_d \times B) \tag{37}$$

where V_d is the dust velocity relative to the comet, B is the interplanetary magnetic field, e is the electronic charge, and $E = -V_{sw} \times B$ is the interplanetary motional electric field acting on the charged grain. For submicron particles, the Lorentz force for a ϕ value of a few volts is sufficient to produce strong perturbations on the orbital motion. Furthermore, during unusual episodes of enhanced energetic electron impact, the electrostatic charging effect could become strong enough to permit fragmentation of the dust particles. Some inhomogeneous structures of cometary dust tails not explained by the syndyname or synchrone formulation may be a consequence of such an effect.

At large solar distances where surface sublimation activities completely subside, the direct interaction of the solar wind plasma with the nucleus surface could lead to differential electrostatic charging of the surface. Against the weak gravitational force of the cometary nucleus, it is possible that some sort of dust levitation and surface transport process might occur. The details are very poorly understood, however.

F. X-Ray Emissions

Very intense extreme ultraviolet radiation and X-ray emissions at the coma of Comet C/Hyakutake 1996 B2 were discovered by chance by the German Roentgen X-ray satellite (ROSAT). Many more cometary X-ray extended sources were subsequently detected. Such X-ray emissions were now understood to be produced by charge transfer process between the neutral gas in the cometary comas and the solar wind minor heavy ions such as O^{5+} and C^{6+}. As a result of the charge transfer effect, a heavy ion such as O^{6+} will lose a positive charge to neutral molecules such as H_2O. The product ion, namely C^{5+}, will usually be excited to a higher electronic state. The radiative transition back to the ground state will lead to the emission of a photon of a few hundred eV. It is for this reason that the gas comas of comets have been found to be very powerful emitters of X-rays. The spatial distribution of the cometary X-ray emission is usually very symmetrical with respect to the sun–comet axis. It generally has a crescent-shaped morphology displaced toward the sunward side. This is simply because of the depletion of the solar wind ions as they stream toward the cometary center.

IX. ORIGIN

There are good dynamical reasons why the Oort cloud of new comets should have peak concentration between around 10^4 AU and 3×10^4 AU. First, with a stellar number density (n^*) of about 0. 1 pc^{-3} in the solar neighborhood and an average encounter velocity (V^*) of about 30 km s^{-3}, the cumulative velocity perturbation over the age of the solar system would be large enough to release comets located at $5–10 \times 10^4$ AU away from the sun. At the same time, stellar perturbations could maintain a steady influx of new comets only if their orbital distances are beyond 2×10^4 AU. Objects closer to the sun would be subject to stellar perturbations only rather infrequently. For example, over the age of the solar system (t_{sy}), the closest penetration distance (r) by a passing star can be approximated to be

$$\pi r_*^2 n_* v_* t_{sy} \approx 1 \tag{38}$$

Since $r_* \approx 10^3$ AU, comets with smaller orbital distances would be very stable against stellar perturbation. On the other hand, because of a lack of injection mechanism, these comets in the inner region would not be observable. It is thus possible that a rather massive inner Oort cloud may exist at solar distances between a few hundred and 1000 AU. Because particles in this system should be related to the condensation and accretion processes in the solar system without suffering from significant scattering effects by the passing stars, they should most likely be confined in a relatively flat configuration. Figure 37 illustrates the most up-to-date scenario according to current thinking. Between 10^3 AU and $2–3 \times 10^4$ AU, there will be sporadic scatterings from time to time due to stellar passages. Once this happens, a burst of comets will be injected into the solar system that would last a few million years. During this interval, the flux of new comets could be strongly enhanced in comparison to the steady-state injection rate from the outer Oort cloud. Such a transient phenomenon is called a comet shower.

Records of biological mass extinctions show that a certain periodicity (≈26 Myr) may be identified. Several lines of evidence (for example, the iridium enrichment at the Cretaceous–Tertiary boundary) have seemed to indicate that mass extinctions of this sort could be related to catastrophic impact events by either comets or asteroids of kilometer size. This possible connection is strengthened by the reported evidence of a periodicity in the ages of impact craters of similar value and in phase with biological extinction events.

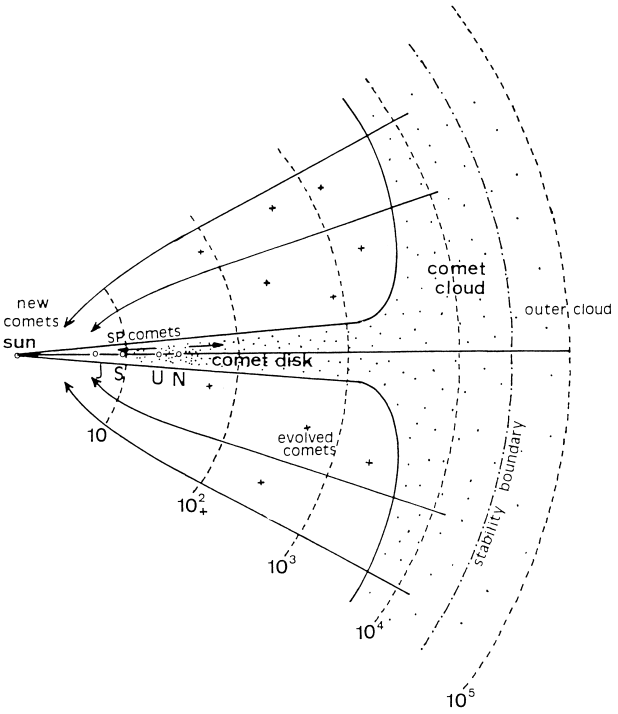


FIGURE 37 A schematic view of the structure of the cometary reservoir including the flat disk of inner Oort belt and the spherical shell of outer Oort cloud. [From Fernández, J. A., and Ip, W.-H. (1990) *in* "Comets in the Post-Halley Era" Newburn, R., Neugebauer, M., and Rahe, J., (eds.), Reidel, Dordrecht.]

Several interesting ideas have been proposed to explain the interconnection between the mass extinction/cratering events and cometary showers. These include (1) a solar companion star Nemesis with a 26-Myr orbital period, (2) periodic excursions of the solar system through the galactic plane, and (3) a trans-Neptunian planet X. The exact cause is under active investigation. In fact, whether the periodic crater events are statistically significant is still an open question. Furthermore, asteroid impacts might have contributed to the majority of these craters.

In addition to stellar perturbations, interstellar molecular clouds have been found to contribute to the destruction of the Oort cloud. The perturbation effect during the penetration passage by a giant molecular cloud with a total mass of 5×10^5 M and a radius of about 20 pc could be most dangerous to the survival of the Oort cloud. However, its occurrence is rather infrequent (approximately 1–10 encounters during the solar system lifetime). Molecular clouds of intermediate sizes with masses of a few 10^3–10^4 $M_\odot$ would be far more numerous and could produce rather strong perturbations comparable to the passing stars. The time interval between encounters of the solar system with intermediate molecular clouds has been estimated to be about 30 Myr.

One other important perturbation effect concerns the galactic tidal force. Because of the disk mass in the galaxy, the tidal force per unit mass, perpendicular to the galactic plane, acting on a comet as it moves to a galactic latitude ϕ is

$$f_{\text{tide}} = 4\pi G \rho_{\text{disk}} r \sin\phi \tag{39}$$

where ρ_{disk} is the mass density of the galactic disk. The change in the comet's angular momentum, which is proportional to $f_{\text{tide}} \cos\phi$, thus will have a latitudinal dependence of the form

$$\Delta L \propto \sin 2\phi. \tag{40}$$

Figure 38 shows the expected boundary of the Oort cloud as a result of the galactic tidal effect. The $\sin 2\phi$ dependence is clearly indicated. Statistical study of the frequency distribution of the galactic latitudes of new comets shows that there is indeed a minimum near the equatorial region with maxima at approximately $\pm 45^\circ$. There are also reports of clustering of the aphelion points of new comets in the galactic coordinate system. Such an effect may be the result of close encounters of passing stars in these angular positions.

The total mass and number of comets stored in the outer and inner Oort clouds are still very uncertain. On dynamical grounds, it is estimated that there should be approximately 10^{11}–10^{12} comets in the outer Oort cloud between 1×10^4 and 5×10^4 AU. As for the inner Oort cloud between 10^3 and 10^4 AU, a similar (or larger) number of comets could exist there. With an average mass of

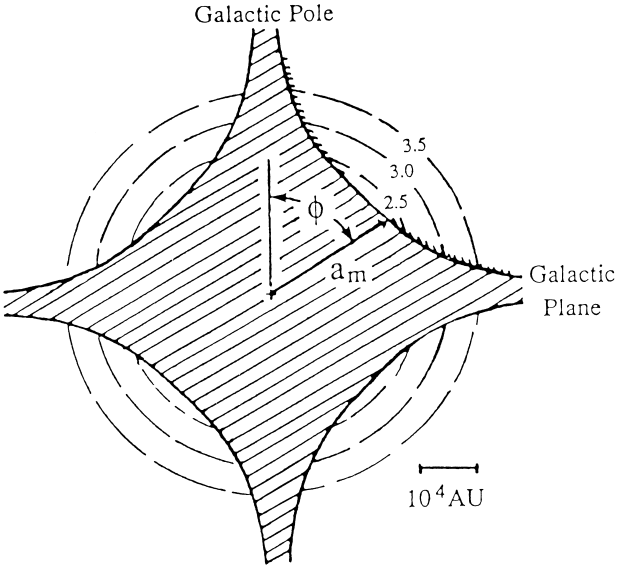


FIGURE 38 The latitudinal variation of the outer boundary of the Oort cloud as a result of galactic tidal effect. [From Morris, D. E., and Muller, R. A., (1986). *Icarus* **65,** 1.]

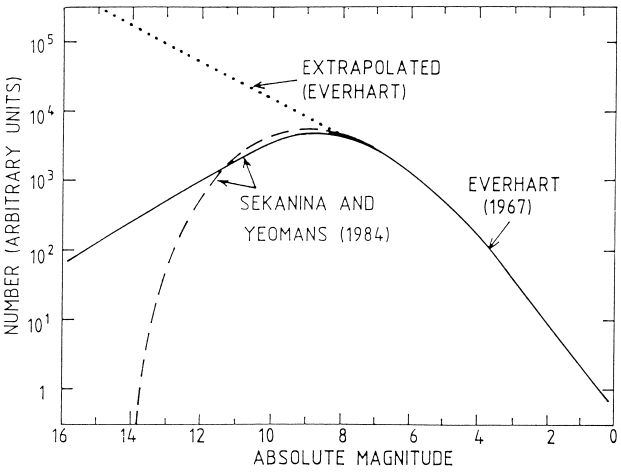


FIGURE 39 Distribution of the absolute brightness (H_{10}) of the long-period comets according to several statistical calculations. [From Everhart, E. (1967). *Astron. J.* **72,** 1002, and Sekanina, Z., and Yeomans, D. K. (1984). *Astron. J.* **89,** 154.]

10^{17} g for the comet nucleus, the total cometary mass (M_c) is thus on the order of 3–30 $M_\oplus$.

The actual values of M_c and number of comets (N_c) depend sensitively on the size distribution of comet nuclei. Figure 39 gives some estimates of the distribution of the absolute brightness distribution (H_{10}) of new comets. There is an apparent deficiency of comets with $H_{10} > 8$. This effect might be simply a matter of incomplete sampling, because comets in this range are relatively small. On the other hand, the cosmogonic process or physicochemical effects during the evolutionary history of the cometary nuclei could also contribute to the variations of the absolute brightness distribution. The determination of the H_{10} value of a comet depends on fitting its brightness variation as a function of power law in solar distance,

$$H = H_{10} + 5\log_{10}(\Delta) + 10\log_{10}(r) \qquad (41)$$

where r is the apparent magnitude of a comet when it is at heliocentric distance r and geocentric distance Δ. In Eq. (41) an inverse fourth-power dependence of the comet's brightness on r is assumed. Since the coma activities and surface properties of a comet could have strong effects on its brightness variation, it is by no means certain that Eq. (41) can be applied to all cases. Comprehensive surveys of distant comets at large solar distances (>3 AU), where the outgassing process generally ceases, are urgently needed to improve our knowledge in this aspect. The application of the following brightness–mass relation

$$\log_{10} m = -16.9 + 0.5H_{10}\,\text{kg} \qquad (42)$$

would mean that the cumulative number of the long-period comets has a power law dependence of $N(m) \propto m^{-1.16}$ for $m > 10^{14}$ kg. In comparison, the main-belt asteroids have a mass distribution given by $N(m) \propto m^{-1}$. Since the assumed brightness–mass relation has large uncertainties, it is difficult to attach too much significance to the similarity of the spectral indexes for the mass distributions of comets with $H_{10} < 6$ and of the main-belt asteroids. In any event, if the cometary population has gone through extensive collisional fragmentation in the solar nebula before ejection into the distant Oort region, a cumulative mass distribution of the form of $N(m) \propto m^{-1}$ should result.

Several possible routes exist for the formation of comets in the primordial solar nebula. For example, it has been proposed that comets could have condensed and agglomerated in fragments on the outskirts of a massive solar nebula. Or, interaction of strong outflows from the young sun with the neighboring dusty environment could have facilitated the formation of icy planetesimals in the outer solar system. Finally, for a solar nebula model with small mass (<0.1 $M_\oplus$), gravitational instability in the central dust layer could lead to the formation of icy planetesimals outside Neptune's orbit.

One other mechanism that has been quantitatively studied by numerical methods is the gravitational scattering of small planetesimals from the Uranus–Neptune accretion zone. According to model calculations, planetesimals with a total mass of a few $M_\oplus$ could be ejected into near-parabolic orbits at the end of the accretion process of Uranus and Neptune. Those with aphelion distances reaching beyond 10^4 AU would have their orbital inclinations isotropized; however, the ones with smaller aphelia would tend to avoid frequent stellar perturbations and thus maintain a flat disklike distribution (see Fig. 37).

The presence of a flat cometary belt received some support from computer simulation experiments in which the orbital evolution of comets under the influence of planetary perturbations were traced. It was found that the short-period comets are most likely supplied by a trans-Neptunian source populations of low orbital inclinations. In the case of an isotropic inclination distribution for the parent population (i.e., the long-period comets) a significant fraction of the short-period comets would be captured into retrograde orbits. Since this is not observed, the obvious conclusion is that the short-period comets must be supplied by a belt (or disk) of comets outside the orbit of Neptune. The required number of comets between 40 and 200 AU is on the order of 10^8. Under steady-state condition of orbital diffusion, there should be about 10^6 such comets orbiting between Saturn and Uranus before they are captured into short-period comets.

A very interesting development from the point of view of cometary origin has to do with the discovery of Kuiper belt objects (KBOs) in trans-Neptunian orbits by David Jewitt and Jane Luu in 1992. These objects, presumably of icy composition, have sizes of a few hundred kilometers.

The total number of the KBOs with diameters >100 km has been estimated to be on the order of 10^5. Their orbits are located mostly beyond Neptune's orbit and, just like Pluto, a significant fraction of the discovered population is trapped in the 2:3 orbital resonance with Neptune. A number of the KBOs have been found to be in orbits of large semimajor axis (∼100 AU). The observational statistics suggested that the total number ($\sim 3 \times 10^4$) of such scattered KBOs with diameters >100 km at the present time could be comparable to the classical KBOs. The original population of the KBOs could be as much as a factor of 100 bigger than the present one. A combination of planteary gravitational perturbations and collision destruction served to reduce the number of KBOs to the present value. It is most likely that the small collisional fragments of the KBOs could serve as an important source population of short-period comets in the past as well as the present time.

X. PROSPECTS

After the preliminary studies of comets Giacobini–Zinner and Halley via flyby reconnaissance, the future impetus of cometary research will be set by long-term investigations afforded only by rendezvous missions and, as a final goal, the return of samples of comet material in their pristine condition to Earth for laboratory analyses. Because of the potential threats of the catastrophic collisional effects of near-Earth objects (NEOs), there is also considerable interest in studying the physical properties of cometary nuclei. It is for these reasons that a number of comet missions have been planned and will be realized in the near future. These include NASA's Stardust mission for coma dust sample return, the CONTOUR mission for multiple comet flyby observations, and the Deep Impact project for active cratering experiment on a target comet. The Deep Space 1 spacecraft employing ion drive propulsion system will encounter Comet Borrelly in 2002. The title of the most daring and comprehensive comet mission belongs to the Rosetta mission of the European Space Agency. The Rosetta spacecraft will be launched in 2003 with Comet Wirtanen as its final target. After rendezvous, the spacecraft orbiting around the cometary nucleus will deploy a small spaceship for landing on the surface of the nucleus. The lander will provide *in-situ* measurements of the chemical and material properties of the cometary ice and dust mantle in great detail. Last but not least, the interior of the cometary nculeus will be probed by the radar tomography experiment on the mother spaceship.

The foregoing new initiatives, which will be among the major thrusts in planetary research, are expected to produce scientific data in the next decade. The outlook for cometary study is therefore extremely promising. Certainly even more ambitious projects might lie ahead at the beginning of the 21st century. At this point, we should mention the potential industrial and commercial interests in cometary materials as important resources in space.

The important contributions from the International Halley Watch network during the comet Halley epoch and more recently the worldwide observations of comet Hyakutake and comet Hale–Bopp in 1996 and 1997 have shown that coordinated ground-based observations are an indispensable part of cometary research. Besides optical observations, infrared and UV measurements using satellite-borne telescopes would be very informative in monitoring the coma activities, dust emission, and gas and dust compositions of a wide variety of comets at different stages of outgassing. For this, a dedicated planetary telescope capable of performing simultaneous imaging and spectroscopic observations in the wavelength range from ultraviolet to infrared would be most desirable. For example, there are many issues related to the thermodynamics and hydrodynamics of coma expansion that require IR observations at high spatial and spectral resolution. The extremely variable structures of cometary ion tails are closely related to momentum and energy-transfer from the solar wind to cometary ions; well-planned observations capable of measuring the ion densitis, plasma velocities, and spatial configurations such as ion rays and plasmoids are urgently needed. The results from comet Halley have shown that the effects involved are extremely complicated. The correlations of remote-sensing observations and *in-situ* measurements carried out in the missions just mentioned will be most useful from this point of view.

The number density and size distribution of comets in the outer solar system are critical to our understanding of the origin of the short-period comets and the distant Oort cloud. Surveys by a new generation of space telescopes in optical and IR wavelengths will possibly provide the definitive answer. The observations of extrasolar systems exemplified by Beta Pictoris will yield complementary knowledge on the formation of the solar nebula and its disk of planetesimal population.

Laboratory simulations of ice condensation and gas trapping at low temperatures have provided very interesting insights into the physical nature of the cometary ice and the surface sublimation processes. Further progress in this area is expected.

Theoretical studies, as always, are the basic tools in laying the foundation for understanding different physicochemical phenomena and their interrelations. The five areas of ongoing theoretical efforts (i.e., dust mantle modeling, gas thermodynamics and hydrodynamics, comet–solar-wind interaction, coma chemistry, and cometary

orbital dynamics) are following a trend of increasing sophistication. In combination with the opportunities of new observations and space missions in the next decade, modern cometary physics has just reached a new era.

ACKNOWLEDGMENT

This work was supported in part by NSC 90-2111-M-008-002 and NSC 89-2112-M-008-048.

SEE ALSO THE FOLLOWING ARTICLES

PRIMITIVE SOLAR SYSTEM OBJECTS: ASTEROIDS AND COMETS • SOLAR SYSTEM, GENERAL • SOLAR PHYSICS • STELLAR STRUCTURE AND EVOLUTION

BIBLIOGRAPHY

Brandt, J. C., and Chapman, R. D. (1981). "Introduction to Comets," Cambridge Univ. Press, Cambridge U.K.

"Comet Halley 1986, World-Wide Investigations, Results, and Interpretation," (1990). Ellis Horwood, Chichester, U.K.

Comet Halley Special Issue (1987). *Astron. Astrophys.* **187**(1, 2).

Crovisier, J., and Encrenaz, Th. (1999) "Comet Science," Cambridge Univ. Press, Cambridge, U.K.

Devine N., Fechtig, H., Gombosi, T. I., Hanner, M. S., Keller, H. U., Larson, M. S., Mendis, D. A., Newburn, R. L., Reinhard, R., Sekanina, Z., and Yeomans, D. K. (1986). *Space Sci. Rev.* **43,** 1.

Huebner, W. F., ed. (1990). "Physics and Chemistry of Comets in the Space Age," Springer-Verlag, Berlin.

Krishna Swamy, K. S. (1997). "Physics of Comets," 2nd ed., World Scientific, Singapore.

Mendis, D. A., Houpis, H. L. F., and Marconi, M. L. (1985). "The physics of comets," *Fundam. Cosmic Phys.* **10,** 1–380.

Newburn, R. L., Neugebauer, M., and Rahe, J. (eds.). (1991). "Comets in the Post-Halley Era," Kluwer Academic Publishers, Dordrecht.

Impact Cratering

William B. McKinnon
Washington University

GLOSSARY

Breccia lens Layer of broken and pulverized rock (breccia) that partially fills simple craters.

Catastrophic disruption Impact so energetic that it breaks the target body apart instead of forming a crater.

Complex crater Impact crater of relatively large dimension and shallow depth, characterized by a central uplift and peripheral slumping or terracing that develops during the modification stage as a result of yielding of the underlying rocks (or ice).

Coupling parameter Single scalar measure of the impactor that determines the characteristics of the cratering flow field.

Ejecta Material ejected from a crater during its formation.

Excavation stage Stage in the cratering process, following initial shock compression of the impactor and target, in which the crater bowl forms.

Hugoniot Locus of points describing the pressure-volume-energy relations or states that may be achieved within a material by shocking it from a given initial state.

Impactor Cosmic object that strikes a plane or satellite surface.

Late heavy bombardment Early period of heavy cratering expressed on the most ancient (3.8–4.5-Gyr-old) surfaces of the moon, Mercury, and Mars.

Modification stage Stage in the cratering process for large craters that promptly follows the excavation stage and results in such morphologies as flat floors, slump terraces, central peaks, peak rings, or multiple rings.

Multiringed basin Large circular depression, usually hundreds of kilometers in diameter, with one or more zones of concentrically arranged mountains caused by large-body impact.

Regolith Layer of fragmentary debris produced by meteoritic impact on the surface of any celestial body.

Secondary crater Impact crater produced by a fragment or fragments ejected from a larger primary crater.

Simple crater Impact crater of relatively small diameter characterized by a uniformly concave-upward shape and maximum depth in the center and lacking a central uplift.

Transient crater Crater formed at the end of the

excavation stage, which if the crater is large enough, promptly collapses to a complex form.

IMPACT CRATERS form when a smaller object strikes a larger one at hypervelocity. The resulting crater has a simple bowl shape at small physical scales, but several more complex morphologies can arise at larger physical scales. Impact craters are a dominant landform of the solar system, identified on every solid body studied to date except Jupiter's moon Io. The effects of cratering have been profound, if only because all solid planets and satellites formed by the collision and accretion of numerous smaller bodies. Space exploration has documented the fundamental role of impact cratering in the geological and atmospheric evolution of earth and the other planets, and especially in the biological evolution of earth. The level of activity in the field of impact studies continues to increase. The recognition of impact cratering as a fundamental geologic process, particularly when contrasted with the impact-versus-volcanic origin debates in the 1950s, represents a revolution in geologic thought.

I. INTRODUCTION

Impact craters are relatively rare on earth, so debate originally centered on the origin of lunar craters observed through the telescope. The lack of *in situ* examination coupled with a naturally dim understanding of planetary surface and evolutionary processes allowed many dubious scenarios to be entertained. The few relatively well-preserved impact structures on earth were considered to be volcanic, but because there was no evidence for lava or other volcanic materials, they were termed cryptovolcanic. Even G. K. Gilbert, a famous geologist who advocated an impact origin for lunar craters and found abundant meteoritic iron at Meteor Crater, Arizona (both in 1891), concluded that Meteor crater was volcanic in origin.

Gilbert was led astray by an incomplete understanding of cratering mechanics. This situation began to change after World War II, when analogies to large explosion craters were developed, notably by R. B. Baldwin and E. M. Shoemaker. Their arguments helped convince early planetary scientists H. C. Urey and G. P. Kuiper and others that lunar craters were created by impact. Geologic evidence accumulated supporting the impact hypothesis, such as the discovery of shocked polymorphs of quartz and other minerals at crater sites, but it was not until the early 1960s, with the return of spacecraft images of the moon showing an abundance of craters at every scale, that the volcanic hypothesis was laid to rest. Even at this time the appreciation of impact cratering as a general phenomenon was incomplete. When Mariner 4 returned images of the cratered highlands of Mars in 1965, researchers were surprised. In contrast, cratered terrains were expected in images of the satellites of the outer planets returned by Voyagers 1 and 2 from 1979 to 1989. Analysis of crater morphologies, spatial distributions, and size-frequency distributions is now a major tool for understanding the geologic histories of planetary bodies, including the earth.

II. MECHANICS

Excavation of an impact crater form begins when a smaller object, the impactor, strikes a larger target, initiating a rapid (but orderly) series of events. Momentum and energy are transferred to the target, and a shock wave is set up which propagates through the target. This shock wave irreversibly deposits internal energy and kinetic energy, which results in the heating, melting, and vaporization of the impactor as well as target material near the impact point and the creation of a flow field in ruptured target material. Expansion of the flow field creates the crater.

The pressure P, density ρ, shock velocity U, particle velocity u, and internal energy E (per unit mass) of the shock-compressed material are determined by the Rankine-Hugoniot equations, which express the conservation of mass, momentum, and energy, respectively, across the shock front:

$$\rho_0 U = \rho(U - u) \tag{1a}$$

$$P - P_0 = \rho_0 U u \tag{1b}$$

$$E - E_0 = \frac{P + P_0}{2}\left(\frac{1}{\rho_0} - \frac{1}{\rho}\right) = \frac{1}{2}u^2 \tag{1c}$$

P_0, ρ_0, and E_0 are the unshocked values, and these equations apply in the frame of reference where unshocked material is at rest. Shock and particle velocities are empirically (and usually linearly) related for a given material and must be measured. Such a $U - u$ curve is one example of a shock Hugoniot. Typical impact velocities in the solar system are on the order of 10 km s^{-1}, which lead to initial shock pressures in the range of 10^{2-3} GPa (1–10 Mbar). Equation (1) plus Hugoniot data, when coupled to the hydrodynamic equations of motion and appropriate boundary conditions (at the impactor-target interface and the impactor and target free surface), allow a complete description of the initial stages of the cratering process.

Impact crater formation is conventionally divided into stages. The first, compression, begins with the initial contact and ends when the resulting shock wave reaches the back of the impactor (see Fig. 1). Unloading also begins immediately, at the free surface near the impactor-target interface. An unloading or rarefaction wave propagates

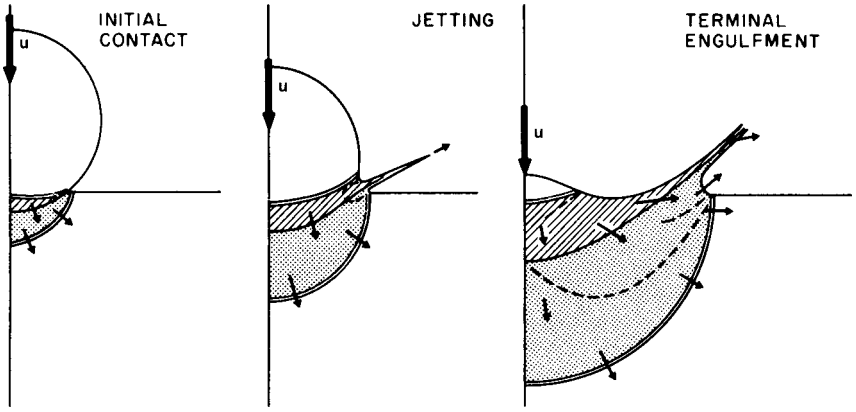


FIGURE 1 The compression stage of impact mechanics. The left panel shows the shock wave geometry just after initial contact with a spherical impactor moving at velocity *u*. Shocked impactor and surface (target) material is lined and stippled, respectively. Arrows indicate particle velocities. Rarefaction waves (dashed lines) unload the shocked zone, causing jetting. Jetting segues into the beginning of formation of the ejecta curtain. [From Chapman, C. R., and McKinnon, W. B. (1986). Cratering of planetary satellites. *In* "Satellites" (J. A. Burns and M. S. Matthews, eds.), pp. 492–580. Univ. of Arizona Press, Tucson, Arizona.]

inward, and the material undergoing rarefaction is subjected to a high pressure gradient and accelerated upward and outward in a jet of what is usually melt or vapor.

Momentum and energy transfer are completed during the first part of the excavation stage. Here, the shock wave continues to propagate into the target, but the rarefaction front from the free surface of the target (and impactor) eventually catches up and combines with the original hemispherical shock in the target. The resulting geometry resembles a detached shock. The overall effect is to rotate the velocity vectors of the flow field, from basically radially outward from the point of initial contact, toward the original target surface (Fig. 2). The pressures behind the detached shock are relatively high, and the original mechanical integrity of the target material is destroyed. Material close to the impact point may be vaporized or melted, but most target material is thoroughly fractured and conminuted. The flow field created is approximately steady-state and incompressible, and the crater bowl, or transient cavity, opens up as the flow field expands. Expansion of crater dimensions is a power-law function of time and is accomplished by both ballistic ejection and displacement (Fig. 2).

The detached shock expands and dissipates. The flow field expands hydrodynamically until arrested by some combination of gravity, strength, or viscosity. Except for small craters in competent materials, the ultimate limitation on crater size and shape is the dissipation of flow field kinetic energy, either frictionally or viscously, or its conversion to gravitational potential energy.

Crater dimensions at the end of the excavation stage are typically one to two orders of magnitude greater than that of the original impactor. Excavation lasts orders of magnitude longer than compression and scales as the square root of the diameter. Thus, a 1-km-diameter crater may form in ~10 sec (on the earth), and a 100-km-diameter crater in ~100 sec.

The crater at the end of the excavation stage is termed the transient crater. It may also be the final one, but for craters larger than a threshold diameter (which varies from planet to planet), the transient crater is not in mechanical equilibrium and promptly collapses under the influence of gravity. This modification stage results in the creation of many different morphological features, which are discussed later. For larger craters, the end of the excavation stage and the modification stage are one continuous process, because both phases are gravity dominated.

III. SCALING

Early scaling relationships derived from explosion crater studies assumed that the final diameter depends solely on impactor energy. Energy and momentum are both important in impact cratering, however, as might seem obvious. The coupling of kinetic energy and momentum into a restricted region near the impactor, nevertheless, allows the impactor to be characterized by a single measure or variable, the point-source coupling parameter. The coupling parameter is intermediate in dimensionality between energy and momentum. Its form must be determined experimentally for each material of interest. For nonporous materials, which are appropriate in most planetary

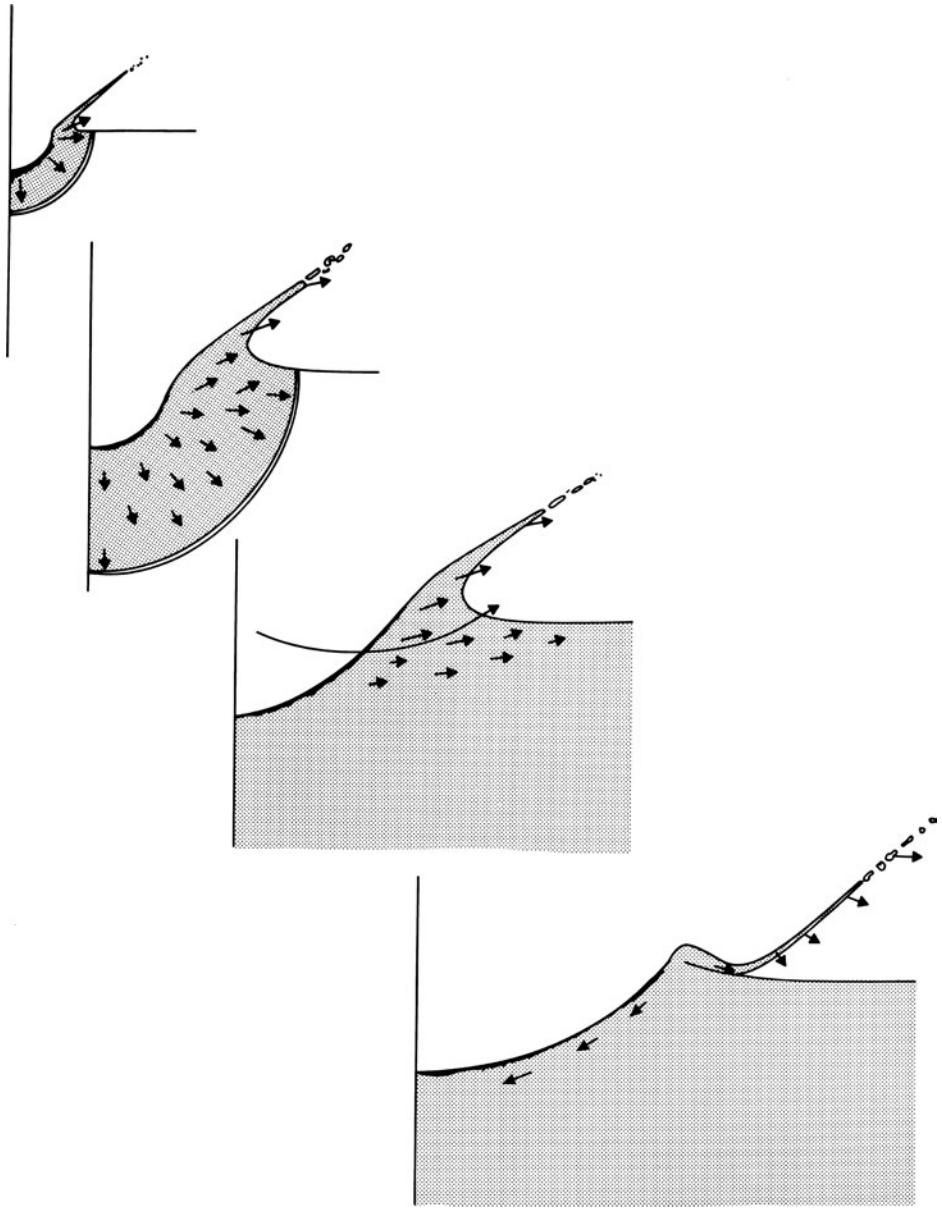


FIGURE 2 Four instantaneous views of the excavation stage. The shock front and flow field velocity vectors are indicated by double lines and arrows, respectively. In this example, impactor remains are indicated by the dark lining of the crater surface. Growth of the transient cavity is nearly hemispherical with material in roughly the upper one-third of the crater (above the streamline illustrated in the third panel) being ejected. Eventually maximum dynamic depth is reached, but lateral expansion (shearing along the crater wall) continues for craters in nonliquids. Stagnation of the entire crater flow results in the classic bowl shape, raised rim, and overturned flap of ejecta. Unstable material lining the crater slides inward to form the breccia lens.

situations, the coupling parameter has dimensions closer to energy.

A convenient way to express crater scaling is through dimensionless variables, in particular, the gravity-scaled size

$$\bullet_2 = 3.22ga/u^2 \qquad (2)$$

where a is the spherical equivalent radius of the impactor, u is its velocity, and g is the gravitational acceleration. The best estimate for the scaling of large crater formation in solid rock (or ice) is

$$V\rho/m = 0.2\,\bullet_2^{-0.65} \qquad (3)$$

where vertical incidence is assumed, and V is the excavated volume below the original ground plane, ρ is the target density, and m is the impactor mass. In terms of transient crater radius R,

$$R(\rho/m)^{1/3} = 0.8\,\bullet_2^{-0.22} \quad (4)$$

As an example, a 1-km-diameter impactor striking the earth at 25 km sec^{-1} (a typical collision velocity for an asteroid) creates a 34-km-diameter crater. Formation time is close to $0.8V^{1/6}g^{-1/2}$. Smaller craters can be affected by the strength, or cohesion, of the target, but craters on the terrestrial planets and icy satellites greater than ~1 km in diameter should be dominantly affected by gravity.

The volume of target melted and vaporized, V_m, depends on conditions during the compression stage, and appears to scale with impact energy alone. Based on numerical models and observations of melt volumes at terrestrial craters in hard rock, $V_m/m \sim 7.5\ (u/20$ km sec$^{-1})^2$ appears to be a good rule of thumb as long as $u > 12$ km sec^{-1}. Because of this scaling, the ratio V_m/V is not constant, but rather, an increasing function of size and gravity.

IV. OBSERVATIONS

Craters, whether in rock or ice, or on planets and satellites of greatly varying gravity, form similar size-dependent morphological sequences. The smallest craters are called simple craters because they have simple bowl shapes. The depth-to-diameter-ratio for such craters, when freshly formed in rock, is always close to 0.2. Those formed in ice (as on the satellites of the giant planets) are somewhat shallower. They all exhibit raised rims built of deformed rock (or ice) units overlaid with ejecta. The trajectories of the ejecta cause the stratigraphy of the target to be emplaced in an inverted sequence at the rim, an occurrence sometimes called an overturned flap. Away from the rim crest, ejecta is laid down more violently and is more thoroughly mixed. The ejecta forms a blanket that thins away from the crater, becoming discontinuous at a distance of about one crater radius. Along with discontinuous ejecta, individual and groups of secondary craters are distributed even further out. Secondary craters form from more coherent individual blocks and fragments that are ejected from the growing crater (probably from the near-surface region). The distance they travel depends on the target body's surface gravity, and although secondary craters are seen on the larger icy satellites, they are generally absent on small satellites. The most distant ejecta is distributed as bright (or sometimes dark, depending on the composition of the terrain) albedo streaks or rays.

The floors of simple craters are covered with a breccia or rubble lens that is the accumulation of material unable to escape from the crater during the final stages of lateral expansion (Fig. 2). It overlies a thinner more heavily shocked unit that lines the crater floor. Terrestrial studies show that this unit includes solidified shock-melted rock (and similar units can be imagined for craters formed on the icy satellites). For terrestrial craters, shock effects are often the principal determinant of a crater's impact status. These effects range from shattercones in homogeneous rocks (implying shock pressures of 5–10 GPa), through planar deformation features in quartz and feldspar (~10–30 GPa), to conversion of quartz and feldspar to higher pressure polymorphs such as coesite, stishovite, and maskelynite (the latter a so-called diaplectic glass) at ~30–40 GPa, and melting at still higher pressures (~60 GPa). The presence of impact-melted rock becomes more pronounced for the larger complex craters. These craters have undergone modification in response to gravity, have lower depth-to-diameter ratios than simple craters, and generally have more complex morphologies (see Fig. 3, for example).

For each planet and satellite there is a distinct size above which craters are complex and below which they are simple (for example, about 3-km and 15-km diameter for the earth and moon, respectively). This transition size is inversely correlated with surface gravity and is lower for icy bodies than for rocky bodies. The smallest complex craters have flat floors. Somewhat larger complex craters show slump blocks along their rims, sometimes organized into wall terraces. An uplift in the central region of the crater, corresponding to the downward and inward slumping movement, produces central peaks (Fig. 4). At still larger scales, the central peaks themselves are apparently not stable and collapse into a mountain ring form (peak rings, see Fig. 3), or on the larger icy satellites, into what are termed central pits.

For the very largest impacts, the response of a planet (or satellite) in the modification stage is widespread. The rigid outer shell, or lithosphere, of the planet is fractured by a coherent system of faults that encircle the collapsing crater but at large distances beyond the rim. Displacements on these faults give rise to great circumferential mountain ranges and valleys. These impact structures are the multiringed basins, the most magnificent crater forms of all. They occur on only the largest bodies: the moon, Mercury, Mars, Venus, and the icy Galilean satellites Europa, Ganymede, and Callisto. They are generally, but not exclusively, created by collision of the largest asteroids or comets (on the order of 100 km in diameter) and can be over a thousand kilometers in extent themselves. The largest involve entire hemispheres of their respective worlds, notably on Ganymede and Callisto, whereas more compact examples can be found on Europa and Venus.

Impact craters, once formed, are subject to further modification. On airless worlds, they can be eroded by further

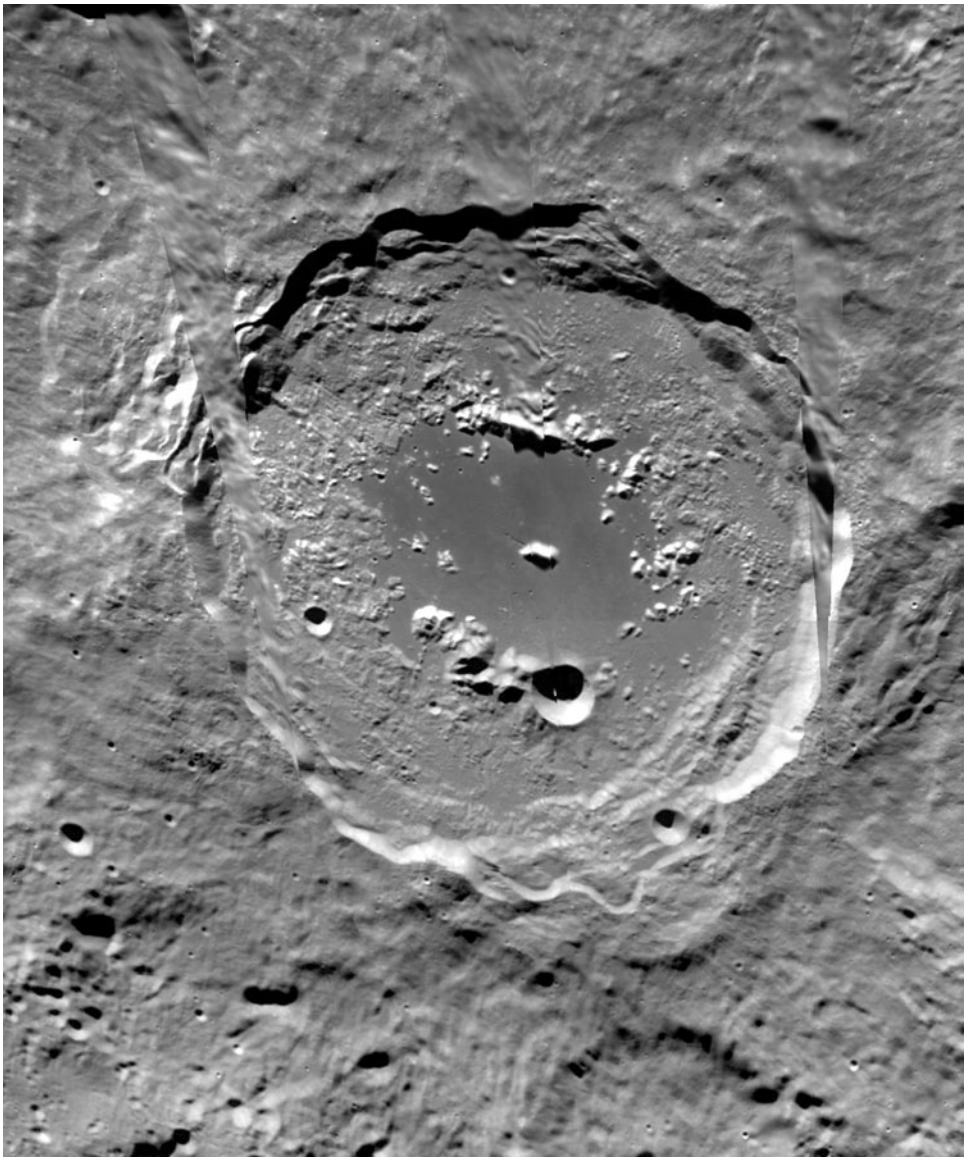


FIGURE 3 Clementine mosaic of the lunar farside crater Antoniadi (~135 km in diameter and several kilometers deep). Antoniadi is transitional in form, possessing both a central peak and a peak ring. Slump blocks and terraces and a lineated ejecta blanket can be seen. The floor outside the peak ring is underlain by impact melt and highly shocked rubble, and the central floor has been subsequently flooded with mare basalt.

bombardment; indeed, a perpetual rain of small impactors churns up such surfaces, creating a fragmental soil layer or regolith. Impact craters can also be eroded and/or buried by the action of wind and water, particularly on the earth and Mars, or buried by volcanic units or broken up by faults. On large, icy satellites, their topography may be reduced by slow viscous creep of the surface. It is characteristic of the larger craters on the moon, Mars, and Mercury to act as preferential sites for the later eruption of basaltic lavas. On the moon these form the well-known maria, or seas.

Striking, bilaterally symmetric ejecta patterns result as the impact angle (with respect to the vertical) is increased, and the crater itself may become elliptical in outline if the impact angle is within a few degrees of horizontal. Martian ejecta blankets often indicate extensive ground-hugging flow away from the crater, most likely due to groundwater or melted ice in the ejecta. Extensive, runout ejecta flows exist on Venus, thought to be due to entrained atmosphere or an enhanced fraction of impact melt in the ejecta (due to the high surface temperatures of Venusian rock and the scaling effect discussed previously).

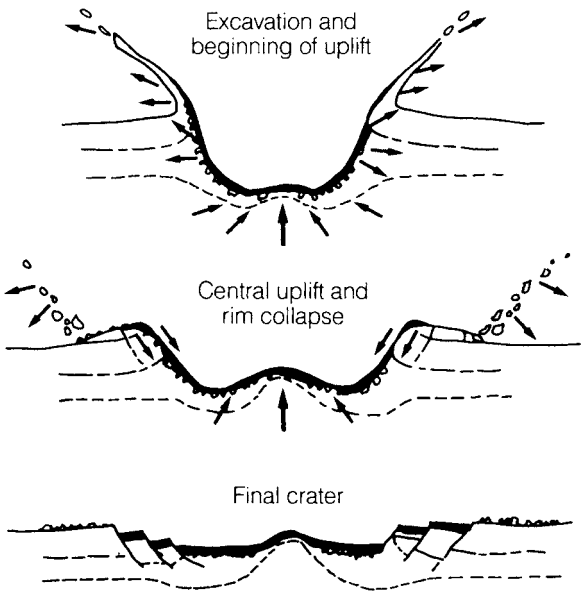


FIGURE 4 Illustration of the formation of a complex crater. Uplift of the crater floor begins even before the rim is fully formed. As the floor rises, collapse of the rim creates a wreath of slump terraces surrounding the crater. In larger craters, the central peak itself collapses and creates a peak ring. [From Melosh, H. J. (1989). "Impact Cratering: A Geologic Process," Oxford Univ. Press, New York.]

Parabolic, west-opening radar-dark deposits are also associated with the youngest craters on Venus, caused by impact debris lofted into the planet's high-altitude westward zonal winds. Sufficiently large impacts on Europa apparently penetrate through surface ice to the ocean below, as no central crater forms, just a flat to irregular circular feature (or crater palimpsest) surrounded by multiple rings. Conversely, smaller impactors are not able to penetrate the thick Venusian atmosphere intact, and so form irregular crater clusters or radar-dark, shock-pulverized splotches.

V. THE SOLAR SYSTEM CRATERING RECORD

Understanding the distribution of impact craters in space and time remains one of the most challenging tasks of cratering science. The distribution of craters on the earth (Fig. 5) is very nonuniform. This is partly a reflection of more detailed exploration in the northern hemisphere, but it also reflects a concentration of impact craters on the oldest and most geologically stable regions of North America and Eurasia. There are always many more small impact craters than large ones, so only a few, deeply eroded and/or buried, large structures survive on the earth.

The level of geological activity on the moon, compared to the earth, is very low, and an enormous number of impacts survive. When the number of craters per square kilometer, or crater density, is plotted against surface age as dated by traditional radioactive isotope methods, an astonishing fact emerges (Fig. 6). Most of the moon's craters formed a long time ago, and nearly all formed in the first 700 Myr or so of solar system history. This torrential rain of debris is called the late heavy bombardment, and can be thought of as the final sweep up of the material that originally built the planets, but whether this bombardment was

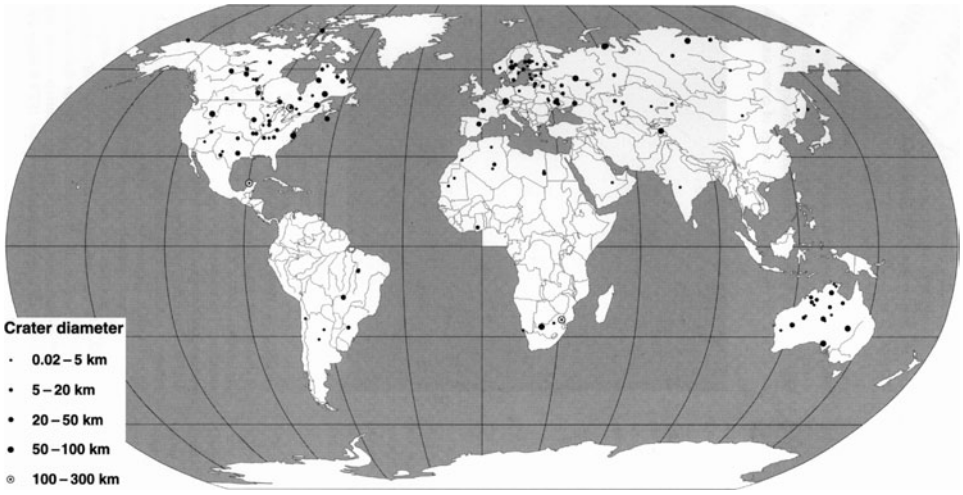


FIGURE 5 Areal distribution of the ~150 known impact structures on the earth. [Adapted from French, B. M. (1998). "Traces of Catastrophe: A Handbook of Shock-Metamorphic Effects in Terrestrial Meteorite Impact Structures, LPI Contribution No. 954, Lunar and Planetary Institute, Houston.]

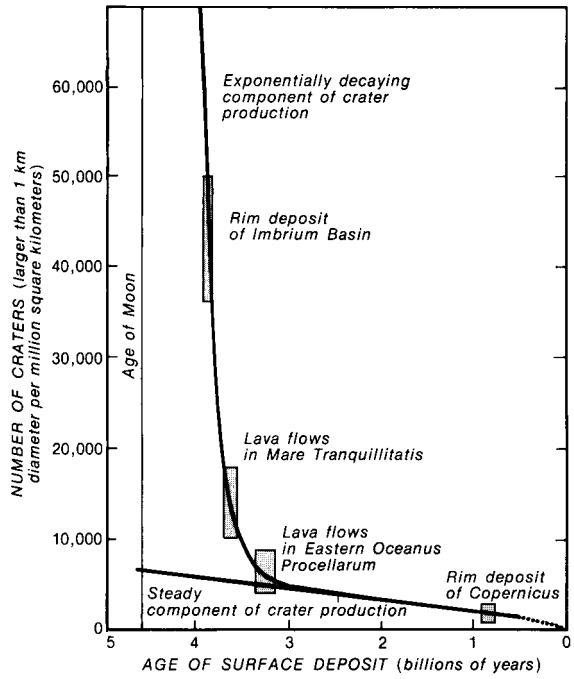


FIGURE 6 The variation of crater density on lunar surfaces of different ages. Widths of the small rectangles, which correspond to Apollo landing sites, indicate the uncertainty in age of each radiometrically dated surface. The heights indicate the statistical uncertainty in crater density. The high cratering rate of the "late heavy bombardment" dropped rapidly between 3.9 and 3.2 Gyr ago, giving way to a lower, more steady rate of crater production. [Adapted from Shoemaker, E. M., and Shoemaker, C. S. (1999). *In* "The New Solar System" (J. K. Beatty, C. C. Petersen, and A. Chaikin, eds.) 4th edition, pp. 69–86, Sky Publishing Corporation, Cambridge, Massachusetts.]

the tail end of a continuously declining impact flux or a specific event, or cataclysm, is not known . Curiously enough, no rock units from earlier than 3.96 Gyr ago survive on earth.

Only a few places on the moon have been visited and sampled, so the information in Fig. 6 can be turned around and used to date other portions of the lunar surface through their impact crater densities. The probability of having a crater form anywhere on the moon is uniform, so terrains of equal age should have equal crater densities. Using impact craters as a chronometer, the stratigraphic column of the moon has been worked out.

It would be marvelous to be able to extend the moon's impact timetable to other solar system bodies, that is, to establish an interplanetary correlation of geologic time. To do this, it must be shown that the same population of impacting objects has struck each body. It is thus necessary to look at the detailed characteristics of the impact flux. The distribution of crater sizes is strongly weighted toward small craters, as mentioned previously. The cumulative number of craters N greater than diameter D, per unit area, generally follows a steep power-law function close to

$$N(>D) \propto D^{-2} \tag{5}$$

The number of craters in each logarithmic size bin, or the crater frequency, goes as D^{-3}.

Figure 7 shows the crater population frequencies for various solar system bodies divided by a D^{-3} power-law. This has the effect of flattening all the curves (removing their steep slopes) and emphasizing their differences. These normalized curves also indicate the relative proportion of a surface covered by craters of different sizes. The well-determined curves for the heavily cratered terrains on the moon, Mars, and Mercury all resemble each other. Hence, the late heavy bombardment occurred on all the planets of the inner solar system.

The case for the outer solar system is more controversial. The largest icy satellites, Ganymede and Callisto, are missing large craters (>100 km in diameter) compared with the moon. This may either mean that large craters are preferentially erased on these satellites (perhaps by viscous relaxation) or that a different population of impactors was dominant in the Jupiter region and beyond. This different population is likely to be comets, in contrast with asteroids in the inner solar system, and in recent epochs predominantly Kuiper Belt comets at that. High resolution Galileo images also show a dearth of small (100 m) craters, possibly indicating a lack of small comets or that small comets disintegrate as their orbits evolve in toward the sun. Some of Saturn's small satellites show crater populations that are strongly peaked in the 10-km-diameter range. These may indicate a population of secondary debris generated within the Saturn system, perhaps by the catastrophic disruption of a small satellite.

The case for Venus is particularly interesting. There, some 950 individual craters can be identified, mostly greater than 30 km in diameter, which attests to the efficiency of the atmospheric shield. More or less randomly distributed, these craters indicate a close to global resurfacing event circa 500 Myr ago, and only modest geological activity since then. At the other end of the spectrum, even the highest resolution Galileo images of Io reveal no craters whatsoever, implying that this hyperactive satellite's surface is no more than ∼1 Myr old.

VI. THE IMPACT CRATER REVOLUTION

Impacts are now felt to play a major role in nearly every aspect of solar system evolution. The following is an incomplete list of recent discoveries or active research areas.

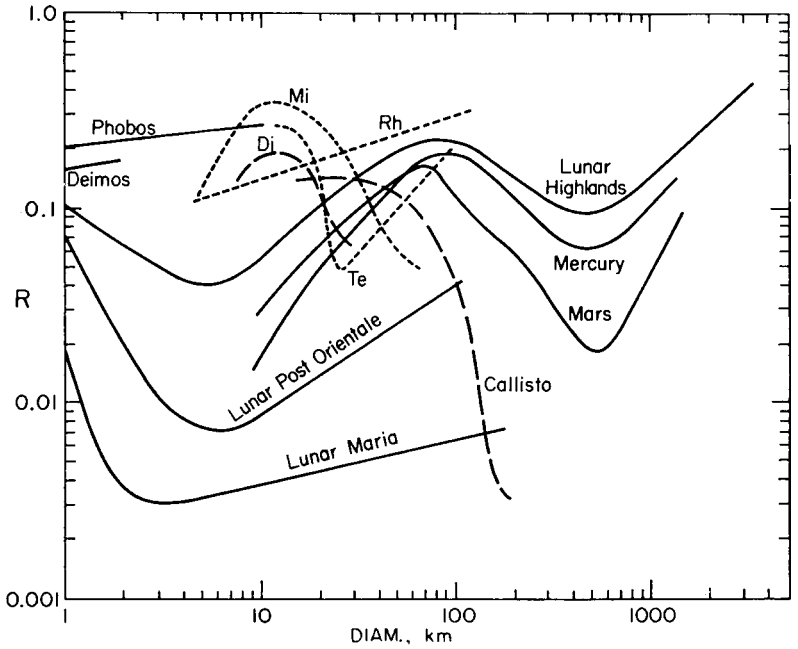


FIGURE 7 Characteristic relative-plot diameter-frequency relations for various satellites and terrestrial planets. These are approximate average relations, generally for heavily cratered terrains. Phobos and Deimos are the moons of Mars, Callisto orbits Jupiter, and Mi, Rh, Di, and Te identify Mimas, Rhea, Dione, and Tethys, respectively (all satellites of Saturn). The curves for the older terrains on the Uranian satellites resemble that of Rhea. [From Chapman, C. R., and McKinnon, W. B. (1986). Cratering of planetary satellites. *In* "Satellites" (J. A. Burns and M. S. Matthews, eds.), pp. 492–580. Univ. of Arizona Press, Tucson, Arizona.]

1. Collisional accretion is thought to be the mechanism of formation for solid planets and satellites and for the cores of the giant planets (the latter are thought necessary for giant planet formation).
2. The impact of a Mars-sized object early in the earth's history may have driven enough material of the right composition into permanent orbit to form the moon.
3. The impact of a very large object with a differentiated proto-Mercury may have stripped it of its silicate mantle, leaving it with an anomalously large iron core.
4. Collision of an earth-sized object with Uranus may have tipped its rotation axis substantially and spun out a disk of material from which its satellites formed.
5. Triton may have been captured from solar orbit by Neptune when Triton collided with an original Neptune satellite.
6. A proto-Pluto could have been collisionally disrupted by collision with a similar object, leading to formation of the Pluto-Charon binary pair.
7. Unreaccreted debris from collisions in the asteriod belt may be the cause of asteroid satellites. The binary asteroid Antiope, in particular, may be the result of an off-center collision by two more-or-less equal-sized precursors.
8. Small satellites whose orbits have evolved inside the Roche limit of a giant planet may be catastrophically disrupted by cometary impact, creating spectacular ring systems such as we see today around Saturn and Uranus.
9. Impacts with planetary and asteroid surfaces have launched meteorites and lunar samples to the earth. Certain (mostly young) meteorites have been deduced to come from Mars.
10. Early comet and carbonaceous asteroid bombardment may have supplied the earth with its inventory of water.
11. Asteroid and comet impact may have partially stripped away the early atmospheres of the terrestrial planets; in particular, Mars' atmosphere may have been much more massive and allowed liquid water on its surface.
12. Comet impacts may have stripped off early atmospheres of Europa, Ganymede, and Callisto, but left Titan's atmosphere intact due to lower collision speeds.
13. Early cometary bombardment may have supplied the earth with its initial inventory of organic molecules, thereby setting the stage for the origin of life.
14. Early giant impacts on the earth may have frustrated the origin of life, such as by boiling the oceans and sterilizing the surface.

15. Microbes may be launched in ejecta from the earth, potentially seeding other worlds of the solar system such as Mars. If life arose independently on a clement Mars, such a process could have worked in reverse.
16. Comet Shoemaker-Levy 9 was temporarily captured by Jupiter, tidally fragmented during a close pass in 1992, and telescopically discovered in 1993, and over a dozen individual, phenomenal impacts into Jupiter's atmosphere were observed in 1994. The fireballs and resulting dark markings represented the most striking change in the planet's appearance since the invention of the telescope.
17. Beds of glassy spherules have been discovered in Archean rocks in South Africa and Australia and interpreted as the distal ejecta of large, ancient craters (possibly from a single crater in one instance).
18. Clementine topography has revealed the broadest and deepest impact basin in the solar system, on the lunar farside (the South Pole-Aitken basin, ~2600 km wide and 12 km deep), which probably excavated into the lunar mantle.
19. Several submarine craters have been located on the continental shelves, including one in Chesapeake Bay, as well as an impact disturbance in the seabed of the South Atlantic that may mark the impact of a ~1-km-diameter impactor that did not penetrate the ocean.
20. A large comet or asteroid strike at the end of the Cretaceous, for which there is abundant physical and chemical evidence as well as a precisely dated, 180-km-diameter, peak-or multi-ring crater buried in the northern Yucatan (and named Chicxulub), must have initiated a sequence of events that led to the extinction of about 90% of all the species on earth at the time, including the dinosaurs and ammonites. The evolution of Homo sapiens ultimately depended on the mammalian radiation that followed.
21. Evidence has been presented that major impacts occurred at the end of the Permian, Jurassic, and Eocene, though none is as conclusive a case as for the terminal Cretaceous. Nevertheless, the course of biological evolution on global and regional scales has probably been profoundly affected in a stochastic way by impacts.

Just as the paradigm of plate tectonics unified the geological sciences in the 1960s by giving an accurate picture of how the earth behaves as a system, so is impact cratering, or more generally, the collision of solar system bodies, giving a unified picture of terrestrial and planetary today. It is no longer valid to speak of earth science and planetary science as distinct entities. The earth and planets are part of one system, a system that interacts and evolves via impact.

SEE ALSO THE FOLLOWING ARTICLES

ASTEROID IMPACTS AND EXTINCTIONS • COMETARY PHYSICS • LUNAR ROCKS • MOON (ASTRONOMY) • PLANETARY GEOLOGY • PLANETARY SATELLITES, NATURAL • PRIMITIVE SOLAR SYSTEM OBJECTS: ASTEROIDS AND COMETS • SOLAR SYSTEM, GENERAL

BIBLIOGRAPHY

Dressler, B. O., and Sharpton, V. L., eds. (1999). Large meteorite impacts and planetary evolution II. *Geol. Soc. Am. Special Paper* **339**.

Dressler, B. O., Grieve, R. A. F., and Sharpton, V. L., eds. (1994). Large meteorite impacts and planetary evolution. *Geol. Soc. Am. Special Paper* **293**.

Grieve, R. A. F., and Therriault, A. (2000). Vredefort, Sudbury, Chicxulub: Three of a kind? *Ann. Rev. Earth Planet Sci.* **28,** 305–338.

Holsapple, K. A. (1993). The scaling of impact processes in planetary sciences. *Ann. Rev. Earth Planet Sci.* **21,** 333–373.

Mark, K. (1987). "Meteorite Craters," Univ. of Arizona Press, Tucson, Arizona.

McKinnon, W. B., Chapman, C. R., and Housen, K. R. (1991). Cratering of the uranian satellites. *In* "Uranus" (J. T. Bergstralh, E. D. Miner, and M. S. Matthews, eds.), pp. 629–692, Univ. of Arizona Press, Tucson, Arizona.

McKinnon, W. B., Zahnle, K. J., Ivanov, B. A., and Melosh, H. J. (1997). Cratering on Venus: Models and observations. *In* "Venus II—Geology, Geophysics, Atmosphere, and Solar Wind Environment" (S. W. Bougher, D. M. Hunten, and R. J. Phillips, eds.), pp. 969–1014, Univ. of Arizona Press, Tucson, Arizona.

Melosh, H. J. (1989). "Impact Cratering: A Geologic Process," Oxford Univ. Press, New York.

Melosh, H. J., and Ivanov, B. A. (1999). Impact crater collapse. *Ann. Rev. Earth Planet Sci.* **27,** 385–415.

O'Keefe, J. D., and Ahrens, T. J. (1999). Complex craters: Relationship of stratigraphy and rings to impact conditions. *J. Geophys. Res.* **104,** 27,091–27,104.

Pierazzo, E., and Melosh, H. J. (2000). Understanding oblique impacts from experiments, observations, and modeling. *Ann. Rev. Earth Planet Sci.* **28,** 141–167.

Ryder, G., Fastovsky, D., and Gartner, S. (1996). The Cretaceous-Tertiary event and other catastrophes in Earth history. *Geol. Soc. Am. Special Paper* **307**.

Spencer, J. R., and Mitton, J. (1995). "The Great Comet Crash: The Impact of Comet Shoemaker-Levy 9 on Jupiter," Cambridge Univ. Press, Cambridge, England.

Lunar Rocks

Arden L. Albee
California Institute of Technology

GLOSSARY

Craters Circular depressions excavated by the impact of meteoroids from space. Large craters (>200 km) are called basins.

Extrusive rocks, volcanic rocks Fine-grained rocks formed by rapid cooling of magma erupted to the surface.

Igneous rocks Crystallized molten rock (magma).

Impact breccia Fragmented and molten rock ejected outward from the crater forming widespread ejecta blankets or rays.

Incompatible elements Elements that concentrate in the residual melt, not being included in the crystallizing minerals.

Intrusive rocks, plutonic rocks Coarse-grained rocks, formed by slow cooling of magma below the surface.

Magma Consists of silicate melt, and may include various silicate and oxide mineral phases, volatile-filled bubbles, and molten globules of metal and sulfide. The chemical elements differ in their affinity for these phases and partition themselves between them in a systematic manner.

Magmatic differentiation Processes by which phases can be separated to produce rocks of differing composition. Crystals of light minerals may float or heavy minerals may sink, in the silicate melt to produce layered cumulate rocks.

LUNAR ROCK studies provide the cornerstone for the scientific findings from lunar exploration by spacecraft since 1959. These findings have completely changed our understanding of moon and its evolution as well as that of earth and the other inner planets. We now understand that the lunar surface features are predominantly the result of impact by numerous huge projectiles during the first half billion years of lunar history and that most of the younger and smaller craters were also formed by impacts, not by volcanism. The role of volcanism is restricted predominantly to the filling, between three and four billion years ago, of the mare basins that resulted from the impacts. The moon did not form by slow aggregation of cold particles that slowly heated up. Instead, it was covered in its early life by molten rock from which a Ca–Al–Si-rich crust formed by accumulation of the mineral plagioclase. As the outer part of moon became rigid, the sources of volcanic lava migrated downward to depths below 500 km.

Since three billion years ago, volcanic activity, has been infrequent and localized. Geochemical similarities show that the moon and the earth must have formed in the same general region of the solar system with a relationship not yet understood.

I. LUNAR EXPLORATION

A. Pre-Apollo Knowledge

Between 1959 and 1976 the United States and the Soviet Union dispatched more than 40 spacecraft missions to explore the moon. These missions (both unmanned and manned) photographed and sensed the moon remotely from orbit, landed and established stations and instruments on the surface, roved across the surface in wheeled and tracked vehicles, drilled holes to collect samples and make measurements, and collected and returned almost 400 kg of rock and soil for detailed study in terrestrial laboratories. Figure 1 shows the locations on the surface of moon of the Apollo and Luna missions.

At the start of this period of exploration, the knowledge of our nearest neighbor was based on continuation of Galileo's pioneering use of the telescope in 1610. Galileo provided the first topographic description of the moon. He noted the great dark smooth areas, which he called maria or seas, and the areas pitted with circular depressions, which he called craters. These intensely cratered areas, which appear white in contrast to the dark maria and also rise above them, are now known as terrae or highlands. Figure 1 illustrates this contrast between the dark flat maria and the cratered highlands. Bright streaks extend as radial rays for hundreds of kilometers from some youthful-appearing, well-defined, bowl-shaped craters.

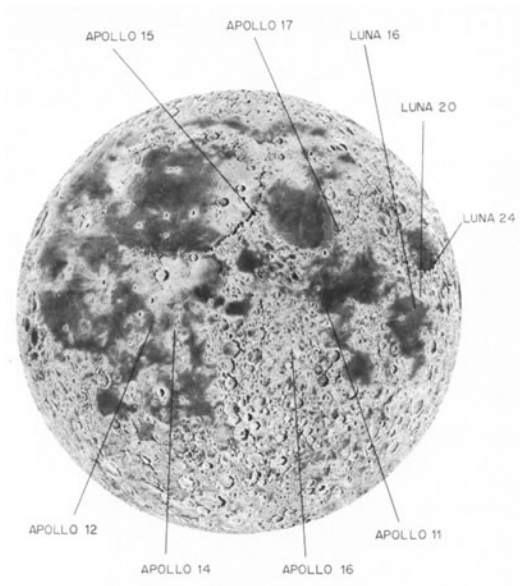


FIGURE 1 Landing sites of the Apollo and Luna missions shown on a relief map of moon. The dark colored smooth mare basins contrast sharply with the lighter colored irregular highlands even as seen with the naked eye. [From Burnett (1977). *Rev. Geophys. Space Phys.* **13**(3), 13–34. Copyright © by the American Geophysical Union.]

Prior to the return of data by spacecraft most observers argued for a common origin, either impact or volcanic, for all features. Early in the 1960s researchers initiated systematic study of the moon on a stratigraphic basic, recognizing that impacts did not merely excavate basins and raise circular ridges or rings, but also ejected rock material to be deposited around the crater or basin in discrete, three-dimensional strata, or ejecta sheets, of fragmented and melted rock. The lunar stratigraphy is based on the time sequence of such strata. These strata overlap one another in the order that they were deposited, just as do beds deposited in the oceans on the earth. Moreover, the oldest exposed surfaces have a higher abundance of craters, resulting from a continual flux of impacting projectiles over longer duration. Surfaces can be correlated in age and their relative age determined by the abundance and character of craters upon them. As on the earth, volcanism also forms beds of lava, which flow out upon the surface from fissures, and beds of pyroclastic material, which are erupted from vents. The maria contain ponds of lava that fill very large depressions formed by earlier impacts. This stratified character now seems obvious, but it became only gradually accepted as telescopic and spaceflight studies advanced during the 1960s. However, this stratigraphic approach provided the scientific basis for targeting the Apollo landings to insure that they would sample rock units of differing relative ages from both the maria and the highlands.

B. Analysis of Lunar Rock Samples

Hundreds of scientists from many countries and many disciplines have been studying the samples, photographs, and instrumental data returned from the moon by the Apollo and Luna programs. Samples of rock and soil were collected and documented so as to maximize the scientific understanding gained at each landing site. These studies have placed significant limits on chemical and physical parameters, on the timing of many events, and on the rate of many processes, and have given insight into the natural processes that formed the moon and shaped its surface. Determination of the mineral and rock characteristics, along with a detailed chemical analysis in terms of major, minor, and trace elements, can be used to deduce processes that produced the various rock types and formed the major rock units. Isotopic techniques can be used to date major events in the history of a rock unit.

The study of the lunar samples required the coordinated efforts of a variety of scientific disciplines using many kinds of instrumentation. Studies of the mineralogy, mineral chemistry, texture, and bulk chemical composition of rocks are used to determine physical and chemical histories. Mineral composition is particularly important, since mineral assemblages reflect both major element chemistry and conditions of formation. Trace element chemistry is used to determine signatures of specific geochemical processes. Precise isotopic analyses are used to study a wide variety of chronologic and geochemical problems. Long-lived radioactive species and their decay products (U–Th–Pb, K–Ar, Rb–Sr, Nd–Sm) are used to measure isotopic ages for rocks and to establish an absolute chronology. The stable isotopes of O, Si, C, S, N, and H provide geochemical tracers that are used in conjunction with chronologic data. Isotopic anomalies left by the decay of extinct short-lived radioactive isotopes provide evidence for very early conditions and time scales. Analysis of the rare gases (He, Ne, Ar, Kr, Xe) and their isotopes provide information on the interaction of the surface with the sun's radiation and the space environment. The samples were examined for evidence of magnetization, a clue to the history of the moon's magnetic field, and for a variety of physical properties, such as density, porosity, thermal conductivity, and seismic wave velocity.

These studies required the development of new precision for the analysis of small samples. They heavily utilized the techniques of optical and scanning electron microscopy, the electron microprobe analyzer, X-ray fluorescence, neutron activation analysis, gas mass spectrometry, isotope-dilution, solid-source mass spectrometry, and ion microprobe analysis. The lunar rocks are certainly the most intensively and extensively studied materials in the history of science—with the exception of the Martian meteorites, which have subsequently been studied with the same techniques.

C. Global Remote Sensing

The remote sensing data obtained by the Apollo orbiters and by telescopic observation of the near side have been vastly augmented in the 1990s by several spacecraft missions. The Clementine mission in 1994 mapped most of the lunar surface with images at a variety of resolutions and wavelengths, with a laser altimeter, and with radio tracking. The Lunar Prospector mission in 1998–99 provided global magnetic field maps and global abundance maps of 11 key elements, including water. The Galileo (1990 and 1992) and Cassini (1999) missions, on their way to the outer planets, also pointed their instruments toward the moon.

II. THE LUNAR SURFACE

A major scientific task has been to understand the nature of the major surface features of the moon and to recognize the processes that have affected the evolution of the surface. The lunar samples provide an understanding of the surface in terms of the rocks and their composition and chronology. Spectral characteristics of the remote sensing data are correlated with specific geologic units and used to map the global distribution of the various rocks that make up the lunar crust.

A. Major Surface Features

Even the naked eye, as illustrated on Fig. 1, detects marked differences in reflectivity over moon's surface. The dark-colored, smooth maria are topographically low basins within the light-colored highlands, which stand more than a kilometer above the maria and have an irregular surface due to the large number of craters. Mountain ranges occur in concentric rings around huge circular basins. Both external and internal processes have shaped this surface. Impacts of projectiles from space have formed craters of all sizes, including huge basins, and the ejecta thrown out from the impacts have produced secondary craters and extensive mantles of fragmented and melted rock. Basins dominate the highlands—more than 40 basins larger than 300 km in diameter have been recognized (Fig. 2) and referred to 15 age groups that can be relatively dated by the abundance of craters on their ejecta blankets. The younger and larger of these basins (e.g., Orientale, Imbrium, and Crisium) have associated ejecta blankets that can be mapped over much of the moon, as shown in Fig. 3, and help to provide the basis for the stratigraphy (i.e., the time sequence of units, shown in Fig. 4).

Such basins occur in roughly equal numbers on both the near and far side of the moon; basins on the near side are partially filled by dark-colored lava flows, whereas those on the far side, in general, are not (Fig. 5). The lava-filled maria have surface features such as flow fronts, channels, domes, depressions, and scarps, which are commonly observed on the earth to form during the flow of lava or during its cooling. Locally, very dark-colored thin deposits represent pyroclastic material erupted from volcanic vents. Craters are less abundant on the smooth mare plains, and these younger craters are better defined and bowl-shaped. The youngest craters (e.g., Tycho and Aristarchus) have thin, bright-colored radial rays of ejecta.

B. Impact Processes

Bombardment of the lunar surface produced craters ranging in diameter from $<10^{-6}$ to $\sim 10^{6}$ m. The larger impact

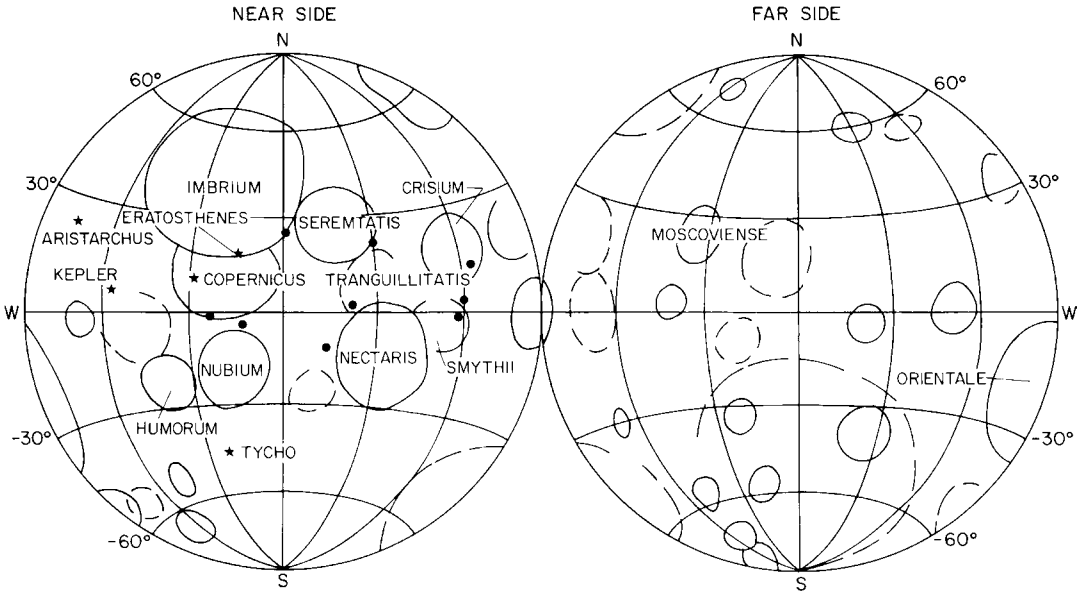


FIGURE 2 Distribution and names of major lunar basins and prominent rayed craters; (⋆) rayed craters; (●) Apollo/Luna landing sites.

events produced major basins, breaking up the outer parts of the lunar crust, transporting material great distances across the lunar surface, and forming layered deposits of impact breccia (i.e., fragmented and recompacted rock). Some of these deposits have been extensively transformed by heat derived from the impact. Clastic fragments are sintered, partially melted, or intimately mixed with impact-derived melt. Some volcanic-textured rocks probably crystallized from impact-melted liquids. The lunar highlands are everywhere broken up, probably to

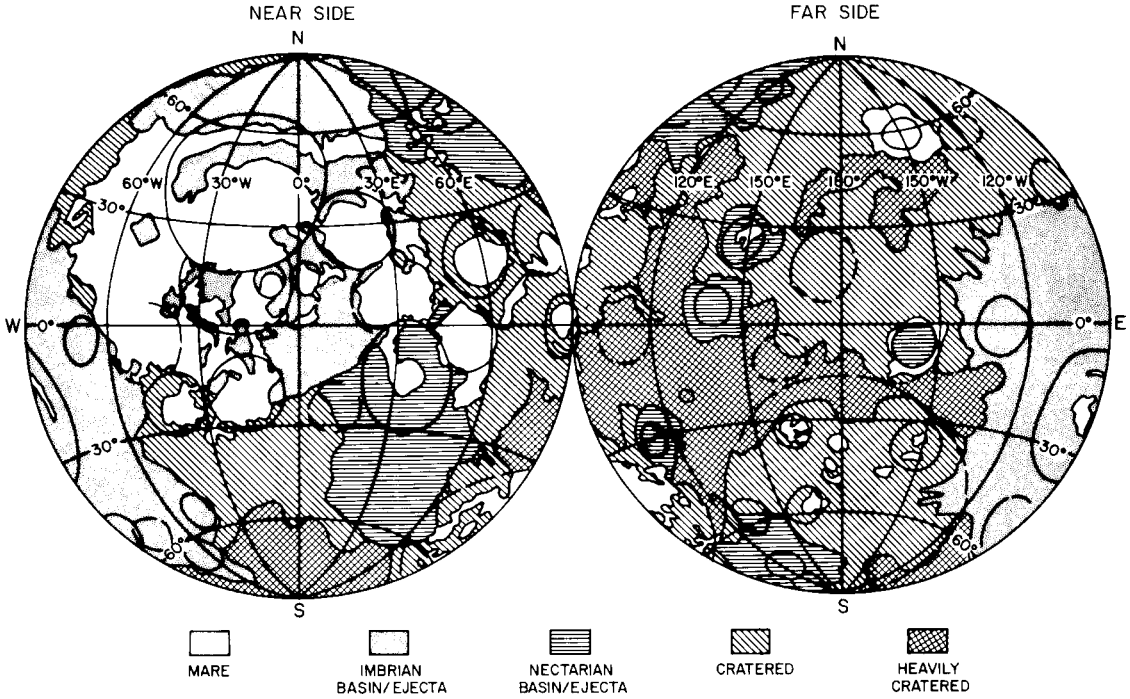


FIGURE 3 Map of major geologic provinces of the moon. Ejecta blankets from the Imbrian and Nectarian basins can be seen to cover much of the surface.

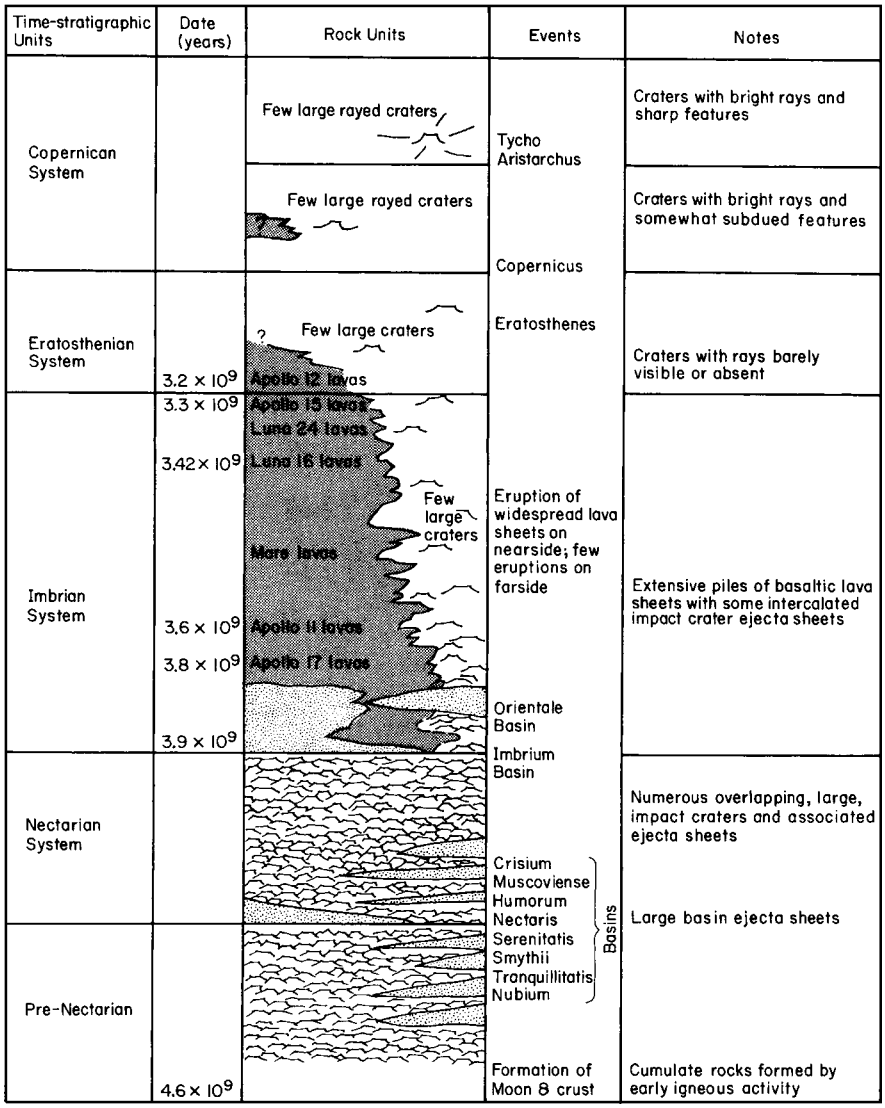


FIGURE 4 Lunar stratigraphic column summarizing the time history of events that have shaped the moon's surface. The positions in the column of the Imbrian and Nectarian ejecta blankets provide a basis for widespread correlation, as do the mare lavas that have been dated by isotopic techniques.

a depth of tens of kilometers, by this process, but the younger mare plains are less heavily and less deeply cratered.

Repetitive bombardment by all sizes of meteorites over long periods of time has produced a generally fine-grained, stratified layer of debris, called regolith or lunar soil, which now blankets nearly all parts of the lunar surface. The regolith differs in thickness from place to place, but is typically a few meters thick on the maria. The debris is composed mostly of the rock types that immediately underlie the regolith, though these rock fragments may themselves be breccias from an impact blanket. Part of the rock fragments in the debris have been derived from a considerable distance, having been dispersed across the lunar surface by impacts (and thereby increasing our knowledge of the lunar surface from the few sample sites). The regolith also contains abundant impact-produced glass and a few percent of meteoritic material from the impacting bodies. The impact of micrometeorites upon the regolith produces tiny puddles of silicate melt that cool rapidly to agglutinates of glass intimately mixed with rock and mineral grains. The abundance of such impact-produced glass agglutinates in a layer reflects the length of time that the layer was exposed to micrometeorite bombardment at the surface before being covered by a new layer thrown out from a nearby impact.

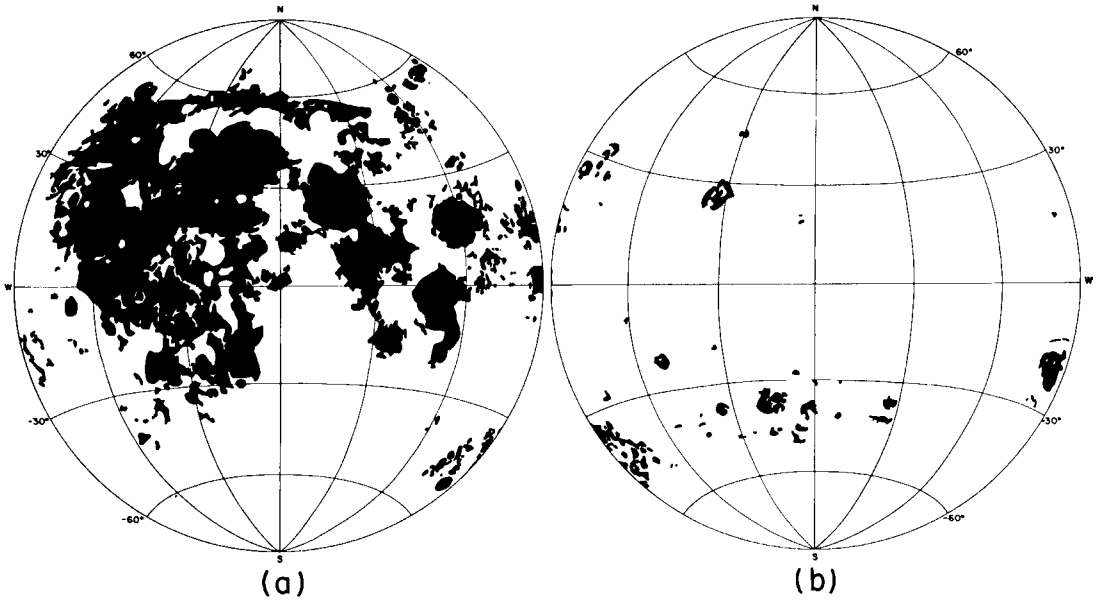


FIGURE 5 Distribution of basaltic flows in the mare basins on the (a) near side and (b) far side of the moon.

C. Highland Breccias and Ancient Rocks

The repeated impact episodes are reflected in the complexities of the fragmental rock (breccia) samples returned from the lunar highlands. The shock of the impact results in intense fragmentation and melting. The molten material may quench to glass during the ejection, or the hot mixture may remelt, sinter, and recrystallize during deposition and cooling in a thick ejecta blanket. The breccia samples range from friable aggregates to hard, sintered material with spherical vesicles that were bubbles filled by a gas phase prior to solidification. Many samples show multiple generations of impacts; fractured fragments of ancient rocks are within irregular fragments of breccia that are themselves contained in a mixture of fragments and melt rock.

The major minerals within the highland breccias are anorthite-rich plagioclase ($CaAl_2Si_2O_8$), orthopyroxene ($[Mg,Fe]SiO_3$), and olivine ($[Mg,Fe]_2SiO_4$); these occur both as mineral fragments and as plutonic rocks made up predominantly of these minerals. The high content of anorthitic plagioclase and the low abundance of iron and titanium oxide minerals is responsible for the light color and for the characteristically high calcium and aluminum composition of the lunar highlands. The words anorthosite, norite, and troctolite are used in various combinations as adjectives or nouns to describe coarse-grained rocks made up of various combinations of these three minerals. Hence the acronym ANT is commonly used to describe this suite of rocks. Such rocks are found on the earth in layered igneous bodies that have crystallized from a silicate melt or magma very slowly deep beneath the surface. The term magma includes not only the complex silicate melt, but the various crystallizing minerals, and may include bubbles of volatiles and globules of sulfide or metal melt. Plagioclase-rich rocks such as the ANT suite do not form by simple crystallization of magma, but represent accumulation of early crystallizing minerals by floating or settling, as evidenced by terrestrial examples of cumulate rocks. Remote sensing maps indicate that anorthosite is the dominant rock type of the highlands.

Despite the complex history, a number of fragments of ANT rocks collected from the breccia have yielded isotopic ages greater than 4.4 billion years (Gy), indicative of crustal formation dating back almost to the origin of the solar system. The existence of an early crust was also inferred from geochemical evidence. The rare earth element europium, unlike the other rare earth elements, is highly concentrated in plagioclase during crystallization of a silicate melt. This element has a relatively high abundance in the highlands rocks and is relatively underabundant in the lunar basalts. These complementary anomalies are ascribed to extensive early differentiation of the primitive lunar material into a plagioclase-rich crustal cumulate of crystals and a more mafic melt, which eventually became the source of the lunar basalts. Hence, it is inferred that much of the outer part of the moon was molten that is, a magma ocean during the early part of lunar history.

This early differentiation seems also to have been responsible for another compositional class of material rich in K, rare earth elements, and P (KREEP). These elements are representative of the "incompatible elements"

(which also include Ba, U, Th, and Rb) that do not enter the crystal structure of the major lunar rock-forming minerals and hence become concentrated in the residual liquid during final crystallization of a magma. The abundance of KREEP ranges greatly in the samples of highland breccias and regolith, occurring as both small rock fragments and glass. However, the uniformity of abundance pattern and the isotopic systematics, albeit partially disturbed in some cases, suggest a rather homogeneous source, one that was enriched in the incompatible elements at about 4.4 Gy. Orbital measurements of gamma rays have shown that material rich in K, Th, and U is concentrated in the region of Mare Imbrium and Oceanus Procellarum. The KREEP-rich material may have been distributed from these regions into the regolith by impact scattering.

D. Mare Volcanism

Four of the Apollo missions landed on the mare plains and returned samples of mare volcanic rocks. The lunar mare rocks are basalts similar in texture and chemical composition to volcanic basaltic rocks on the earth. The lunar basalts consist chiefly of the silicates—clinopyroxene ($[Ca,Fe,Mg]SiO_3$), anorthitic plagioclase ($CaAl_2Si_2O_8$), olivine ($[Mg,Fe]_2SiO_4$)—and the iron–titanium oxides, ilmenite and spinel. The mineralogy is generally similar to that on earth except for the very low content of sodium in the plagioclase, the high abundance of ilmenite, and the total absence of hydrous alteration minerals. Textures range from fine-grained and partially glassy, such as might be expected in a basalt that chilled quickly at the surface of a flow, to coarser-grained, interlocking textures that result from slower cooling within a flow.

Terrestrial basalts are formed by partial melting in the olivine and pyroxene-rich rocks of earth's mantle. As melting occurs, the early formed silicate melt has the composition of basalt, and it separates and rises to the surface. The detailed composition of a basalt provides a probe into the temperature, depth, and composition of the source; basalt can be readily dated, and the textures provide information on the cooling and crystallization history. Mare basalts differ from terrestrial basalts in some detailed elemental abundances. Lunar basalts (1) contain no detectable H_2O; (2) are low in alkalis, especially Na; (3) are high in TiO_2; (4) are low in Al_2O_3 and SiO_2; (5) are high in FeO and MgO; and (6) are extremely reduced. Lunar basalts contain no trivalent Fe, and the reduced ions (Fe metal, trivalent Ti, and divalent Cr) may be present.

The petrologic and chemical variation within the lunar basalts of the various missions is summarized in Fig. 6; their isotopic ages fall between 3.15 and 3.96 Gy. The samples can be divided into two broad groups: an older, high-titanium group (ages 3.55–3.85 Gy; TiO_2, 9–14%) and a younger, low-titanium group (ages 3.15–3.45 Gy; TiO_2 1–5%). Laboratory experiments indicate that the high-titanium mare basalts could be derived from partially-melted titanium-rich cumulates at depths of about 100 km and that low titanium basalts could be produced by partial melting of olivine and pyroxene-rich rock at depths of 200–400 km.

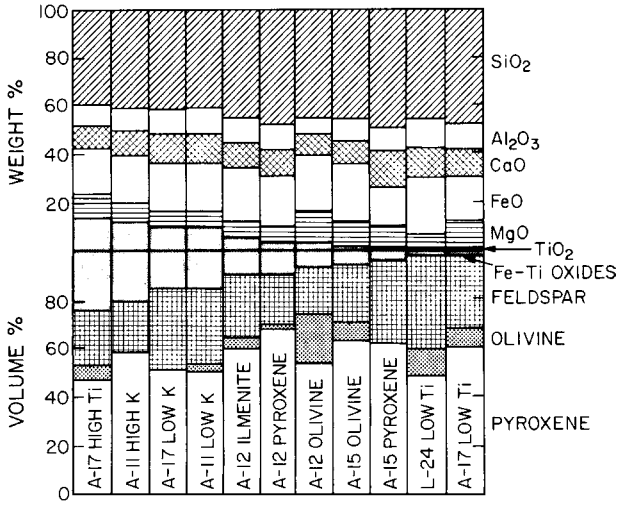


FIGURE 6 Mineral and chemical compositions of the basalt groups identified at the various landing sites. All consist predominantly of pyroxene and plagioclase feldspar, but Fe–Ti oxides (and TiO_2 content) and olivine differ in abundance.

Earth-based telescopic spectra indicate that a wide variety of volcanic flows cover the maria, only a third of which were sampled by the Apollo missions. Further, the multispectral images from Clementine suggest that the high Ti basalts, so abundant in the returned samples, cover only a small fraction of the surface lava flows on a global basis.

E. Lunar Surface History

Isotopic analyses of samples from its surface show that the moon, as a body, was formed at 4.6 Gy. It is similar in age to the earth and to meteorites derived from other bodies in the solar system. Isotopic ages have been determined for several events on the lunar geological time scale. Most of the recognizable major geologic events, formation of the basins and infilling of the mare basins, occurred during the first 1.5 Gy of lunar history.

During the first half billion years of lunar history intense meteorite bombardment pulverized and mixed the upper zones to great depths, culminating in the sequential excavation of the great multiringed basins with their widespread deposits of impact breccia (Figs. 3 and 4). Many feldspathic impact breccias have recrystallization ages greater than 3.9 Gy, and fragments of coarse-grained rock from within the breccias have ages of over 4.4 Gy,

dating back to the formation of the lunar crust. The bombardment rate declined rapidly after 4 Gy, approaching the current rate.

Commencing about 3.9 Gy vast floods of basaltic lava poured out on the lunar surface within the basins, particularly those on the near side (Fig. 5). These basaltic eruptions occurred significantly later in time than the excavation of the basins and filled in the topographically low areas of the basins. Although these basalts cover about 17% of the moon's surface, they constitute less than 1% by volume of the crust. The measured isotopic ages range from 3.9 to 3.2 Gy, but photographic and remote sensing studies indicate that dark-colored basalts on the western near side are somewhat younger and were not sampled by the missions.

Except for minor lava flows and a relatively small number of large impact craters, the face of the moon has remained largely unchanged for the last 3 Gy. A few of the very young impact craters have been dated by use of returned samples.

III. THE LUNAR INTERIOR

Our knowledge of the lunar interior consists of inferences based on geophysical data and the detailed studies of the surface samples. Especially important have been the moonquake and artificial-impact recordings by the landed seismometers. The major results (the existence of a crust and the low level of dynamic activity of the interior) have major implications for our understanding of the moon.

A. Chemical Composition of The Interior

To a first approximation lunar chemistry is dominated by the crustal formation process, and the interior composition can only be inferred. Neither the mare basalts nor the plagioclase-rich rocks of the highlands have compositions that are consistent with the mean density of the moon. Laboratory experiments show that both these rock types would undergo phase changes to higher density materials at fairly shallow depths, which would yield too high a lunar mean density. This evidence, along with the high radioactivity of the surface rocks, the heat flow measurements, and isotopic data, demonstrate that the moon has a crust with a composition different from its bulk composition. Consideration of the mean density indicates that the moon must have less metallic iron than the earth, and that any metallic core must be very small.

The mare basalts were derived by partial melting of the upper mantle and provide a sensitive probe to the composition of the lunar interior and, by inference, the bulk moon. The bulk composition of moon is not primitive, that is, it does not have the composition of the sun or of carbonaceous condrite meteorites, that are thought to be very primitive because of their similarity to the solar composition. Apart from the obvious paucity of gaseous elements, such as hydrogen and helium, several differences are known.

1. Volatiles

Volatiles are those elements that condense from the solar nebula below about 800°C: The total lack of hydrated minerals in lunar rocks indicates that water has been in very low abundance throughout lunar history. This low water content is also indicated by the very low attenuation of seismic waves in the lunar crust. The abundance of indigenous carbon is also strikingly low; crystalline rocks contain up to a few tens of parts per million by weight of carbon. Regolith samples contain up to an order of magnitude more, but this is largely trapped from the solar wind as it impinges on the surface. Nitrogen is equally rare. The moon is, and probably always has been, an exceedingly inhospitable environment for organic chemicals, prebiotic or biotic. Lunar rocks are also consistently deficient in alkali metals (Na, K, Rb) by up to a factor of ten relative to their terrestrial counterparts. The elements of higher volatility such as Pb, Bi, Tl, Br, etc., are even lower in relative abundance (by factors of 10–100). Most, if not all, of these low abundances probably reflect initial deficiency, not subsequent loss.

2. Refractories

Refractories are those elements with condensation temperatures above 1200°C. Lunar surface rocks, in general, contain the refractory elements (Al, Ca, Sr, Ti, Zr, Th, U, and rare earths) in such high abundances (up to several hundred times chondritic abundances) that it appears certain, especially in conjunction with the heat flow evidence in Section III.B that the bulk moon is enriched in these elements relative to chondritic meteorites. Such an enrichment is consistent with the apparent initial deficiency of volatile elements.

3. Siderophiles

Siderophiles are those elements that tend to enter iron metal rather than oxides, sulfides, or silicates. Metals such as Au, Ir, Re, Ni, etc., are depleted in lunar surface rocks 10^{-4} to 10^{-5} times their solar abundance. They are also much more depleted than on the earth; gold, for example, being more depleted on the moon by two orders of magnitude. This difference suggests that the moon lost its noble metals under physical and chemical conditions quite

different from those of the earth. The low abundance of siderophile elements, coupled with the low density of the moon, suggests the separation of a metal phase from the lunar material prior to the formation of the moon as a body.

4. Lunar Oxygen Isotope Compositions

Ratios of the abundances of the three natural isotopes of oxygen—^{16}O, ^{17}O, ^{18}O— are quite different in lunar rocks from those found in basaltic meteorites derived from asteroids or from Mars, which in turn differ from those found in chondritic meteorites. Moreover, ratios of the abundances of these elements found in lunar and terrestrial rocks show a distinct trend, one offset from the trend line for Mars meteorites. It follows that the earth and moon share some common genetic link.

B. Internal Temperature and Radioactivity

Indirect evidence of temperatures in the moon's interior comes from the relative attenuation of seismic waves at various depths, the depth of present seismic activity, the electrical conductivity profile deduced from the interaction of the moon with abrupt changes in the interplanetary magnetic field, and measurements of the near-surface heat flow.

The high attenuation of seismic waves penetrating to depths greater than 800 km and the occurrence of all sizable moonquakes near that depth suggests that partial melts may exist below that depth in the present moon. If so, temperatures must be about 1500°C at that depth. Electrical conductivity profiles of the interior are interpreted as indicating interior temperatures of 1300–1500°C.

Two measurements of the surface heat flow yielded values that are about one-third that of the average heat flow on the earth, but are higher than would be predicted by thermal history calculations for a moon with a bulk uranium and thorium content like that of chondritic meteorites. This greater abundance of uranium supports the indication from the surface samples that the whole moon is enriched in refractory elements relative to the chondrites and the earth. Such a high abundance of long-lived radioactive elements, would lead to total melting of the interior, unless they are mainly concentrated near the surface; this provides further evidence that a substantial portion of the moon's interior is radially differentiated.

C. Seismicity

A large number of very small moonquakes have been detected by the Apollo seismic network. The total seismic energy release within the moon appears to be about 80 times less than that in the earth. The moonquakes are concentrated at great depth—between 600 km and 1000 km—which is deeper than earthquakes. The difference in distribution and magnitude of seismic energy relase on the earth and the moon is probably related to fundamental differences in interior dynamics. Correlation of deep moonquakes with lunar tides indicates that tidal stresses must play an important role in triggering them.

D. Gravity and Shape of Moon

The gravity field and the shape of the moon have been determined from orbital measurements using a laser altimeter and radio tracking. The near side is relatively smooth with 5–6 km of relief whereas the far side is rougher with up to 16 km of relief. The large South Pole-Aitken basin on the far side is about 12 km deep. Broad-scale gravity differences indicate regional variations in the moon's internal density. Calculations using reasonable choices of density indicate that the crust averages about 70 km in thickness, varying from a few tens of kilometers beneath some mare basins to over 100 km in some areas of the highlands. Furthermore, the center of mass is about 2 km closer to earth than the center of figure; this can be attributed to greater crustal thickness on the far side.

Striking positive gravity anomalies, the mascons, coincide with major ringed basins. The filled multiringed basins on the near side are mascons, whereas, the unfilled multiringed basins on the far side are negative anomalies. Thus, the mascons are associated with the filling rather than the excavation of the basins. The excess mass, in conjunction with the fact that the surface of the maria lies below that of the surroundings, can be understood by the replacement of lower density crustal rocks by higher density basalt. The mascons are associated with topographic basins that have persisted for more than 3 Gy. To support these mascons requires that the outer few hundred kilometers of the moon have great strength, and seismic data support this interpretation.

E. Magnetic Field

The moon has a very small magnetic field at the present time. However, the returned lunar samples contain natural remanent magnetism of lunar origin, which is probably due to cooling of fine particles of metallic iron in an ambient magnetic field. Local remanent magnetic fields measured at Apollo landing sides and over large regions scanned by the satellites indicate that magnetic fields were present at the lunar surface during a period of igneous activity. The localized surface fields with scales of 1–100 km suggest that a crust, initially magnetized in a field of global scale, was broken up by impacts. Fields 50–500 times less than earth's fields, but 20–200 times greater than the field in the solar wind, are required to produce such fields. One

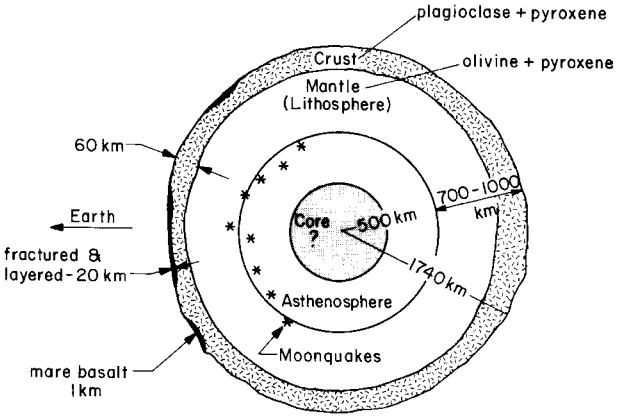


FIGURE 7 Diagrammatic section of the lunar interior.

interpretation is that the moon had a liquid core, rich in metallic iron, which behaved as a dynamo during the time when the mare basalts were crystallizing. The only alternative explanation would be that there existed an external field of unknown origin.

F. Internal Structure

A variety of evidence indicates that the moon has been radially differentiated and is a layered body with a crust, mantle, and possibly core as shown in Fig. 7. It is covered by an extremely dry, highly heterogeneous layer to a depth of about 25 km. This layer is related to the extensive fracturing of the crust by cratering processes, but layered sequences of impact blankets and lava flows contribute to the general complexity of this zone. Seismic velocities are nearly constant from 25- to 65-km deep with a value consistent with a plagioclase-pyroxene rock. A large velocity increase at about 65 km marks the base of the crust. The 600–900-km thick upper mantle has low attenuation of seismic waves and nearly constant velocities. These and other properties are consistent with a pyroxene-rich composition with olivine, spinel, and possibly plagioclase in small amounts. Below about 700 km the shear-wave velocity decreases, attenuation increases, and deep moonquakes occur; this may represent a partially molten zone. The moment of inertia and overall density of the moon and other data do not rule out the existence of a small iron-rich core, but do place an upper limit of about 500 km on its radius.

IV. THE MOON–SPACE INTERFACE

Unlike the earth, which is partially protected by its magnetic field and atmosphere, the flux of particles and radiation from the sun and of meteorites upon the lunar surface has imprinted evidence of the history of the solar system upon the surface materials. The complex conditions at this interface with space play a major role in giving the surface of the moon its appearance, and in giving the lunar regolith its many novel characteristics.

A. Record of Lunar Surface Environment

Lunar soil properties cannot be explained strictly by broken-up local rock. Distant impacts throw in exotic material from other parts of the moon. About 1% of the soil appears to be of meteoritic origin. Vertical mixing by impacts is important; essentially all material sampled from lunar cores shows evidence of having resided at the surface.

The layers of the regolith preserve a record of meteorite, solar particle, and cosmic-ray bombardment. The surface layers exposed to space contain a chemical record of implanted solar material (rare gases, H, and other elements transported from the sun in the solar wind). During the impact processes, lunar surface material became heated, and lead and other volatile and gaseous elements were mobilized and then trapped in the regolith, enriching it in such elements.

Theoretical calculations have shown that water molecules, formed in these processes or derived from meteoritic influx, could be cold-trapped in permanently shadowed basins near the lunar poles. The neutron spectrometer on Lunar Prospector has mapped the global abundance of hydrogen. The results support the theory and indicate that a significant amount of water ice is contained in the regolith near each pole.

Radiation damage effects dominate the physical properties of the surface layers. The surfaces of rock samples record the flux of micrometeorites by the presence of numerous small impact craters, many of them microscopic in size.

The abundance of atomic nuclei produced by cosmic-ray bombardment provides a measure of the time duration that rock material resided in the upper few meters of the regolith before being turned under or covered by impact debris. Individual rock fragments found on the surface have been shown to have resided with a few meters of the surface for times (exposure ages) varying from a few million to 500 million years. A few parts of the regolith are known to have remained undisturbed to meter depths for several hundred million years. The ages of several craters have been determined by measuring exposure ages on blocks thrown out from the crater.

B. Solar Record

The isotopic composition of the heavy rare gas component of the solar wind has been measured in soils and

lunar samples. Amorphous surface films, produced by solar wind bombardment, are observed on many lunar grains. Comparison with artificial irradiations has set limits on fluctuations in the energy of the ancient solar wind.

The concentrations of hydrocarbons (mainly methane and ethane) correlate with the solar wind irradiation of different samples of lunar soil regolith. It is believed that these compounds formed by trapping and reaction in the surficial damaged layers. Since interstellar space contains both dust clouds and sources of energetic particles, this may be an important model for organic synthesis in the galaxy. Large isotopic fractionation effects for oxygen, silicon, sulfur, and potassium are present in the surface layer.

Dramatic progress has been made in the study of solar flare particles. The energy spectrum of heavy solar particles has been established. It has been shown the lowest energy solar cosmic rays are highly enriched in very heavy nuclei compared to normal solar material, the first demonstration of preferential heavy ion acceleration by a natural particle accelerator.

Information has been obtained on solar flare activity on the sun over geologic time by measuring the induced radioactivities and nuclear particle tracks produced in the outer layers of surface material. The average solar flare activity has not changed appreciably over the past few million years, suggesting that major climatic changes during this time are not related to largescale changes in solar activity, as had previously been postulated.

V. THE EVOLUTION OF MOON

A major result of Apollo lunar science has been the understanding of the evolution of another planetary body for comparison with that of the earth. The moon and probably the earth went through a period of major igneous activity and a high rate of impact during their first 1.5 Gy. However, since then the moon has been thermally and mechanically quiescent; if our exploration of the moon had taken place 2.5 Gy earlier, few aspects of the moon would have been different. The importance of understanding this very early history lies mainly in its implication for the earth and the rest of the solar system. Nowhere else have we had access to as detailed an ancient record of planetary evolution. Our knowledge of this evolution is summarized in the cartoon in Fig. 8.

A. Chemical and Isotopic Composition

The bulk composition of the moon is not primitive; it does not have the composition of the sun or of the primitive chondritic meteorites. The moon seems to contain more than its cosmic share of refractory elements and much less

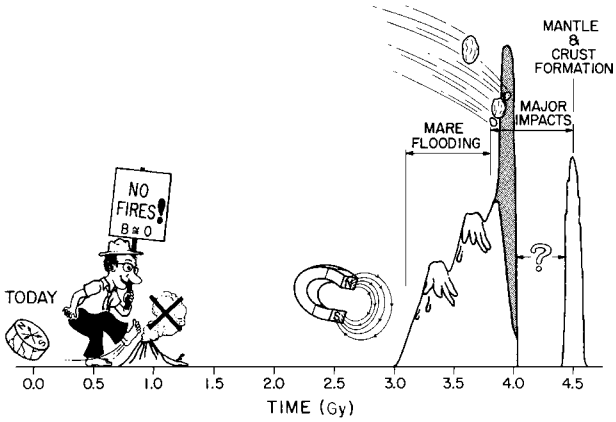


FIGURE 8 A cartoon showing the chronology of major events in lunar history. [From Wasserburg *et al.* (1977). *Philos. Trans. R. Soc. London A* **285,** 7–22.]

than its cosmic share of volatile elements. Yet, the composition of the moon cannot be explained simply by cessation of equilibration with the vapor of the solar nebula at some high temperature. The amount of oxidized iron (in silicates and oxides) and the abundances of volatile metals suggest equilibration of part of the condensate at 200°C. Thus the moon seems to be a mixture of materials that separated from the nebula over a wide range of temperature, like the chondrites and presumably many other bodies in the solar system.

Measurements of oxygen isotopic compositions in meteorites have provided evidence of compositional inhomogeneites in the solar nebula. The earth, moon, and the basaltic meteorites could have formed from a common oxygen isotope reservoir, separate and distinct from that of the chondritic meteorites. The basaltic meteorites—achondrites derived from asteroids or planets—tend to show trends or correlations in their isotopic ratios that reflect the partitioning of oxygen isotopes between silicate minerals during melting and crystallization. Meteorites from the same parent body fall on a single trend line and, for example, the trend line for earth and moon rocks is distinct from that for Martian meteorites.

B. Crustal Formation and Early Differentiation

The moon has differentiated into layers. Seismic and gravity measurements indicate that a crust (∼60-km thick) of low density (∼3.0 g/cm^3) material overlies a higher density (∼3.35 g/cm^3) mantle. Isotopic, chemical, and petrologic data indicate that the crust originated by magmatic differentiation very early in lunar history. There was no knowledge of a lunar crust before the Apollo missions; indeed, the moment of inertia suggested that the moon was nearly uniform, and hence undifferentiated. One widely

held view at this time was that the moon had accreted cold and had not been heated sufficiently by internal radioactivity to melt.

No samples of "primitive" material were returned by Apollo. However, a variety of igneous cumulate rocks (the ANT suite of rocks) were returned. These probably formed in the early differentiation of the lunar crust, since isotopic ages of greater than 4.4 Gy have been measured on a number of such samples.

Petrologic, geochronologic, geochemical, and geophysical constraints support the idea that accretional melting involved at least the outer few hundred kilometers. Crystallization could have formed a chilled surface layer underlain by (1) an anorthiterich cumulate layer; (2) a pyroxeneolivine–spinel cumulate layer; and (3) trapped residual melts, rich in Ba, U, K, REE, P, Th, and highly concentrated in an intermediate level. The layering would have been disrupted by early impacts, but such magmatic differentiation produced a body systematically zoned in both major and minor elements.

C. Impact History

From 4.6 to 3.9 Gy lunar petrologic history was dominated by impact brecciation and heating to great depths. The large mare basins represent only the last stages of the impact of large bodies on the lunar surface. The profound cratering on all scales of the highlands provides a record of this period of impact and cratering. The Apollo age dates indicate that the Imbrium basin, the second youngest of the large mare basins, was formed only 0.5 Gy after the moon's accretion. During the first half billion years, the rates of impact may have been so great at times that most of the large craters that formed early were totally obliterated by later impacts. Similar catastrophic events are now seen to have affected the earth and the other inner planets.

The craters on the mare surface are a record of impacts during the last 3 Gy. The average cratering rate has been much lower than during the first half billion years. Only a limited number of medium-sized (20–100 km) craters such as Copernicus and Tycho formed during this period. Study of particles from the regolith at the Apollo 12 site suggests that Copernicus may have formed 850 million years ago.

D. Mare Filling and Volcanism

Many large multiringed basins, especially those on the near side, were filled with basaltic flows in the interval 3.9–3.2 Gy. These eruptions occurred significantly later than the formation of the basins. The individual flows are thin and very extensive, reflecting the low viscosity of the lava; they differ slightly in composition from one another. Plagioclase-rich basalts and KREEP-rich basalts with ages greater than about 3.9 Gy may have been derived either by partial melting at shallow depths or from impact-produced melts. From 3.8 to 3.6 Gy titanium-rich mare basalts were extruded, and experimental studies suggest that these were derived from partially melted titanium-rich cumulates at depths of about 100 km. Low-titanium mare basalts were extruded at 3.4–3.2 Gy, produced by partially melted olivine pyroxenite at depths of 200–400 km. Various compositions of flows and flows somewhat younger than 3.2 Gy, have been identified by remote sensing methods, but were not sampled. It seems likely that the regional lunar volcanic activity stopped by about 2.3 Gy.

E. Thermal History

Lunar heat flow, electrical conductivity, seismic velocities and the degree of attenuation at various depths, lunar viscosity, seismicity, the presence of mascons, and tectonism provide evidence on the present-day temperatures and are manifestations of the internal structure. Magnetized lunar rocks, the early differentiated lunar crust, the abundance of heat-producing elements, and the chronology of lunar magmatic activity indicate the thermal state earlier in lunar history. The outer 700 km is relatively cool, strong, and inert—and has been so for about the last 3 Gy. However, the outer part of the moon, at the very least, melted early in lunar history during the closing stages of accretion by impact. This melting permitted extensive magmatic differentiation with upward concentration of radioactive heat sources into the lunar crust. Below 700 km, the moon may still be partially molten with temperatures between 1000 and 1600°C.

VI. THE ORIGIN OF MOON

To understand the early history of the solar system, it is necessary to know the bulk chemical composition of the planets and their satellites, and also to know the chemical rules that govern the assembly of planetary bodies from a cosmic cloud of dust and gas. The analysis of the lunar rocks is the first step in the chemical and isotopic mapping of the solar system beyond the earth by direct study of returned samples. Similar analyses performed on chondritic meterorites have provided considerable insight into the behavior of the elements during the condensation of the solar nebula.

Prior to lunar exploration it was commonly assumed that the bulk compositions of the inner planets were closely similar to the composition of chondritic meteorites, which are near solar abundance in composition except for the most volatile elements. The sample studies

have shown dramatically that the moon is not chondritic in composition, a fact which seriously challenges our earlier assumption about the makeup of the inner planets.

It now appears that the moon and the earth accumulated from material quite changed from solar abundances. These changes consist of enrichment in elements, which are believed to have condensed from the solar nebula at high temperatures and depletion of many less refractory elements as well as the very volatile gases. The moon is apparently enriched in high-temperature condensates, possibly similar to samples preserved as inclusions in the Allende meteorite and similar carbonaceous chondrites. However, such inclusions have oxygen isotope compositions so different that they must have formed in a different part of the nebula.

Despite the detailed studies, the problem of lunar origin remains. How is it possible for a planet to become enriched in the refractory condensates? What physical processes operate in the solar nebula to concentrate refractory-rich dust from volatile-rich dust? Much more work is obviously required in order to understand this complicated chemical fractionation.

However, our conception of the process that generated the planets has been altered. We now consider that a variety of planetary objects were formed, which ranged from accumulations of material enriched in refractory elements (the terrestrial planets) to relatively unfractionated bodies, represented by the outer planets and all the low-density objects in the solar system. The earth no longer appears peculiar, but seems to belong to a compositional class of planets. The compositions of the planets provide a basis for reconstructing the accumulation processes that governed planet formation. It now appears possible to determine the duration of the accumulation processes and the time scale required for chemical segregation within the planets. The moon preserves the early stages of internal segregation and also shows that the actual accretion process, and major collisions, extended for a much longer period that was considered before. Whatever the origin of the impacting objects (local or distant), the fact that catastrophic impacts occurred as recently as 3.9 Gy on the moon suggests that the earth suffered similar bombardment, which strongly affected its early geological record.

The lunar studies of the past few decades have not produced conclusive evidence on the moon's origin. However, since a major conference in 1984, a giant impact or collision hypothesis has emerged that seems to satisfy most of the dynamical and chemical constraints on the origin of the earth–moon system. The variants of this hypothesis propose that during the final stages of the accretion of the terrestrial planets a body (or bodies) larger than Mars collided with earth and spun out a disk of hot material from which the moon formed.

SEE ALSO THE FOLLOWING ARTICLES

METEORITES • METEORITES, COSMIC RAY RECORD • MOON (ASTRONOMY) • PLANETARY GEOLOGY • PLANETARY SATELLITES, NATURAL

BIBLIOGRAPHY

Basaltic Volcanism Study Project (1981). "Basaltic Volcanism on the Terrestrial Planets," Lunar and Planetary Institute. Houston, Texas.

Burnett, D. S. (1977). Lunar Science: The Apollo legacy, *Rev. Geophys. Space Phys.* **13**(3), 13–34.

Hartmann, W. K., Phillips, R. J., and Taylor, G. J. (eds.) (1986). "Origin of the Moon," Lunar and Planetary Institute. Houston, Texas.

Hood, L., and Zuber, M. T. (in press). Recent refinements in geophysical constraints on lunar origin and evolution *In* "Origin of the Earth and Moon" (R. Canup and K. Righter, eds.), Univ. Arizona Press, Tucson.

Newson, H. E., and Taylor, S. R. (1989). Geochemical Implications of the formation of the moon by a single giant impact. *Nature* **238,** 29–34.

Special issue of Clementine Mission results (1994). *Science* **266,** 1835–1848.

Special issue of Lunar Prospector Mission results. (1998). *Science* **281,** 175–1496.

Spudis, P. D. (1999). "The Moon. The New Solar System," (J. K. Beatty, C. C. Petersen, and A Chaikin, eds.), Cambridge Univ. Press, Cambridge, U.K.

Stevenson, D. J. (1987). Origin of the moon—the collision hypothesis. *Ann. Rev. Earth Space Sci.* **15,** 271–316.

Taylor, S. R. (1982). Planetary Science: A Lunar Perspective, Lunar and Planetary Institute. Houston, Texas.

Wasserburg, G. J., Papanastassiou, D. A., Tera, F., and Huneke, J. C. (1977). I. The accumulation and bulk composition of the moon: Outline of a lunar chronology. *Philos. Trans. R. Soc. London A.* **285,** 7–22.

Wilhelms, D. E. (1986). Geologic history of the moon, *U.S. Geol. Surv. Prof. Pap.*, no. 1348.

Meteorites, Cosmic Ray Record

Robert C. Reedy
Los Alamos National Laboratory

GLOSSARY

Complex history Case where a meteoroid was exposed to appreciable intensities of cosmic rays both before and after it suffered a collision in space that changed its geometry.

Cosmogenic Product of cosmic-ray interactions.

Electron volt Unit of energy, whose symbol is eV, that is necessary to raise one electron through a potential difference of 1 volt. It is equal to 1.602×10^{-19} joules. Nuclei in the cosmic rays typically have energies ranging from about a mega-electron-volt (MeV) (10^6 eV) or less to many giga-electron-volts (GeV) (10^9 eV) per nucleon.

Exposure age Length of time that a meteorite has been irradiated by a sufficient intensity of cosmic-ray particles to produce an observable record.

Meteoroid Object in space that, after it survives passage through the Earth's atmosphere and hits the surface, is called a meteorite.

Noble gases Elements helium, neon, argon, krypton, and xenon, which are gases at normal temperature and pressure, almost never form compounds, and are usually very rare in meteorites. They are sometimes called rare gases or inert gases.

Radionuclide Atomic nucleus that is unstable and decays to a different nucleus by the emission of radiation, such as ^{10}Be, which emits an electron and becomes a ^{10}B nucleus. Each radionuclide decays with a specific half-life, the time required for half of an initial population to have decayed away.

Shielding Location of a sample inside a meteoroid and the size and shape of the meteoroid, which are important in determining the rates at which cosmogenic products are produced in that sample.

Simple history Case where a meteoroid, having previously been deeply buried in its parent object far from the cosmic rays, received its entire cosmic-ray record without experiencing a collision that changed its geometry.

Spallation Type of nuclear reaction induced by an energetic particle in which one or more nuclear fragments (nucleons, α particles, and so on) are removed from a nucleus, such as the production of ^{53}Mn from ^{56}Fe.

A product of such a reaction is called a spallogenic nuclide.

Thermoluminescence (TL) Property of certain minerals that, after electrons are excited into traps by ionizing radiation, they emit light when moderately heated.

Track Path of radiation damage made by heavily ionizing particles, such as iron nuclei, in a dielectric medium, such as olivine or other meteoritic minerals. These paths normally are observable only by electron microscopes but can be dissolved and enlarged by certain chemicals that preferentially etch along the damage zone.

VH or VVH Label applied to two groups of cosmic-ray nuclei with VH (very heavy) nuclei having $20 \leq Z \leq 28$ and VVH (very, very heavy) nuclei having atomic numbers of 30 and greater.

THE ENERGETIC particles in cosmic rays interact with meteorites in space and produce a variety of products. Many of these products can be measured and identified as having been made by the cosmic rays. Most cosmogenic products in meteorites are made by the high-energy galactic cosmic rays, although occasionally products made by solar energetic particles are observed. These cosmogenic products include radionuclides with a wide range of half-lives, stable isotopes rare in meteorites (such as noble gases), radiation-damage tracks of heavy cosmic-ray nuclei, and trapped electrons observable by thermoluminescence. These cosmogenic products are made by known processes, and their concentrations and distributions in meteorites can be fairly well predicted. The cosmic-ray record in meteorites show that cosmic rays in the past are not much different from those today, although cosmic-ray intensities have varied over time periods of $\sim$10 to $\sim 10^8$ years. Meteorites have been exposed to cosmic rays for periods ranging from about 10^4 to over 10^9 years, and some fell to the Earth more than 10^6 years ago. The cosmic-ray exposure ages for some classes of meteorites cluster in groups, which suggest that those meteorites were all produced by an event at that time. The exposure histories of some meteorites are complex. These exposure records give information on the recent evolution of meteorites in space.

I. INTRODUCTION

A. General

The solar system is filled with more than just the planets and other large objects such as asteroids and comets. It contains many small stony, iron, or stony-iron objects called meteoroids, which are referred to as meteorites if they survive passage through the atmosphere and reach the Earth's surface. Also moving through the solar system is a variety of energetic particles, the cosmic rays. Most of these cosmic-ray particles possess sufficient energy (above about 10^6 eV) that they can penetrate and interact with matter. These interactions produce atomic displacements and ionization in solid media; they also can induce nuclear reactions. Many of these cosmic-ray-produced effects persist for long periods of time and can be observed in meteorites and identified as being the products of cosmic-ray interactions. The collective ensemble of these cosmic-ray-produced effects in a meteorite forms a record that provides us with information about the cosmic-ray interactions and about the histories of both the cosmic rays and the meteorites.

There are wide ranges for the mineralogies and chemistries of meteorites. Within each of the three major classes of meteorites (stones, which are silicate rich; irons, which are metal rich; and stony-irons, which are approximately half silicates and half metal), there are a number of distinct types. There are two main categories for stones: chondrites, which are fairly primitive and named after the spherical or ellipsoidal chondrules that they contain, and achondrites, which have chemically evolved relative to chondrites.

A meteorite is considered a "fall" if it was observed to hit the Earth; otherwise it is called a "find," and the length of time that it was on the Earth's surface, its "terrestrial age," is unknown. Stony meteorites are normally hard to detect among terrestrial rocks and are fairly rapidly destroyed by weathering (in $\sim 10^4$ years as determined from cosmogenic nuclides), so most stones are falls (except for those found in unusual regions of the Earth). The converse is true for iron meteorites; most are finds that fell a long time ago. Since 1970, a large number of meteorites of many types have been found on several ice fields in Antarctica. Since 1990, many meteorites have also been collected in deserts such as the Sahara.

During its passage through the Earth's atmosphere, a meteorite is subjected to high pressures and its outermost layers are substantially heated. The pressures often cause a meteorite to break into fragments. The heated outer layers are removed by a process called ablation. Both processes affect the meteorite's geometry and make the study of a meteorite's cosmic-ray record more difficult. Meteorites with low-ablation regions or that fell as a single or mainly in one piece are rare but are valuable in studies of the interactions of cosmic rays with meteoroids.

Here, we shall consider only two types of energetic particles that irradiate meteoroids in space: the galactic cosmic rays (GCR), which come from outside the solar system, and solar energetic particles or "solar cosmic rays" (SCR), which are produced irregularly by solar

activity. We shall not discuss the irradiation of meteorites by the low-energy particles in the solar wind, although some meteorites do contain a record of solar-wind particles. Both the GCR and SCR particles are mainly protons, about 10% α particles (^{4}He nuclei), and 1% heavier nuclei.

The nuclei heavier than about calcium ($Z = 20$) produce tracks in certain meteoritic minerals. The protons and α particles induce a variety of nuclear reactions with atomic nuclei in meteorites, many of which produce nuclides, such as radionuclides and noble-gas isotopes, that are not normally present in a meteorite. All primary and most secondary cosmic-ray particles are ionizing radiations that also produce effects observable by thermoluminescence. The cosmic-ray particles penetrate at most a few meters into a meteoroid before they are stopped by ionization energy losses or removed by nuclear reactions. Thus, meteoritic material that was shielded by more than a few meters of solid matter contains essentially no record of the cosmic rays, and a cosmic-ray record only tells us about a meteorite's most recent history.

A wide variety of techniques are used to study the cosmic-ray record of a meteorite. A number of models predict the concentrations of these cosmogenic products expected in meteorites and are used to interpret the record. Meteorites with simple histories in space can be used to study the nature of the cosmic rays in the past. Temporal variations in the intensities of the GCR particles have been observed for periods ranging from over an 11-year solar cycle to over the past 10^9 years. The cosmic-ray record can be used to infer the size and shape of the preatmospheric body of a meteorite. It also can tell us how long a meteorite was exposed to the cosmic rays. Often the cosmogenic products were made both before and after a meteoroid was altered in space by a collision with another object.

B. History of the Field

Although the cosmic radiation was first detected in 1911 by V. Hess from balloon flights with ionization chambers, its exact nature was not known for several decades. By the late 1930s, it was known that the cosmic rays are atomic nuclei moving at high energies. In the late 1940s, W. Libby and co-workers established the use of cosmogenic radiocarbon, ^{14}C, to date terrestrial samples. The activities of ^{14}C measured in objects of known ages agreed well with predicted values, showing that the intensities of the cosmic rays had been fairly constant over the past several thousand years. Also in the 1940s, helium was measured in a number of iron meteorites, but it was assumed that all of the helium was made by the radioactive decay of uranium and thorium. C. Bauer argued in a short note published in 1947 that most of the helium was produced by cosmic radiation. Soon F. Paneth and others measured that about 20% of the helium in iron meteorites was ^{3}He, thus confirming the origin of the helium as the product of nuclear reactions between the energetic particles in the cosmic rays and the iron. At this time, accelerators were starting to produce protons with energies of a few GeV, and nuclear scientists were systematically studying spallation reactions similar to those that make helium in iron meteorites.

In the decade after Paneth's initial measurements of cosmogenic helium in iron meteorites, the study of cosmogenic nuclides in meteorites advanced rapidly. Cosmogenic noble-gas isotopes other than helium were measured in several stony and iron meteorites. Then a number of cosmogenic radionuclides were observed in meteorites, starting with the detection of 12.3-year ^{3}H by low-level gas counting techniques. Shortly thereafter, other cosmogenic radionuclides, such as ^{26}Al, ^{10}Be, and ^{60}Co, were observed in meteorites, and the cosmogenic pair of ^{3}H/^{3}He was used to determine the exposure ages of meteorites.

Around 1960, a number of systematic studies of cosmogenic nuclides in meteorites were done. Noble-gas isotopes were measured in many pieces from slabs of several iron meteorites, such as Carbon and Grant. Many radionuclides were measured in several freshly fallen meteorites, such as the iron Yardymly (then called Aroos) and the stone Bruderheim. Several research groups also measured noble-gas isotopes in a suite of meteorites. New techniques to measure cosmogenic radionuclides, such as nondestructive γ-ray spectroscopy, were developed.

In parallel with the rapid growth in cosmogenic-nuclide measurements, a number of theoretical models were developed. Simple models in which the primary cosmic-ray particles are exponentially attenuated and secondary particles are produced and removed were used by several investigators, such as P. Signer and A. Nier, who applied such a model to their noble-gas data for the Grant iron meteorite. Numerous experiments at accelerators in which thin or thick targets were bombarded with high-energy protons established production ratios for many cosmogenic nuclides. J. Arnold, M. Honda, and D. Lal estimated the energy spectrum of primary and secondary cosmic-ray particles in iron meteorites and calculated production rates of cosmogenic nuclei using cross sections for many reactions.

These measurements and models for cosmogenic nuclides in meteorites were applied to a number of studies, such as the constancy of cosmic rays over time. The ratio of the measured radionuclide activities to the calculated values showed no systematic trends for half-lives less than a million years. H. Voshage and H. Hintenberger found that the ^{40}K/^{41}K ratios measured in iron meteorites were inconsistent with the ratios for other radioactive/stable pairs of

cosmogenic nuclides, implying that the fluxes of cosmic rays have been higher over the past 10^6 years than over the past 10^9 years. E. Fireman, R. Davis, O. Schaeffer, and co-workers used the measured $^{37}Ar/^{39}Ar$ ratios in studies of the spatial variation of cosmic rays between 1 and about 3 astronomical units (AU) from the sun (the region of space in which most meteorites probably traveled). A variety of pairs of radioactive and stable nuclides, such as $^3H/^3He$ and $^{39}Ar/^{38}Ar$, were used to determine the lengths of times that meteorites were exposed to cosmic rays.

During the 1960s, additional measurements were made and the ideas used to interpret the observations were refined. Measurement techniques for the tracks produced in certain minerals by heavy nuclei were developed and applied to meteorites. Some meteorites, those with high concentrations of trapped gases and tracks, were realized to have been exposed to energetic solar particles on the surface of some parent object. The orbits of three stony meteorites, Pribram, Lost City, and Innisfree, were accurately determined by several photographic networks, and all had aphelia in the asteroid belt and perihelia near 1 AU. Bombardments were done at accelerators to simulate the cosmic irradiation of meteorites and used to predict the profiles for the production of nuclides in meteorites.

In the early 1970s, the studies of cosmogenic nuclides in meteorites declined as most investigators were studying lunar samples. Techniques for studying nuclear tracks were developed and used to study the irradiation history of meteorites. Some new methods for measuring cosmogenic nuclides were perfected using lunar samples. The lunar-sample studies confirmed meteoritic results about the galactic cosmic rays and gave us our first detailed knowledge of cosmogenic nuclides produced by the solar cosmic rays. In the late 1970s, interest in the studies of meteorites increased, especially for stones. Measurements of the distributions of the cosmic rays in the solar system by various satellites helped in interpreting the meteoritic cosmic-ray record.

In the 1980s, it was recognized that some meteorites have been ejected from planetary objects, the Moon and Mars. Many meteorites were being recovered from ice fields in Antarctica and from several deserts and arid regions around the world, which greatly expanded the numbers of meteorites available for study. Thermoluminescence was now being routinely used for many studies. Cosmogenic nuclides were being measured in samples much smaller than previously possible because of the use of improved or new measurement techniques (such as accelerator mass spectrometry). By the end of the 20th century, the study of the cosmic-ray records of meteorites was a mature field with gradual but steady advances in all of its aspects.

II. COSMIC RAYS AND THEIR INTERACTIONS WITH METEORITES

A. Nature of Cosmic Rays

In the solar system nuclei move with a variety of energies and compositions. Energies range from slow-moving atoms and molecules to highly relativistic nuclei in the galactic cosmic rays. Here we are only considering the cosmic rays, not the lower energy particles such as those in the solar wind. Table I summarizes the typical mean fluxes and energies of the particles in the two types of cosmic rays, the galactic cosmic rays (GCR) and the solar cosmic rays (SCR), which are often called solar energetic particles. The nuclei in both types of cosmic rays are mainly protons and α particles (with a proton/α-particle ratio of about 10–20), with about 1% heavier nuclei ($Z \geq 3$). The elemental distribution of these heavier nuclei tends to cluster in a number of groups, such as one for carbon, nitrogen, and oxygen, one near magnesium through calcium, and one around iron.

The temporal and spatial distribution in the inner solar system (out to about 5 AU), where most meteoroids orbit, of both types of cosmic-ray particles are strongly influenced by interplanetary magnetic fields that originate at the Sun. Earth-based observations established the role of the Sun in controlling the variations of the cosmic rays. The Sun's activity generally varies with an 11-year period. During periods of maximum solar activity, sunspots on the solar surface are common and many particles and fields are emitted from the Sun. The Sun is relatively quiet for the other half of an 11-year solar cycle. About once every 200 years, the Sun goes through a period of several decades or more where its activity level remains either very high or very low. From studies of sunspots and of the terrestrial record of ^{14}C, we know that the last such solar activity anomaly was a quiet period from about 1645 to 1715, the Maunder Minimum. Recently, experiments on many spacecraft have provided important results on the nature of cosmic rays and their distribution in space.

The Sun is an important source of energetic particles in the inner solar system. These particles are produced as

TABLE I Typical Energies, Fluxes, and Interaction Depths of Cosmic-Ray Particles

Radiation	Energies (MeV/nucleon)	Mean flux (particles/cm^2/s)	Depth (cm)
GCR protons and α particles	100–3000	3	0–100
GCR nuclei, $Z > 20$	100–3000	0.03	0–10
SCR protons and α particles	5–100	~100	0–2
SCR nuclei, $Z > 20$	1–50	~1	0–0.1

a result of solar activity, such as coronal mass ejections. Since 1942, more than 100 particle-accelerating events have been observed near the Earth. Most particle-emitting events occur when the activity level of the Sun is high. During the parts of a typical 11-year solar cycle when solar activity is low, very few SCR particles are produced. Near the Earth, the fluences of protons with $E > 10$ MeV during SCR-producing events range from below 10^5 to over 10^{10} protons/cm^2. Most SCR particles have energies below 100 MeV/nucleon, although some solar particle events have many nuclei with energies above 1 GeV/nucleon. From measurements of SCR-produced radionuclides in the surface layers of lunar rocks, the average fluxes of solar protons with energies above 10 and above 100 MeV over the past few million years are known to be about 70 and 3 protons/(cm^2 s), respectively. Although solar particle events are fairly rare, the average fluxes at 1 AU of SCR particles with energies between 10 and 100 MeV/nucleon greatly exceed those for GCR particles with the same energies. Only above about 500 MeV/nucleon does the GCR dominate. The average fluence of SCR particles probably varies with distance from the Sun according to an inverse-square law, so the average flux of SCR particles to which most meteorites are exposed at 2–3 AU is less than that observed near the Earth.

The GCR particles originate far from the solar system. The sources of the energy that accelerated most of them to high velocities probably involve supernovae, although the exact mechanisms for the acceleration are not known. As the GCR particles diffuse or are transported to the solar system, additional acceleration or other interactions can occur. The transit times of most GCR particles from their sources to the solar system are $\sim 10^7$ years and during that time they pass through $\sim$5–10 g/cm^2 of matter. As these GCR particles enter the solar system, their spectrum is modulated by the interplanetary magnetic fields, which originate at the Sun. Solar modulation is the dominant source of GCR variability to which meteorites and the Earth are exposed. Over a typical 11-year solar cycle, the flux of lower energy GCR particles ($E \sim 100$ MeV/nucleon) changes by about an order of magnitude, and the integral flux of GCR particles above 1 GeV/nucleon varies by about a factor of 2. Even during a typical solar minimum, the intensities of GCR particles are reduced compared to what they are in interstellar space near the solar system. The ^{14}C record on Earth showed that the GCR-particle fluxes were quite high during the Maunder Minimum. Measurements by charged-particle detectors on the Pioneer and Voyager spacecraft that flew past Jupiter showed that the fluxes of GCR protons with $E \geq 80$ MeV increased with distance from the Sun by about 2–3% per AU. Thus, in the ecliptic plane, orbital variations affect the production rates of cosmogenic products much less than the cosmic-ray changes over an 11-year solar cycle.

B. Interaction Processes and Products

The energy, charge, and mass of a cosmic-ray particle and the mineralogy and chemistry of the meteorite determine which interaction processes are important and which cosmogenic products are formed. Energetic nuclear particles interact with matter mainly in two ways: ionization energy losses and reactions with the nuclei. All charged particles continuously lose energy by ionizing the matter through which they pass. Some of the radiation damage produced during ionization by cosmic-ray particles is accumulated in meteorites and can be detected as thermoluminescence (TL). In a number of meteoritic minerals, such as olivine, the paths traveled by individual nuclei with atomic number (Z) above about 20 and with energies of the order of 0.1 to 1 MeV/nucleon contain so much radiation damage that they can be etched by certain chemicals and made visible as tracks. The nuclear reaction between an incident particle and a target nucleus generally involves the formation of new, secondary particles (such as neutrons, protons, pions, and γ rays) and of a residual nucleus that is usually different from the initial one.

Low-energy and high-Z nuclei are rapidly slowed to rest by ionization energy losses. High-energy, low-Z particles lose energy more slowly and usually induce a nuclear reaction before they are stopped. A 1-GeV proton has a range of about 400 g/cm^2 and a nuclear-reaction mean free path of $\sim$100 g/cm^2, so only a few percent go their entire range without inducing a nuclear reaction. Few GCR particles penetrate deeper than about 1000 g/cm^2 (the thickness of the Earth's atmosphere or about 3 meters in a stony meteorite) because they are removed by nuclear reactions or stopped by ionization energy losses. Only a few particles, such as muons, reach such depths, and it is hard to detect the few interaction products made there. Only in the few meters near a meteorite's surface is a cosmic-ray record readily determined. Depths are often expressed in units of g/cm^2, which is the product of the depth in centimeters times the density in g/cm^3, because it is the areal density of nuclei that is important for the main nuclear interactions that occur in meteorites. [Densities (in g/cm^3) of meteorites range widely: about 2.2 for CI carbonaceous chondrites, 3.1 to 3.4 for most achondrites, 3.4 to 3.8 for ordinary chondrites, about 5 for stony-irons, and near 7.9 for iron meteorites.]

Because of the variety of the cosmic-ray particles and of their modes of interactions, the effective depths of the interactions and their products vary considerably, as noted in Table I. This diversity in the types of products and their depths is very useful in studying the cosmic-ray record

of a meteorite. The meteorite that is found on the Earth's surface is not the same object that was exposed to cosmic rays in space. Meteoroids can be suddenly and drastically changed in space by collisions, which remove parts of the old surface and produce new surfaces that usually have not been exposed to the cosmic rays. Micrometeoroids and solar-wind ions gradually erode a meteorite's surface. In passing through the Earth's atmosphere, the outer layers of a meteoroid are strongly heated and removed, a process called ablation. The meteoroid usually fragments into several pieces. The cosmic-ray record of a meteorite is often used to determine its preatmospheric history.

1. Thermoluminescence

Thermoluminescence (TL) in meteoritic minerals is a consequence of the filling of electron traps when electrons are excited into the conduction band by ionizing radiation and some become trapped in metastable energy states. The concentration of trapped electrons is a function of the ionization rate, the nature of the minerals involved, and their temperature record. Thermoluminescence gets its name because light is emitted when a sample is heated. The light is produced when the electrons are thermally released from the traps and combine with luminescence centers. The TL from ordinary chondrites is mainly in the wavelengths near blue (about 470 nm). The natural radioelements, uranium, thorium, and potassium, expose ordinary chondrites to doses of about 10 mrads per year. (Natural radioactivity is the source of radiation that is used in determining when certain terrestrial samples were last heated to a high temperature.) The GCR is the dominant source of ionizing radiation in almost all meteoritic minerals, about 10 rads/year. The natural TL in a meteoritic mineral, like the activity of a cosmogenic radionuclide, is usually at an equilibrium value where the rate for trapping electrons is equal to the rate for thermal draining.

The TL sensitivity of meteoritic samples can vary greatly and depends on bulk composition, the abundance and composition of glasses, and metamorphism. (D. Sears and co-workers have used the wide ranges of TL sensitivity in primitive types of ordinary chondrites to determine the amount of their metamorphic alterations.) Thus it is useful to normalize the natural TL observed in a sample to the TL induced by an artificial dose of radiation. Comparisons of the TL observed in natural samples to that produced artificially in the same sample show that electrons in traps that can be released by temperatures below about 200°C are relatively rare in meteorites. Below about 200°C to 300°C, the retention of TL varies widely among meteorites. The high-temperature ($T \sim 400$°C) TL observed in most meteorites is usually the same as that from an equivalent dose of 10^5 rads, which is the level acquired from the GCR in about 10^4 years.

2. Tracks

Unlike nuclear reactions, which occur in all of the phases of meteorites, observations of the solid-state damage due to charged-particle irradiation is almost always limited to the crystalline dielectric phases of the minerals present. This solid-state damage can be produced by energetic heavy nuclei from a variety of sources, such as fission, the solar wind, and cosmic rays. Here, we are primarily concerned with the damage paths produced in certain meteoritic minerals by cosmic-ray nuclei with $Z \geq 20$. These nuclei are divided into two groups: $Z = 20$ to 28, termed the iron or the VH group, and $Z \geq 30$, called the VVH group. This classification is based primarily on the observed abundances of these nuclei in the cosmic rays. In the cosmic rays, the VVH group of nuclei (mainly $Z = 30$ to 40) is less abundant than the VH group by a factor of $\sim$700 for energies above 500 MeV/nucleon. This ratio is similar to abundance ratios for these elements in the Sun and meteorites. Nuclei with $Z > 40$ are rare in the cosmic rays, about 5×10^{-5} per iron nucleus.

In dielectrics, when the ionization along the path traveled by the particle exceeds a certain critical value, the radiation damage has altered the material sufficiently that the trails can be dissolved and enlarged by suitable chemical treatment. These chemically developed holes are cylindrical or conical, can be seen with an optical microscope, and are called tracks. In meteoritic and other natural minerals, this ionization threshold is exceeded only by nuclei of calcium and heavier elements ($Z \geq 20$). As Z becomes higher, the range of energies that have ionization rates above the critical value is greater. The VH and VVH groups of nuclei form chemically etchable tracks near the end of their range, where their ionization rates are the largest. Iron nuclei ($Z = 26$) with energies between about 0.1 and 1 MeV/nucleon are capable of producing tracks in meteoritic minerals; the corresponding energy range for Kr nuclei ($Z = 36$) is about 0.1 to 5 MeV/nucleon.

The length of the etchable tracks depends on the mineral. Cosmic-ray tracks are mainly studied in three common silicate minerals: olivines, $(Mg,Fe)_2SiO_4$; pyroxenes, $(Mg,Fe)SiO_3$; and feldspars, solid solutions of $CaAl_2Si_2O_8$ and $NaAlSi_3O_8$. The maximum etchable lengths of fresh Fe and Zn tracks in olivine are about 13 and 60 μm, respectively. Fossil tracks of these nuclei in meteorites are somewhat shorter because of partial annealing of the radiation damage along the track. However, tracks of VH nuclei do not appear to fade appreciably over long periods of time, so they can be used to determine exposure ages of meteorites. As the length of a track varies rapidly

with the atomic number of the nucleus, the Z of the nucleus forming a track usually can be determined to within about ± 2 charge units, and tracks made by VVH-group nuclei are clearly distinguished from those of VH nuclei.

Profiles of track density versus depth have been measured in a number of materials exposed in space, including meteorites, lunar samples, and a glass filter from the Surveyor III camera returned by the Apollo 12 astronauts. These measurements and theoretical calculations have been used to obtain curves for the production rates of tracks as a function of meteorite radii and sample depth. The track production profile as a function of depth in a large extraterrestrial object is shown in Fig. 1. The tracks between about 10^{-3} and 0.1 cm are made mainly by heavy SCR nuclei that have energies below about 10 MeV/nucleon; the deeper tracks are made by GCR nuclei with $E > 100$ MeV/nucleon after they have been slowed to energies of about 1 MeV/nucleon. The very surface layers of meteorites, which contain the tracks of heavy SCR nuclei, are almost always removed by ablation when passing through the Earth's atmosphere. The density of cosmic-ray tracks drops rapidly with depth, which make tracks excellent indicators of a sample's preatmospheric depth.

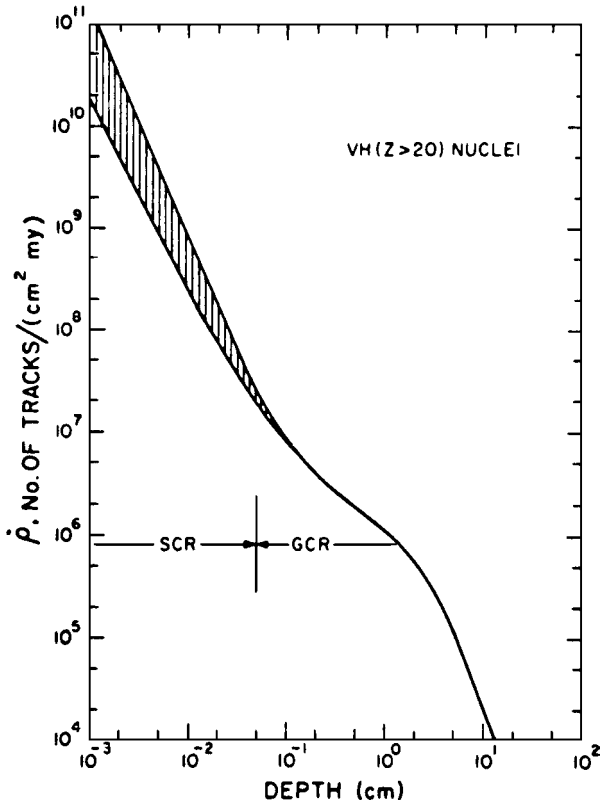


FIGURE 1 Calculated production rates of heavy-nuclei tracks as a function of depth in a large stony meteorite. The shaded region represents the uncertainties in the fluxes of VH nuclei in the solar cosmic rays.

3. Nuclides

Cosmic-ray particles can induce a wide variety of nuclear reactions with any nucleus in a meteoroid. In a nuclear reaction, a particle such as a proton, neutron, pion, or α particle interacts with a target nucleus. A large variety of nuclear interactions is possible. The simplest one is elastic scattering, where the incident particle and the target nucleus are unchanged after the interaction, and only the direction and the energy of the particle are changed. Elastic scattering is the only way that low-energy (below about 0.5 MeV) neutrons can be slowed, as neutrons have no charge and hence cannot be slowed by ionization energy losses. Inelastic scattering is like elastic scattering, except that the residual nucleus is left in an excited state. Elastic and inelastic scattering are important in the transport of particles inside a meteorite, but they very seldom leave a record. The other types of nuclear reactions contribute to the cosmic-ray record only when they produce a product nucleus that can be distinguished from the nuclei normally present in the meteorite. Cosmogenic nuclei in meteorites are very scarce: over the $\sim 10^8$-year exposure age of a meteorite, only about one in every 10^8 nuclei will undergo a nuclear transformation. These cosmogenic nuclides are usually ones that are radioactive or those that are very rare in a meteorite, such as the isotopes of the noble gases.

In a piece of meteorite exposed to cosmic-ray particles, the radius- and depth-dependent production rate of a cosmogenic nuclide, $P(R, d)$, is

$$P(R, d) = \sum_j N_j \sum_i \int \sigma_{ij}(E) F_i(E, R, d)\, dE, \quad (1)$$

where N_j is the abundance of target element j, i indicates one of the primary or secondary particles that can induce nuclear reactions, $\sigma_{ij}(E)$ is the cross section as a function of energy for particle i making the product from element j, and $F_i(E, R, d)$ is the flux of particle i as a function of energy E, meteoroid radius R, and sample depth d. The particles that can induce nucler reactions are those primary ones in the cosmic rays, such as protons and α particles, or the secondary ones made by nuclear reactions inside a meteorite, such as neutrons or pions. The most important parts of Eq. (1) are the expressions for the fluxes of cosmic-ray particles, $F_i(E, R, d)$, that induce the nuclear reactions, and the cross sections, $\sigma_{ij}(E)$, for making the cosmogenic nuclide. The basic shapes for $F_i(E, R, d)$ are fairly well known, especially for the primary cosmic-ray particles at higher energies. The numbers of secondary particles and variety of residual nuclei made by nuclear reactions depend on the energy of the incident particle.

TABLE II Cosmogenic Nuclides Frequently Measured in Meteorites

Nuclide	Half-life[a] (yr)	Main targets
^{3}H	12.3	O, Mg, Si, Fe
^{3}He, ^{4}He	S	O, Mg, Si, Fe
^{10}Be	1.5×10^{6}	O, Mg, Si, Fe
^{14}C	5730	O, Mg, Si, Fe
^{20}Ne, ^{21}Ne, ^{22}Ne	S	Mg, Al, Si, Fe
^{22}Na	2.60	Mg, Al, Si, Fe
^{26}Al	7.1×10^{5}	Si, Al, Fe
^{36}Cl	3.0×10^{5}	Fe, Ca, K, Cl
^{36}Ar, ^{38}Ar	S	Fe, Ca, K
^{37}Ar	35 days	Fe, Ca, K
^{39}Ar	269	Fe, Ca, K
^{40}K	1.28×10^{9}	Fe
^{39}K,^{41}K	S	Fe
^{41}Ca	1.03×10^{5}	Ca, Fe
^{46}Sc	83.8 days	Fe
^{48}V	16.0 days	Fe
^{53}Mn	3.7×10^{6}	Fe
^{54}Mn	312 days	Fe
^{55}Fe	2.73	Fe
^{59}Ni	7.6×10^{4}	Ni
^{60}Co	5.27	Co, Ni
^{81}Kr	2.3×10^{5}	Rb, Sr, Zr
^{78}Kr, ^{80}Kr, ^{82}Kr, ^{83}Kr	S	Rb, Sr, Zr
^{129}I	1.57×10^{7}	Te, Ba, La, Ce
$^{124-132}$Xe	S	Te, Ba, La, Ce, (I)

[a] S denotes that the nuclide is stable.

Most of the cosmogenic nuclides that have been studied in meteorites are given in Table II. They include radioactive nuclides with half-lives ranging from a few days (^{48}V and ^{37}Ar) to millions of year (^{53}Mn and ^{40}K) and isotopes, especially the minor ones such as ^{3}He and ^{21}Ne, of the noble gases. The most commonly measured cosmogenic nuclides include ^{3}He, ^{10}Be, ^{21}Ne, ^{22}Na, ^{26}Al, ^{36}Cl, ^{38}Ar, and ^{53}Mn. In iron meteorites, many stable nuclei lighter than iron, such as ^{45}Sc and the potassium isotopes, can be identified as cosmogenic, usually because they have a different isotopic distribution than normally present in nature. In most pieces of meteorites freshly exposed to cosmic rays by a collision, the concentrations of cosmogenic nuclides are very low. As the cosmogenic-nuclide production rates at a given position in a meteoroid are fairly constant, the concentrations of cosmogenic nuclides steady build up. After several half-lives, the activity of a cosmogenic radionuclide will approach an equilibrium value where its rate of decay equals its production rate. Stable isotopes, such as ^{3}He and ^{21}Ne, continue to accumulate and can be used to determine the total length of time that a meteorite was exposed to cosmic rays, if the production rates of these isotopes are known. However, certain processes, such as shock or other heating events, can liberate the lighter noble gases such as ^{3}He. When a meteorite falls to the Earth's surface, it is exposed to a very low flux of cosmic-ray particles, and the activities of the radionuclides start to decrease at rates proportional to their half-lives. A meteorite's terrestrial age, the length of time it has been on Earth, can often be inferred from cosmogenic radionuclides.

The distribution of cosmogenic nuclides made by SCR particles are much different from those made by the galactic cosmic rays. The relatively low-energy solar protons and α particles are usually stopped by ionization energy losses near the surface of a meteoroid. The SCR particles that induce nuclear reactions usually produce few secondary particles and the product nucleus is close in mass to the target nucleus. An example of an SCR-induced reaction is ^{56}Fe$(p, n)^{56}$Co, the production of ^{56}Co when a proton enters a ^{56}Fe nucleus and a neutron is emitted. The fluxes of SCR particles as a function of depth can be calculated accurately from ionization-energy-loss relations, so a nuclide's production rates can be predicted well if the cross sections for its formation are known. Like densities of heavy-nuclei tracks, the activities of SCR-produced nuclides decrease rapidly with depth, most being made within a few centimeters of the meteoroid's surface. However, ablation seldom leaves a significant amount of SCR-produced nuclides in a meteorite.

The GCR particles producing nuclear reactions can roughly be divided into four components: high-energy ($E > 1$ GeV) primary particles, medium-energy (about 0.1 to 1 GeV) particles produced partially from the first component, a low-energy group ($E < 100$ MeV) consisting mainly of energetic secondary neutrons, and slow neutrons with energies below about 1 keV. The fluxes of the high-energy primary GCR particles decrease exponentially with depth as they are removed by nuclear reactions. The numbers of secondary particles as a function of depth build up near a meteoroid's surface, where most of them are made, but eventually decrease roughly exponentially. In a very big meteoroid or on the Moon, there are about 13 neutrons produced per second per cm^2 of surface area (the solar-cycle-averaged rate). This neutron production rate compares to an average omnidirectional flux of about 3 particles/(cm^2 s) for the GCR primaries in space, showing the importance of the large cascade of nuclear reactions in a meteoroid that produce numerous secondary particles. The total fluxes of GCR primary and secondary particles for several depths in spherical stony meteoroids of different radii are shown in Fig. 2. The energy spectrum of primary GCR particles, labeled (0, 0), is also shown in Fig. 2. Even in a meteoroid of radius 40 g/cm^2 (about

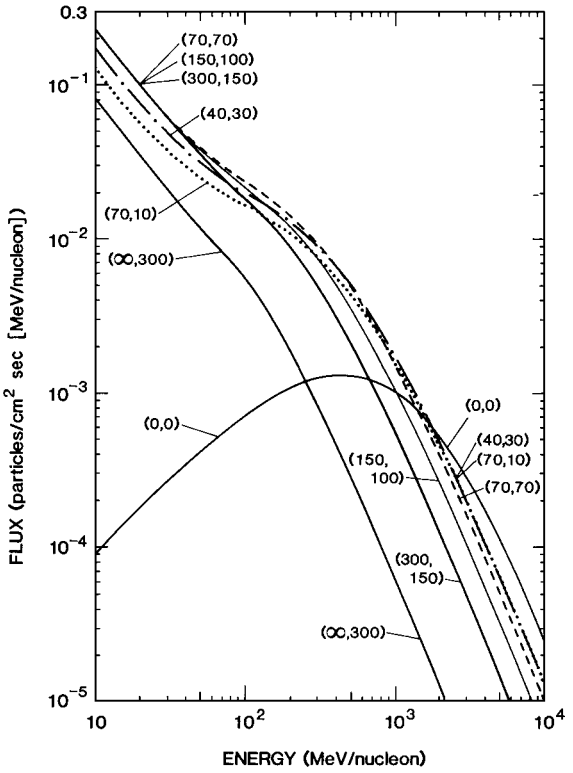


FIGURE 2 Fluxes of both primary and secondary cosmic-ray particles as a function of particle energy for certain depths in stony meteoroids of various radii. Numbers in parentheses are the radius and depth (in g/cm^2). The curve labeled (0, 0) is that for the primary cosmic rays averaged over an 11-year solar cycle.

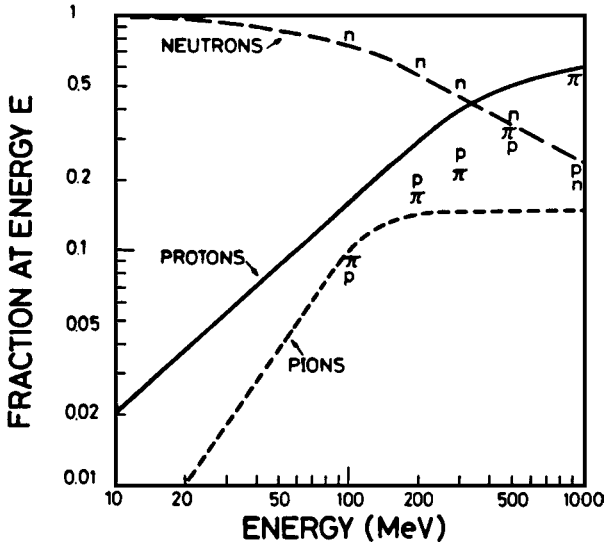


FIGURE 3 The fractions of neutrons, pions, and protons deep (~100 g/cm^2) inside meteorites as a function of energy. The lines are based on lunar Monte Carlo calculations by T. Armstrong and R. Alsmiller; the symbols are from meteoritic estimates by J. Arnold, M. Honda, and D. Lal.

11 cm), there is a significant flux of low-energy particles, mainly secondary neutrons. This dominance of neutrons at lower energies is shown in Fig. 3, which gives the fractions of neutrons, protons, and pions deep inside a piece of extraterrestrial matter.

Neutrons slowed to energies of keV or eV can produce nuclides by neutron-capture reactions, in which only one or more γ rays are emitted after a nucleus captures a neutron, such as ^{59}Co$(n, \gamma)^{60}$Co. However, neutrons can be moderated to such low energies by many scattering reactions only if a meteoroid has a radius of more than about 20 cm. Nuclides made by neutron-capture reactions in meteorites include ^{60}Co, ^{59}Ni, and ^{36}Cl (98.1% of which decays into ^{36}Ar).

Most GCR-induced reactions involve incident particles with energies of about 1 MeV or more, emit one or more nucleons either individually or in clusters such as α particles, and are called spallation reactions. Such products are often referred to as "spallogenic" nuclides. Sometimes the spallation reactions are divided into low-energy and high-energy groups. High-energy spallation reactions involve particles with energies above about 100 MeV, produce numerous secondaries, and can make, in relatively low yields, many different product nuclides. Examples of high-energy spallation reactions are ^{16}O$(p, X)^3$He or ^{24}Mg$(p, X)^{10}$Be, where X can be any one of a large number of possible outgoing particle combinations. Low-energy spallation reactions usually involve particles with energies below 100 MeV and can produce certain nuclides in high yields because both the fluxes of particles and the cross sections [see Eq. (1)] are relatively large. The reaction ^{24}Mg$(n, \alpha)^{21}$Ne is such a low-energy reaction and is the major source of ^{21}Ne in most stony meteorites.

Because the distributions of low-energy and high-energy GCR particles are not the same for all locations inside meteoroids, the profiles of production rate versus depth for different cosmogenic nuclides can vary. A high-energy product such as ^{3}He (see Fig. 4) has a profile that is fairly flat near the surface, whereas a low-energy product, like ^{21}Ne (shown in Fig. 5), builds up in concentration considerably with increasing depth near the surface. The size and shape of a meteoroid are important in determining the production profiles of cosmogenic nuclides. When a meteoroid's radius is less than the interaction length of GCR particles, about 100 g/cm^2, particles entering anywhere can reach most of the meteoroid, and production rates do not decrease much near the center. In much larger meteoroids, many GCR particles are removed before they get near the center, and production rates for increasing depths decrease from their peak values near the surface. The production profiles for cosmogenic nuclides made by GCR particles vary with depth much less rapidly than do profiles for tracks or SCR-produced nuclides; however, ratios of a

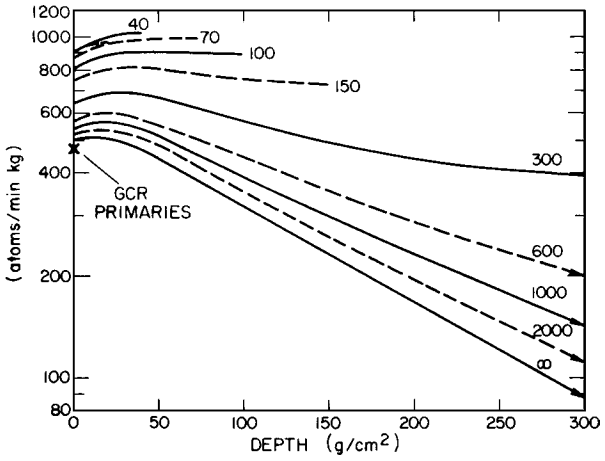


FIGURE 4 Calculated production rates of ^{3}He as a function of meteoroid radius (in g/cm^2) and sample depth in the achondrite Shergotty.

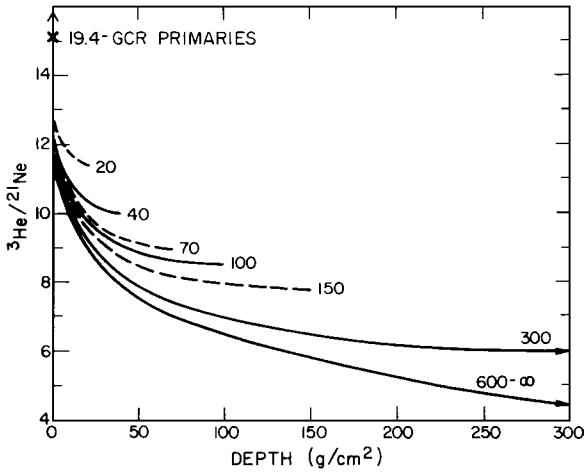


FIGURE 6 Calculated ^{3}He/^{21}Ne production ratios as a function of meteoroid radius (in g/cm^2) and sample depth in the achondrite Shergotty.

high-energy product to a low-energy one, such as ^{3}He/^{21}Ne in Fig. 6 or ^{22}Ne/^{21}Ne, are often used to estimate how much a sample was shielded from the cosmic rays.

As noted in Table II, many cosmogenic nuclides can be made from more than one of the elements that are common in meteorites, so the chemical composition of a sample is often needed to interpret a cosmogenic-nuclide measurement. This diversity of target elements is usually not a problem if one is only studying a single class of meteorites that are chemically homogeneous, such as H-chondrites. However, many cosmogenic nuclides are now studied in very small samples much less than a gram in mass, so chemical variability on that size scale is a concern. Also, when one wants to extend results from one class to another chemical class, elemental production rates for certain cosmogenic nuclides often are needed. Measurements of cosmogenic nuclides in a suite of chemically different meteorites or in mineral separates from one meteorite have been used to infer elemental production rates. Experimental cross sections with pure target elements can also be used to get relative elemental production rates. One phase of meteorites that is fairly simple chemically is the metal phase, which is almost pure iron with ~8% nickel. Metallic iron–nickel pieces occur in many chondrites; these pieces are often separated from the bulk of a chondrite, and cosmogenic nuclides, such as ^{37}Ar and ^{39}Ar, measured in them.

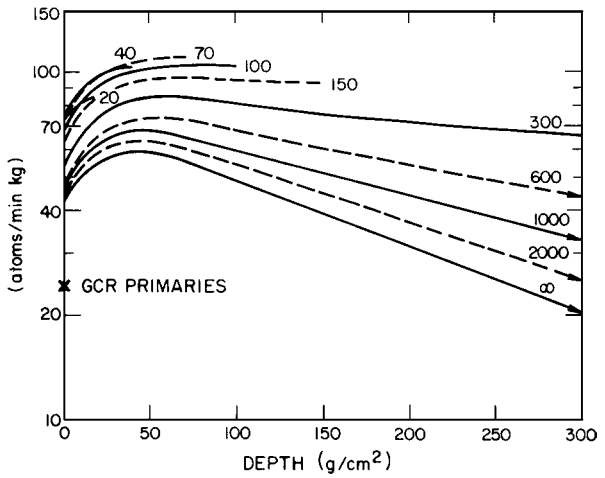


FIGURE 5 Calculated production rates of ^{21}Ne as a function of meteoroid radius (in g/cm^2) and sample depth in the achondrite Shergotty.

III. CALCULATED PRODUCTION RATES

A. Tracks

The densities of the tracks made by heavy cosmic-ray nuclei have been measured in a number of extraterrestrial objects, such as meteorites, lunar samples, and artificial materials exposed in space. These observations, plus a knowledge of the particle fluxes and processes involved in track formation, have resulted in several operational curves for the production rates of tracks as a function of depth in meteoroids of various radii, such as the profile shown for a large-radius object in Fig. 1. The relations for the ranges in meteoritic material of heavy nuclei as a function of their energy are well known. The energy spectra of heavy ($Z > 20$) nuclei in the galactic cosmic rays have been measured by detectors on satellites and inferred from track profiles in meteorites and lunar samples with simple exposure histories. Thus, the production-rate profiles for tracks of GCR nuclei are predicted fairly well.

As the solar cosmic rays are only emitted irregularly from the Sun, there have been only a few chances to

directly observe their spectra in space. Another problem is that the low energies of the SCR heavy nuclei make them hard to detect using instruments on satellites. Each event has its own energy distribution for the SCR particles, so it is hard to determine the long-term-averaged spectrum of heavy SCR nuclei. Lunar rocks with simple exposure histories of short durations have been used by several research groups to infer the spectrum of SCR nuclei with $Z > 20$ averaged over various time periods. The problem here is that the very surface layers in which the SCR nuclei produce tracks are continuously removed by micrometeoroids and solar-wind ions. Typical erosion rates for lunar rocks are of the order of a millimeter per million years. As the tracks from SCR nuclei mainly occur in the top millimeter, the observed track densities very near the surface depend critically on the erosion rate. The range shown in Fig. 1 for the production rates of tracks at depths shallower than 0.1 cm reflects these uncertainties in interpreting such profiles in lunar rocks or in measuring the spectrum of heavy SCR nuclei in space. However, we do know that the tracks from heavy nuclei in the SCR dominate over those from GCR nuclei for depths less than about 0.1 cm. Although such very surface layers of meteorites are removed by ablation, we are still very much interested in tracks of SCR nuclei in meteorites because they occur in grains now inside a meteorite that were once exposed at the very surface of the meteorite's parent body. A steep gradient of track densities over a distance of much less than 0.1 cm is a sign that the tracks were made by SCR nuclei.

B. Nuclides

A number of approaches have been used to predict the rates, ratios, or profiles for the production of cosmogenic nuclides in meteorites. The profiles for the production of trapped electrons observed in thermoluminescence are similar to those for spallogenic nuclides. Experiments at high-energy ($E \sim 1$–6 GeV) proton accelerators have been useful in simulating the cosmic-ray irradiation conditions in meteoroids. Some cosmogenic nuclides are only made by cosmic rays with energies above about 1 GeV where secondary particles are not important, examples being ^{36}Cl, ^{37}Ar, and ^{39}Ar from iron. Cross sections or isotope ratios measured with high-energy protons can be used to predict the production ratios for such nuclides in metal phases of meteorites. The distributions of a number of spallation products have been measured in irradiated targets. When the product nuclide has a mass that is more than about 5 units less than that of the target nucleus (this mass difference is called ΔA), the trends for the product yields behave as a power law with respect to the product mass. This relation for the cumulative yield of all isotopes with mass A, $Y(A)$, is usually expressed as

$$Y(A) = c(\Delta A)^{-k}, \tag{2}$$

where k typically varies between about 2 and 3 is smallest for small meteorites. Measurements of a number of stable cosmogenic nuclides in iron meteorites follow this trend very well. This trend usually fails for very small values of ΔA, such as nuclides made by (p, pn) reactions, and for product nuclides with A less than about 10 (which, like α particles, can be emitted from the excited target nucleus as a fragment by itself). As only a part of the yield for a given A is to a specific isotope, independent yields of radionuclides with mass A are less than $Y(A)$ by a fraction that is related to how close the radionuclide is to the most stable isotope with that mass.

A variety of stationary thick targets have been irradiated with beams of high-energy protons. The distributions of the product nuclides inside such targets varied considerably with depth and with the lateral distance from the beam. High-energy products, such as ^{22}Na from iron, decrease exponentially with depth from very near the target's surface and show little lateral spread. A low-energy product, such as ^{55}Fe from iron, shows a buildup in activity near the surface, reaching its peak at depths of 50 g/cm^2 or more, and then shows an exponential drop in its activity with depth. The lateral spread of low-energy products is very large, reflecting the emission at all angles of the secondary particles that mainly induce such reactions. (In studies of low-energy products, care must be taken in planning thick-target experiments that the target is sufficiently wide to contain all the secondary particles inside the target.)

A thick iron target was bombarded by a collimated beam of 3-GeV protons and the distributions of product nuclides measured by M. Honda. He then transformed the results into that for a hemisphere irradiated by an isotopic flux of particles, an approach valid for large iron meteoroids. In 1967, T. Kohman and M. Bender published a procedure to translate thick-target data to the distribution of nuclides expected inside an isotropically irradiated sphere and applied it to data from several bombardments of thick iron targets. Later B. Trivedi and P. Goel applied the Kohman–Bender model to data for ^{22}Na and ^{3}H in thick silicate targets. These production profiles based on thick-target bombardments have been very useful in studies of cosmogenic nuclides, although they are usually limited to the nuclides measured in the thick target or those made by similar nuclear reactions. Such transformations from thick-target measurements are not valid to very small meteoroids because the cascade of secondary particles in such small objects is not fully developed. In the 1980s, R. Michel and co-workers isotropically irradiated several stone spheres at two accelerators using a machine that translated and rotated the spheres in several directions. Even in a

5-cm-radius sphere, there was a noticeable buildup in the activities of low-energy products.

Several research groups have applied a semiempirical model to measured concentrations of cosmogenic nuclides in meteorites. This approach considers the exponential decrease in the flux of primary cosmic-ray particles and the buildup and decrease in a flux of secondary particles that can produce a specific nuclide. The basic expression of this model for the production rate P of a nuclide as a function of distance x from the surface is

$$P = A_i[\exp(-\mu_p x) - B_i \exp(-\mu_s x)] \qquad (3)$$

where p and s refer to primary and secondary particles, respectively, and $\mu_j = N\rho\sigma_j/A$, j being p or s. Here N is Avogrado's number, ρ is the density, A is the average atomic weight of the meteorite, and μ_j are interaction cross sections for primaries or secondaries (with $\mu_s > \mu_p$). This expression can be integrated for several simple geometries. Values of μ_s and μ_p can be estimated from nuclear systematics or inferred from meteoritic measurements; A_i and B_i are determined from measured profiles. In 1960, P. Signer and O. Nier applied this model to their ^{3}He, ^{4}He, ^{21}Ne, and ^{38}Ar concentrations measured for many locations in a slab of the Grant iron meteorite and used the model to predict spallogenic noble-gas profiles in spherical iron meteorites of any radius. This model has worked quite well in describing the production profiles of noble-gas nuclides in meteorites, in spite of several limitations with the model. It does not allow for energy losses or wide angle scattering of the cosmic-ray particles, processes that do occur in meteorites. It also is limited to two parameters that represent the effective cross sections, μ_s and μ_p. These limitations are less important for high-energy products; thus, the model works fairly well for such nuclides in iron meteorites. In 1990, Th. Graf and co-workers published a version of this model for nuclides in stony meteorites.

Another series of models for the production profiles of cosmogenic nuclides are based on Eq. (1). The most important parts of these models are the expressions for the fluxes of cosmic-ray particles as a function of particle energy E, meteoroid radius R, and sample depth d, $F_i(E, R, d)$. In many versions of this model, all particle types (i) are combined and fluxes are not given for individual particle types. This combination of particle types can be done for most cases as the fractions of various particles usually are not much different from those shown in Fig. 3, being mainly neutrons at lower energies. At higher energies, cross sections for nuclear reactions are less sensitive to the nature of the incident particle. Expressions for the fluxes of cosmic-ray particles can be derived in many ways. For SCR particles, ionization-energy-loss relations can be used to calculate them. For GCR particles at fairly high-energies, the primary spectrum measured in space, $dF/dE = cE^{-2.5}$, is used. In 1961, J. Arnold, M. Honda, and D. Lal used the expression $dF/dE = c(\alpha + E)^{-2.5}$ for $E > 100$ MeV, where α varied with location in the meteoroid. The same authors used a spectrum measured in the Earth's atmosphere for $E < 100$ MeV. R. Reedy and J. Arnold derived spectra for $E < 100$ MeV in the Moon that varied with depth. Cross sections used for the production of a nuclide are measured ones, if possible, or else they can be estimated from nuclear models or other systematics, such as spallation formulas that are extensions of Eq. (2). In most cases where there are no experimental cross sections for neutron-induced reactions, results from proton irradiations are used.

J. Arnold and co-workers first used such a model to calculate cosmogenic-nuclide production rates in iron meteorites. In 1972, Reedy and Arnold extended this model to the Moon. In 1985, R. Reedy applied the lunar flux expressions to stony meteorites, using measured profiles of cosmogenic nuclides to infer the spectral shapes of cosmic-ray particles inside several spherical meteorites. The lunar version of this model has reproduced measurements to depths of about 350 g/cm^2. For meteorites, there were only a few profiles measured for cosmogenic nuclides. The curves shown in Figs. 4–6 were calculated with this model. The profiles calculated for meteoroid radii less than about 40 g/cm^2 or more than about 200 g/cm^2 are less certain than those for intermediate sizes. Although the production rates calculated for many cosmogenic nuclides agree well with measured values, there are some serious disagreements, such as for ^{53}Mn. Such disagreements could be caused by incorrect cross sections, especially for reactions induced by medium-energy neutrons, or poorly estimated particle fluxes.

Production profiles have been calculated from theoretical expressions for nuclear interactions. T. Armstrong and R. Alsmiller used a Monte Carlo code for intranuclear cascades to calculate the distribution of cosmic-ray particles and product nuclides in the Moon. Masarik and Reedy or Michel and co-workers have used Monte Carlo codes to numerically simulate the production of cosmogenic nuclides in a variety of meteorites. Such calculations confirm that neutrons are the most important nuclear particles in making cosmogenic nuclides in all but the very smallest meteoroids and have been used for unusual irradiation geometries (such as ellipsoids) and a wide range of compositions. The bulk composition of an object can affect cosmogenic-nuclide production rates, with rates being higher for most reactions in objects enriched in the heavier elements such as iron.

Rates for neutron-capture reactions have been calculated for the Moon and stony meteorites using several

neutron-transport codes. The major neutron-capture-produced nuclides observed in meteorites are ^{60}Co, ^{59}Ni, and ^{36}Cl (and its decay product ^{36}Ar). (However, for most meteoritic samples, ^{36}Cl is made mainly by spallation reactions.) Given an initial distribution of secondary neutrons, which are made with energies of the order of 1 MeV, these calculations transport the neutrons through the object and follow the scattering reactions that slow the neutrons to thermal energies ($E < 1$ eV). P. Eberhardt, J. Geiss, and H. Lutz first did such calculations for meteorites in the early 1960s. They showed that stones with radii of less than about 30 cm are too small to have an appreciable density of thermal neutrons, as most neutrons escape from such small meteoroids before they can be thermalized. Maximum rates for neutron-capture reactions in meteoroids occur in the centers of stony meteorites with radii of about 80 cm (about 300 g/cm^2). The production rates for neutron-capture reactions vary much more with meteoroid radius and sample depth than do those for spallation reactions (like those in Figs. 4 and 5). This large variability with radius and depth makes neutron-capture products, like track densities, useful in determining sample locations in meteorites.

Production rates, ratios, and profiles of cosmogenic nuclides calculated with all of these various models have been used in studies of meteorites. Each model has its advantages, but each also has some limitations. Most models agree on the basic production profiles that they calculate, giving results that are not much different from those shown in Figs. 4 and 5. These calculated production profiles generally agree well with those measured in meteorites. The model of Signer and Nier is good for cosmogenic noble gases in iron meteorites. There have only been a few profiles for cosmogenic nuclides measured in stony meteorites, such as cores from the Keyes and St. Severin chondrites, a slab of the Knyahinya chondrite, and samples from documented locations on the main mass of the Jilin chondrite. Measurements in lunar cores can be used to test models for the production of nuclides in very big objects, although the lunar cores only extend to depths of about 400 g/cm^2 and have been disturbed by meteoroid impacts on various time and depth scales.

The models based on thick-target irradiations often do not reproduce well the profiles measured in the Moon or large stones such as Jilin. Additional production profiles measured in meteorites with a wide range of preatmospheric sizes and shapes would be useful in testing and developing models. The models all work fairly well for typical-sized meteorites (radii between about 40 and 150 g/cm^2), which is also the region where production rates usually do not change much with radius or sample depth. To a first approximation, a single value for the production rate of a cosmogenic nuclide in any type of meteorite can be used, as often was done in meteoritic studies. A better approach has been to use a measured value to estimate the corrections for production-rate variations because of the meteoroid's size and shape and the sample's location. The isotope ratio ^{22}Ne/^{21}Ne has often been used for such shielding corrections.

Almost all cosmogenic-nuclide production rates used in the studies of meteorites are based on measurements. Because the absolute rates calculated by most models are sometimes not very reliable, the calculated production rates are usually normalized to those measured in meteorites. As the activity of a radionuclide is usually in equilibrium with its production rate, measured activities can directly be used as production rates. Sometimes, as with the thin-target cross sections, only production ratios are determined. A measured activity of a cosmogenic radionuclide and a production ratio can be used to infer the production rate for a stable nuclide, especially if the pair of nuclides are made by similar reactions or if the stable nuclide is the decay product of the radionuclide. Radioactive/stable pairs often used with production ratios include ^{3}H/^{3}He (although ^{3}H measurements show much scatter in meteorites), ^{22}Na/^{22}Ne (with care to correct for variations in ^{22}Na activity over an 11-year solar cycle), ^{36}Cl/^{36}Ar, ^{39}Ar/^{38}Ar, ^{40}K/^{41}K, and ^{81}Kr/^{83}Kr. These radioactive/stable pairs have been frequently used to get exposure ages of meteorites.

Production rates of stable isotopes can also be determined from measured concentrations if the exposure age of the meteorite is known. In meteorites with short exposure ages, the activity of a long-lived radionuclide has often not reached its saturation or equilibrium value. If the activity of a radionuclide is significantly below its equilibrium value and its half-life is known, then the exposure age of the meteorite can be readily calculated. Production rates of several noble-gas isotopes, such as ^{21}Ne in chondrites, have been calculated from measurements for chondrites with exposure ages that were based on the radionuclides ^{10}Be, ^{22}Na, ^{26}Al, ^{53}Mn, or ^{81}Kr. These calculated production rates have used the measured ^{22}Ne/^{21}Ne ratio for shielding corrections, and the nominal production rate refers to a specific ^{22}Ne/^{21}Ne ratio. The rates determined from undersaturation of ^{26}Al are higher by about 50% than those based on the other radionuclides, for reasons that are not well understood but probably represent some previous production of ^{21}Ne.

IV. MEASUREMENT TECHNIQUES

A. Thermoluminescence

The thermoluminescence of a meteoritic sample is the intensity of light emitted as it is heated. The sample is

crushed to a fine powder, magnetic particles are removed, and about 4 mg is placed in a pan and heated in a special chamber. The chamber is filled with an oxygen-free gas. The heating strip usually consists of a nichrome plate, and the temperature is monitored with a thermocouple. The temperature is increased from room temperature to about 550°C at a rate of about 2–8°C/s. (Above about 500°C, the blackbody radiation becomes too strong, although a blue filter has be used to transmit the blue TL from chondrites while absorbing most of the longer wavelength blackbody radiation.) The emitted light is detected with a photomultiplier tube that is shielded from magnetic and electrical fields. A filter is often used to select the wavelength of the light reaching the photomultiplier tube. The plot of the TL light output as a function of the temperature is often called a glow curve.

The sample usually is then exposed to an artificial source of radiation, such as a ^{90}Sr β source, for a test dose of $\sim 10^5$ rads (below the saturation part of the TL versus dose curve). The glow curve for this test dose is a measure of the TL sensitivity of the sample. For each temperature, the product of the test dose (in rads) and the ratio of the natural TL to the artificial TL is called the equivalent dose. Below and above a temperature of about 300–350°C, the character of the TL output from meteoritic samples varies. At higher temperatures, the TL is more stable. At lower temperatures, the intensity of the TL can cover a much larger range of values. The ratio of low-temperature TL to high-temperature TL is sometimes used in studies of meteorites.

B. Tracks

The paths of strong radiation damage that are produced by heavy ($Z > 20$) nuclei in various meteoritic minerals can sometimes be viewed directly by electron microscopy but are usually observed after chemical etching. The use of simple chemicals to attack these radiation-damage paths enlarges the tracks so that they can be viewed with an optical microscope. Usually a fairly thin piece of meteorite or a series of selected mineral grains are mounted on a suitable support, and the top surface is ground and polished. Polished thin sections with thicknesses of a few tens of micrometers are often used for petrological studies of the minerals. For studies of tracks, the thickness must be $\sim$200–300 μm, and the sample mounted on a medium (such as epoxy resin) that can withstand high temperatures and harsh chemicals. The polished thick section is then etched with the appropriate chemicals for a specific length of time to reveal the tracks.

Much work has been done by many researchers in selecting the chemicals and etch times to be used for various materials. Care is needed to etch mainly the radiation damage along the tracks and not the whole surface of the mineral grain. For example, the common meteoritic mineral olivine is often etched using H_3PO_4, oxalic acid, EDTA, water, and NaOH (to adjust the pH) for 2–3 h at 125°C or 6 h at 90°C. The time and temperature are selected so that the tracks are suitably enlarged, but that a minimum of material beyond the radiation damage region is etched. The etched tracks in most meteoritic minerals usually are cones with a narrow angle (a few degrees) at the apex, as the chemical works in from the surface and enlarges the part of the track near the surface more than that at the bottom. The lengths of the etched tracks vary considerably with the energy and atomic number of the heavy nuclei that produced the track, with how much the radiation-damage paths had annealed prior to chemical treatment, and with how much of the track was removed when the grain was polished. Measurements of the complete lengths of etched tracks can be done for tracks inside a mounted thick section, with the chemicals reaching these inner tracks through mineral cleavages or other tracks. The maximum recordable length for tracks of VH-group nuclei are $\sim$15 μm; tracks of VVH nuclei are much longer.

The etched tracks are usually observed with optical microscopes, although electron microscopes are used to observe surfaces with very high track densities ($>10^8$/cm^2). Known magnifications are used to determine the area of the sample scanned for tracks. Several techniques are used to enhance the contrast of the tracks so that they are easier to observe. Often a thin layer of silver or other opaque substance is deposited in the tracks. Occasionally a replica of the holes is made by coating the surface with a material that will fill the holes and removing the coating with the holes now visible in the replica as fingers sticking out from the surface. In scanning the surface being examined, background holes, such as those made by the etching of defects, should not be counted. Backgrounds are usually not a problem for track densities above about 10^4/cm^2. Enough tracks must be counted over a known area to ensure a statistically valid number for the track density. In counting the tracks of a sample, some subjectivity is involved as different individuals can vary in what they consider real or background tracks. As with backgrounds, this is usually not a problem when the track densities are fairly high. Special techniques are often used to identify the densities of the long tracks due to VVH-group nuclei, such as grinding about 10 μm from the surface (to remove tracks of VH nuclei) after a heavy etch and silver plating. Tracks from VH and VVH nuclei in grains for the Patwar meteorite are shown in Fig. 7.

C. Nuclides

A large number of techniques is needed to measure the cosmogenic nuclides found in meteorites. The nuclides range from the stable isotopes of the noble gases to ra-

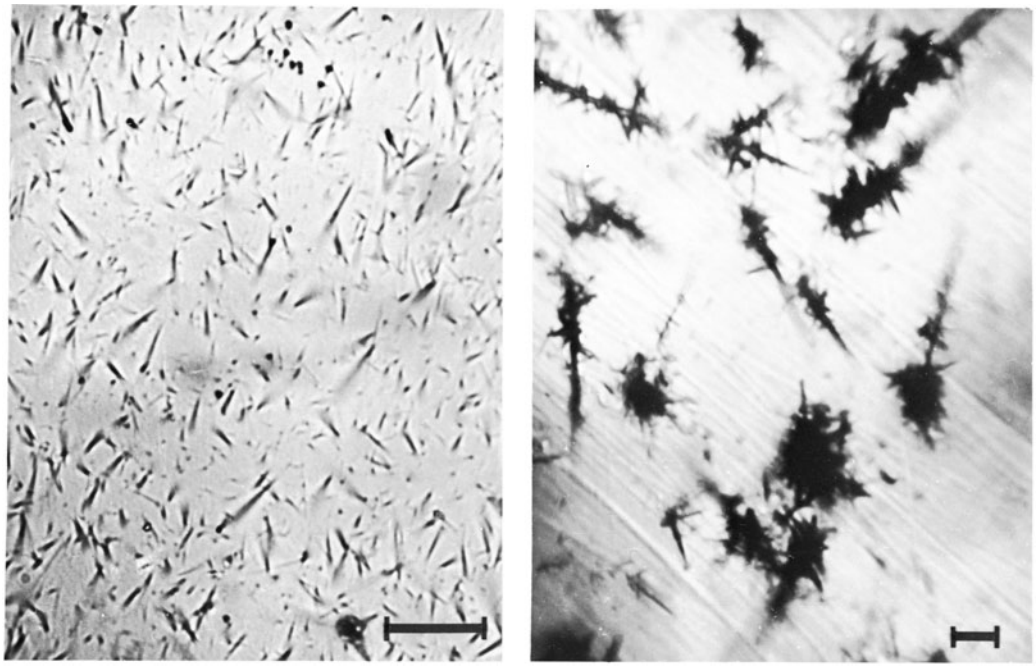


FIGURE 7 Photomicrographs of tracks due to VH and VVH nuclei in pyroxene grains from the Patwar meteorite (a mesosiderite). The low density of long VVH tracks was revealed by overetching the surface and removing the top layer. The lengths of the bars are 10 μm. (The photographs courtesy of Prof. D. Lal.)

dionuclides with a great variety of half-lives. Stable isotopes are usually measured by counting the atoms with a mass spectrometer. Radionuclides were usually observed by detecting the radiation that they emit when they decay, although now several more sensitive techniques are used for the longer-lived ones. Preparation of the sample can vary from almost nothing to detailed chemical separation procedures. Sample sizes have ranged from milligrams to kilograms.

The first measurements of cosmogenic nuclides in meteorites were of noble-gas isotopes and were done using mass spectrometers. In a typical noble-gas analysis, the gases are released from the sample into an evacuated system by heating, and then purified, ionized, and mass analyzed in the magnetic field of a mass spectrometer. Samples typically have masses of the order of a gram. The whole extraction system is heated under a vacuum to remove gases. Often the sample is first warmed to $\sim$100°C to remove gases adsorbed from the atmosphere. The gases can be collected in a single heating, or several fractions can be separated while the sample is heated to successively higher temperatures in a number of steps. Stepwise heating is useful if there are gases of diverse origins, such as trapped solar-wind gases, that are released at different temperatures. Gases not released at lower temperatures are released by completely melting a sample (at $\sim$1600°C). The gases are purified, usually by using charcoal at liquid nitrogen temperature or special getters of metals or alloys of titanium or zirconium. The noble gases often are separated by trapping and releasing from charcoal at different temperatures.

The purified noble gases are ionized, and the ions accelerated with an electric field and separated while passing through a magnetic field. The ions of the desired mass-to-charge ratio are detected, often with an electron multiplier. The mass peaks of interest are scanned a number of times. The backgrounds for the various mass peaks are frequently determined by duplicating the whole process without a sample in the heating crucible. Backgrounds usually result from molecules such as hydrocarbons, HD at mass 3, or doubly charged ions of argon or CO_2. The mass spectrometer is calibrated with sources of known isotopic composition (such as atmospheric gases) to get corrections for mass discrimination. Isotopic ratios for a given element can be determined very precisely. Absolute amounts can be measured using standards to calibrate the spectrometer or by isotopic dilution, adding a known amount of the same element with a well-known, different isotopic ratio prior to a mass analysis.

Solid isotopes of elements such as potassium and calcium have been analyzed mass spectrometrically. In these cases, the sample is usually chemically purified and placed on a special piece of wire that can be heated to very high temperatures. The atoms or molecules to be measured are

thermally released and ionized. The ionized species are mass analyzed in the same way as described above for gases.

Each radionuclide emits one or more characteristic radiations that can be detected by suitable counting techniques. Some radionuclides, such as ^{22}Na and ^{26}Al, emit γ radiations that are sufficiently penetrating and unique that they can be detected nondestructively in a sample. Often a number of cosmogenic or naturally occurring radionuclides emit similar radiations, so the element of interest must be separated from the sample prior to counting. Known masses of the element (called carriers) are usually added to trace the sample through the separation and to determine the final chemical yield. To be certain that there are no contaminations, a sample sometimes is repurified and recounted several times until the ratio of measured radioactivity to the sample mass remains constant. Energy analysis of the emitted radiation can often discriminate against similar radiations from other sources. Standards of known activities are needed to calibrate the counters and determine how efficiently the counter detects the emitted radiation. The efficiencies of the counters must be known to convert the measured counts per minute to disintegrations per minute (dpm). Activities of radionuclides are usually reported as dpm per kilogram (dpm/kg) of the initial sample. In a few cases, the units are dpm per kilogram of the major target element.

The counters are often actively or passively shielded from the cosmic rays and from naturally occurring radiations to reduce the backgrounds from such sources. In active shielding, a special counter covers the main counter, and, if a background radiation (such as a cosmic-ray particle) passes through it and gives a signal, then special electronics are used to cancel any signal from the main counter. Passive shielding uses large masses of nonradioactive material, such as pure lead, to stop background radiations from reaching the counter. In some cases, the radionuclide emits two or more radiations that can be detected separately with a pair of counters, and backgrounds can be greatly reduced by using a coincidence technique where both radiations must be detected within a very short period of time. Counting with low backgrounds due to shielding or coincidences is called "low-level."

Several volatile radionuclides, such as ^{3}H, ^{14}C, ^{37}Ar, and ^{39}Ar, are usually counted as gases inside specialized counters. Known amounts of carrier gases are added before the sample is heated. The gases released from the sample are purified and a known volume is added with counting gases and placed inside a counter designed to detect the radiation from the isotope of interest. Such internal counting is fairly efficient as most radiations are detected.

Cosmogenic radionuclides of nonvolatile elements, such as ^{10}Be and ^{53}Mn, must be counted as a solid, and the radiation detected external to the sample. Weak radiations, such as X-rays or low-energy β particles, are not very penetrating, and both the sample and the window through which the radiation enters the counter must be as thin as possible. Geiger or proportional counters filled with specialized gases are generally used for X-ray or β counting. Corrections must be made for self-absorption of weak radiations in the sample, so the thickness and composition of the counting sample must be known. Such counting techniques often require fairly large samples and the errors are frequently large because the counting rates are fairly low. Gamma rays can be detected using thallium-doped NaI-scintillation crystals and, more recently, with high-resolution solid-state detectors made of germanium. Gamma-ray spectra can be measured with chemically purified samples or large, unaltered pieces of a meteorite.

Coincidences between a γ ray and a β particle, X-ray, or other γ ray are often used for the low-level counting of radionuclides such as ^{22}Na and ^{54}Mn. Two counters are needed, and special electronics determine if both radiations occurred within a very short period of time. The γ-ray spectrum when there is a coincidence has a much lower background than one obtained without requiring a coincidence. For coincidences with β particles or X-rays, the sample usually is chemically separated and prepared in a thin counting geometry. Gamma–gamma coincidence counting can be done nondestructively on pieces of meteorites, usually with two large NaI(Tl) crystals positioned on opposite sides of the sample. Radionuclides that emit two or more γ rays (^{60}Co and ^{46}Sc), a positron (a positive electron that produces two 511-keV γ rays when it annihilates with an electron), or a positron and a γ ray (^{22}Na and ^{26}Al) can be measured very effectively with γ–γ coincidence counting.

Several radioactive or stable cosmogenic nuclides can be measured by neutron-activation techniques. The nuclide is exposed to a high flux of thermal neutrons in a reactor, which produces a radionuclide that can be readily counted. Neutron-activation analysis of stable cosmogenic nuclides has been done for isotopes such as ^{45}Sc that are produced in iron meteorites. Several long-lived radionuclides, ^{53}Mn and ^{129}I, can be measured by converting them to the short-lived radionuclides ^{54}Mn and ^{130}I. This activation method is routinely used for ^{53}Mn using samples as small as a few milligrams and can transform a sample that has only a few decays of ^{53}Mn (which captures an electron and emits a weak X-ray) per minute into one that has thousands of energetic ^{54}Mn γ rays emitted per minute. Before being irradiated with thermal neutrons, the sample most have a low activity of ^{54}Mn, and iron, which can make ^{54}Mn by the (n, p) reaction with ^{54}Fe, must be

chemically removed. The neutrons that irradiate the purified sample must not have many neutrons with energies above 10 MeV, or else ^{54}Mn will be produced from stable ^{55}Mn by the $(n, 2n)$ reaction. Careful monitoring of interfering reactions is done to correct for the ^{54}Mn not made from ^{53}Mn.

In the 1980s, a new technique for counting long-lived radionuclides was perfected, accelerator mass spectrometry (AMS). The atoms themselves are counted (as in a regular mass spectrometer) at an accelerator, not the radiation that they emit. A sample with a low level of radioactivity of a long-lived radionuclide such as ^{10}Be actually contains many atoms of ^{10}Be. (Only about one in 10^6 atoms of ^{10}Be will decay in a given year.) A regular mass spectrometer normally cannot detect the relatively few atoms of ^{10}Be because it is very hard to remove the atoms of the stable isobar ^{10}B prior to the mass spectrometry or to resolve them with the mass spectrometer. If the nuclei can be accelerated to high energies, however, nuclear particle-detection techniques can be used to distinguish ^{10}Be from ^{10}B. The energies required are at least a few MeV, which means that large accelerators, usually tandem Van de Graaffs, are needed. The cosmogenic radionuclides that have been frequently done so far with AMS are ^{10}Be, ^{14}C, ^{26}Al, and ^{36}Cl. A number of analyses of ^{41}Ca, ^{59}Ni, and ^{129}I have also been done with accelerators, and the list of isotopes measured this way keeps growing. The technique of accelerator mass spectrometry is very sensitive, and sample sizes of the order of milligrams are now used. Such small samples are big improvements over the gram-sized or larger samples that were required for counting the decays of radionuclides such as ^{10}Be.

Since about 1970, the new techniques of neutron activation and accelerator mass spectrometry, plus improved γ–γ coincidence counters and noble-gas spectrometers, have resulted in many more analyses of cosmogenic nuclides in meteorites. These measurements are now done with much smaller samples, which allows a number of different nuclides to be studied in fairly small pieces. For example, now the light noble gases (helium, neon, and argon) and the radionuclides ^{10}Be, ^{26}Al, and ^{53}Mn are frequently measured in a sample. It is now easier to do such comprehensive sets of analyses on several samples from a given meteorite, such as from cores drilled in several meteorites or from the surfaces or slabs of large pieces. Unfortunately, AMS is not yet sensitive enough for routinely measuring ^{41}Ca and ^{59}Ni, nuclides that are very hard to measure by conventional counting methods. The radionuclides ^{41}Ca and ^{59}Ni would be very useful in the studies of meteorites' cosmic-ray records because of their half-lives ($\sim 10^5$ years) and because they are often made by neutron-capture reactions, and, except for short-lived ^{60}Co, such products are seldom measured in meteorites. It is hoped that the list of cosmogenic nuclides that can be measured with very sensitive techniques will continue to grow. Some will probably be done by accelerator mass spectrometry. The use of lasers to selectively ionize specific elements or isotopes has some promise, but has yet to been applied to studies of cosmogenic nuclides.

V. HISTORIES OF COSMIC RAYS

A. Heavy Nuclei

The fossil tracks of heavy ($Z > 20$) nuclei in extraterrestrial minerals are the only way to study their record in the past and have provided much of the early characterization of heavy nuclei in the cosmic rays. These studies have provided long-term average values of the fluxes, energy spectra, and elemental ratios of cosmic-ray nuclei for atomic numbers (Z) from 20 to above 90 and in the energy range of $\sim$1–2000 MeV/nucleon. Many of these results depend on knowing additional data that are often model dependent, such as exposure ages, rates for the erosion of meteorites, or their fragmentation by collisions. Studies of large suites of different samples have eliminated many of these possible complications in the studies of tracks. The results from fossil tracks complement, and sometimes extend, what we know about heavy nuclei in the contemporary cosmic rays.

Much is known about the VH or iron group of nuclei in the ancient galactic cosmic rays. Tracks in various samples were formed over the past 10^9 years. The flux can be determined from the track density if the exposure age can be independently ascertained. The depth-versus-density profile can be used to get the energy spectrum. The flux and energy spectra of VH-group nuclei in the past are similar to those measured by satellites. For the VVH group ($Z \geq 30$), the main result is the VVH/VH abundance ratio in the energy region 100–1000 MeV/nucleon, which was 1.5 $(\pm 0.2) \times 10^{-3}$. This ratio and relative cosmic-ray abundance ratios for the charge groups 52–62, 63–75, 76–83, and 90–96 are in fairly good agreement with the corresponding elemental ratios for the Sun or the composition of primitive meteorites.

Most fossil track studies of the VH- and VVH-group nuclei in solar cosmic rays have been done with lunar samples because the very surface layers of meteorites with tracks of SCR nuclei are usually removed by ablation. The lunar-sample work has shown that the VVH/VH ratio for $E < 40$ MeV/nucleon is much higher, by about an order of magnitude, than that observed for GCR nuclei with $E > 100$ MeV/nucleon. Grains in the lunar regolith contain SCR tracks made a long time ago and since deeply buried. The high density and the steep gradient of tracks in a lunar grain are shown in Fig. 8. Lunar breccias also

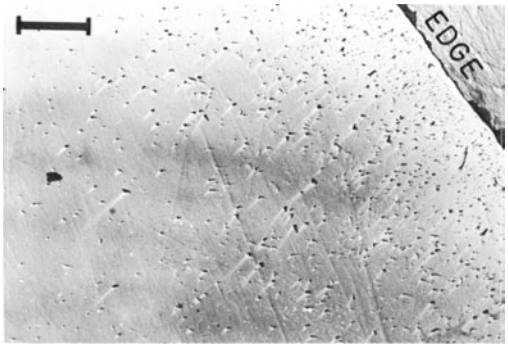


FIGURE 8 An electron micrograph of tracks due to VH nuclei in an Apollo 12 feldspar grain, obtained using the replication technique. Note the high track density and the appreciable track density gradient from the edge. The length of the bar is 1 μm. (Photograph graciously provided by Prof. D. Lal.)

contain grains with tracks made in the distant past. Their track records show that the energy spectra of ancient heavy nuclei in the SCR have not been very different from the recent ones. The determination of the flux of heavy nuclei in the SCR is difficult because it is hard to get an independent estimate of the exposure ages of these lunar soil grains and because erosion rapidly removed the SCR tracks.

The tracks of SCR heavy nuclei have been observed in mineral grains that are inside certain meteorites called "gas-rich" because they contain high concentrations of solar-wind-implanted noble gases. These tracks are present in very high densities and show pronounced density gradients over distances of a few micrometers, similar to the tracks from SCR nuclei seen inside lunar grains (see Fig. 8). These meteoritic grains were irradiated a long time ago while in a regolith similar to that on the moon's surface. Certain carbonaceous chondrites also have grains with tracks made during the early history of the solar system. These meteoritic grains are more isotropically irradiated than lunar grains, suggesting a more effective turning process, possibly because of the lower gravity field on the asteroid-sized parent body of these gas-rich meteorites. There is no way to estimate how long these grains were exposed to the solar wind or heavy SCR nuclei. The spectrum of VH nuclei in these gas-rich meteorites is not different from those observed elsewhere. The ancient VVH/VH ratios have varied, but not by more than a factor of 2 or 3, similar to variations in this ratio seen among contemporary events. These results for solar-wind gases and SCR tracks in modern and ancient grains indicate that the mechanisms that accelerate nuclei from the Sun have not changed much.

B. Spatial Variations

The intensities of the GCR particles have long been expected to vary with distance from the Sun in the inner solar system. Temporal variations over the 11-year solar cycle had been observed on Earth. Spatial variations in the inner solar system were anticipated because the galactic cosmic rays are modulated by the Sun and because the solar fields that interact with the cosmic rays become weaker as they move away from the Sun. Before satellites traveled far from the Earth's orbit at 1 AU, the magnitude of the variation of the GCR flux with distance from the Sun was not known, and first estimates were made using cosmogenic nuclides in meteorites. Photographs by several cameras of the fireball trajectories of three recovered meteorites and of many bright meteors showed that most were in orbits with low inclinations, perihelions near 1 AU, and aphelia in the asteroid belt. These orbits are fairly eccentric, and the meteorites spend most of their time near aphelion 2–4 AU from the Sun, where most of the longer-lived cosmogenic radionuclides are made. Only the very short-lived radionuclides, such as 35-day ^{37}Ar, are made mainly near the Earth.

A number of studies used the measured activity ratios of short-lived to long-lived radionuclides in meteorites to study GCR spatial variations. Some variations in ^{37}Ar/^{39}Ar activity ratios were interpreted as being caused by GCR gradients. With additional analyses of ^{37}Ar, ^{39}Ar, and other radionuclides in many meteorites, these early results on spatial gradients were examined more critically. The activities of long-lived radionuclides, such as ^{39}Ar, showed very little spread in meteorites that probably had a range of aphelia, suggesting that the spatial gradients of the GCR were not large. These meteorites fell during different phases of the solar cycle, so the correlation of the ^{37}Ar activity with other measures of the cosmic-ray flux, such as count rates of cosmogenic neutrons in terrestrial neutron monitors, were examined. The correlation between the ^{37}Ar activity and the neutron-monitor count rates was weak, although the trend was for higher activities when the Sun was less active. Other mechanisms to explain the variations in the ^{37}Ar/^{39}Ar activity ratios have been proposed, such as the perihelions of these meteorites or whether the meteorite was moving toward or away from the Sun when it hit the Earth. The generally accepted conclusion now is that cosmogenic radionuclides do not show any evidence of large gradients of the GCR with distance from the Sun. Over the past decade, experiments on the Pioneer, Voyager, and Ulysses satellites have measured the GCR intensities with distance from the Sun, and the GCR spatial gradients are low, only 2–3%/AU.

Prior to 1990, no satellites have traveled very far from the ecliptic. Only since 1990 has the Ulysses (the International Solar Polar Mission) been exploring the region away from the celiptic. Also, very few spacecraft have explored the solar system inside about 0.5 AU. Unusual

activities or activity ratios of cosmogenic radionuclides are sometimes interpreted as reflecting variations of the cosmic-ray intensities in the solar system. Orbits determined from visual observations of meteor trails indicate that some meteorites may have had fairly high inclinations with respect to the ecliptic, such as about 20° and 28° for Allende and Dhajala, respectvely. In Dhajala, the short-lived radionuclides have activities similar to or slightly higher than those in other meteorites that fell at about the same time, while the long-lived ones tend to be less radioactive than normal. These ratios of short-lived to long-lived radionuclides in Dhajala are higher by about 30–50% than usually observed in meteorites. These high ratios could be caused by a variation in the cosmic-ray intensity with heliographic latitude, although they could reflect shielding changes in the past. The Malakal and Innisfree chondrites have very high activities of ^{26}Al. Malakal has a very low thermoluminescence level that suggests it was heated in an orbit with a small perihelion (∼0.5–0.6 AU). Although the unusual ^{26}Al activity in Malakal could be a consequence of its orbit, data for other cosmogenic nuclides suggest that Malakal (and Innisfree) possibly had complex exposure histories. Thus there are no clear cases of a meteorite's cosmic-ray record showing unusual spatial variations in the intensities of the cosmic rays.

C. Temporal Variations

Unlike changes in the fluxes or spectra of heavy nuclei in the past or spatial variations of cosmic rays, significant temporal variations of cosmic rays have been seen from the concentrations of cosmogenic nuclides in meteorites. These variations in the intensities of the GCR protons and α particles cover periods that range from an 11-year solar cycle to $\sim 10^9$ years. Other time intervals studied with cosmogenic radionuclides include the Maunder Minimum, which occurred about 300 years ago, and 10^5 to 5×10^6 years ago. Temporal variations have been based primarily on radionuclides with a variety of half-lives that extend from 16-day ^{48}V to 1.28×10^9-year ^{40}K. As can be seen from the list of cosmogenic radionuclides in Table II, there are many gaps in the half-lives of such radionuclides studied in meteorites. For example, except for a very few ^{129}I measurements, no radionuclides with half-lives between 3.7×10^6-year ^{53}Mn and ^{40}K have been measured. Several cosmogenic radionuclides in terrestrial samples, such as ^{14}C and ^{10}Be, represent very short time periods because these nuclides are rapidly removed from the atmosphere where they were made and stored in places, such as plants or ice layers, with very little subsequent alteration. In meteorites, radionuclides usually have been produced over their last few half-lives, so the time period examined is determined by the radionuclide's half-life.

Although the variations in the activities of ^{37}Ar meteorites were not strongly correlated with solar activity, such variations have been seen more clearly for other short-lived species. Several research groups have nondestructively measured y-ray-emitting radionuclides, especially short-lived ^{46}Sc, ^{54}Mn, and ^{22}Na, in a number of recent falls. Although the variations in activity with phase of the solar cycle is visible in the raw radioactivities, it is more visible when corrections are made for the shielding of the cosmic rays by the meteoroid's geometry and the sample's depth. Ratios of activities of radionuclides made by similar reactions, such as ^{22}Na/^{26}Al or ^{54}Mn/^{22}Na, eliminate most effects due to shielding and clearly show the variation in radioactivity with solar cycle. The use of shielding corrections based on the measured ^{22}Ne/^{21}Ne ratios also showed the recent temporal variations in short-lived radioactivities more clearly. The shielding-corrected variations in activity are greatest for the shorter-lived radionuclides ^{46}Sc and ^{54}Mn, and the inferred production-rate variations are a factor of 2.5–3 over the 11-year solar cycle. This magnitude in the production-rate variation is consistent with the measured fluxes and spectral variations of GCR particles. The GCR-proton fluxes over a solar cycle vary considerably for $E < 1$ GeV and by a factor of 2 for all energies above 1 GeV.

Although the ^{37}Ar/^{39}Ar ratios measured in the metal phases from many meteorites showed much variation, they provided some evidence for the Maunder Minimum in solar activity. The 269-year half-life of ^{39}Ar is ideal for studies of this period from 1645 to 1715 when there were extremely few sunspots or aurorae and enhanced production of ^{14}C in the Earth's atmosphere. A similar minimum in solar activity, the Sporer Minimum, occurred about 200 years before the Maunder Minimum. As noted above under spatial variations, the activities of ^{37}Ar were usually lower than those measured for ^{39}Ar, whereas the production rates of these radionuclides from iron are essentially equal. M. Forman and O. Schaeffer noted that the mean activity of ^{37}Ar is below that for ^{39}Ar, even if the observed spatial gradient in the GCR (about 3%/AU) is considered, and interpreted this excess ^{39}Ar (∼18%) as having been made by enhanced fluxes of GCR particles during the Maunder and Sporer minima. Measurements of ^{39}Ar in the few meteorites that fell just after the Maunder Minimum could show the effects of the Maunder Minimum more clearly than using recent falls.

The radionuclide pair ^{39}Ar and 3×10^5-year ^{36}Cl are made from iron in nearly equal yields. The measured ratios for the activities of these two radionuclides in iron meteorites are also about unity, so the average fluxes of

GCR over the past 300 years are not very different (<10% variation) from those over the past 3×10^5 years. As the solar-cycle-averaged ^{37}Ar activities are slightly lower than those of ^{39}Ar and ^{36}Cl (although only by about a standard deviation), there is a hint that the GCR fluxes during the past few decades could be slightly lower than in the past 10^3–10^5 years, possibly because there have not been any recent periods of unusually low levels of solar activity, like the Maunder Minimum.

Several studies have looked for possible GCR variations over the past 5×10^6 years. The measured activities of a number of radionuclides in iron meteorites have been compared with calculated production rates to search for such cosmic-ray variability. The ratios of the observed to calculated activities vary by factors of 2 (due probably to both calculational and measurement uncertainties), but show no systematic trends with half-lives. Several research groups have inferred production rates for noble-gas isotopes, such as ^{21}Ne, from meteorites with known exposure ages. As radionuclide activities were used to get these exposure ages (from either the undersaturation of activity or from a radioactive/stable pair), variations in the inferred production rate of ^{21}Ne could be a consequence of GCR intensity variations in the past. Undersaturation of 7×10^5-year ^{26}Al, 1.5×10^6-year ^{10}Be, and 3.7×10^6-year ^{53}Mn and the pairs ^{22}Na/^{22}Ne and ^{81}Kr/^{83}Kr were used by various groups in inferring ^{21}Ne production rates. Results from all radionuclides except ^{26}Al gave similar results; however, the ^{26}Al-inferred ^{21}Ne production rates were considerably higher, by about 50%. Several simple possible explanations for this anomaly, such as an incorrect half-life for ^{26}Al, have been eliminated, although slight (∼10%) adjustments in several half-lives would reduce the magnitude of the discrepancy. The shielding corrections used in these calculations were reexamined, but the discrepancy still persisted. Temporal variations in the cosmic-ray intensities over the past few million years could account for some of this variation in inferred ^{21}Ne production rates, but not all of it, because the half-life of ^{26}Al is intermediate and not that different from those of 2×10^5-year ^{81}Kr and ^{10}Be. A probable explanation for the high ^{21}Ne production rates determined from meteorites that are undersaturated in ^{26}Al is some ^{21}Ne from previous irradiations. Other possible explanations are difficulties in making shielding corrections or unusual cosmic-ray records for the few meteorites that have very low ^{26}Al activities. These meteoritic results for over the past 5×10^6 years are consistent with measurements for cosmogenic nuclides in lunar and terrestrial samples, which imply that cosmic-ray-intensity variations of >30% for periods of $\sim 10^5$ years have been unlikely.

Most stony meteorites have cosmic-ray exposure ages considerably below 5×10^7 years, so studies of cosmic-ray variations for longer time periods have used iron meteorites, which usually have much longer exposure ages. The radioactive/stable pairs that generally have been used to determine exposure ages of iron meteorites are ^{39}Ar/^{38}Ar, ^{36}Cl/^{36}Ar, ^{26}Al/^{21}Ne, ^{10}Be/^{36}Ar, and ^{40}K/^{41}K. The first four radionuclides have half-lives that range from 269 years to 1.5×10^6 years, and the exposure ages of various iron meteorites determined with those pairs generally agree well within the experimental uncertainties, and, like the results discussed above, imply no major variations in the fluxes of cosmic rays over the past few million years. However, the exposure ages determined with 1.28×10^9-year ^{40}K tend to be 45% greater than the ages inferred with the shorter-lived radionuclides.

These higher exposure ages based on ^{40}K were first discovered around 1960 by H. Voshage and H. Hintenberger. Much work has been done since then that confirms the earlier results. (However, Voshage cautions that the key parameter in the ^{40}K work, the ^{40}K/^{41}K production ratio, could possibly be incorrect as it was not directly determined from experimental bombardments as were the other production ratios, although he believes that it is known much better than a factor of 1.5.) The question has been whether the higher exposure ages based on ^{40}K represents something in the history of the cosmic rays or that of the iron meteorites. For example, complex exposure histories for iron meteorites are fairly common. This explanation is unlikely because at least 8 iron meteorites have exposure-age ratios of 1.45 ± 0.10, while other ratios (for irons with complex histories) range from 3 to 8. Another explanation is that these iron meteorites, which have exposure ages of 3×10^8 to 1×10^9 years, have been slowly and steadily eroded by solar-wind particles and micrometeoroids in space. An erosion rate of $\sim 2 \times 10^{-8}$ cm per year could produce a sufficient shielding change to account for the observed differences in the exposure ages. O. Schaeffer, H. Fechtig, and co-workers used laboratory simulations to estimate that the erosion rates of iron meteorites in space to be 2.2×10^{-9} cm per year (30 times slower than for stony meteorites), too low to account for these exposure-age ratios. B. Lavielle, K. Marti, K. Nishiizumi, and co-workers have measured noble-gas isotopes and radionuclides in iron meteorites with reported K ratios and confirmed that exposure ages determined with ^{36}Ar–^{36}Cl (and other radionuclides) are about 28% lower than those determined with K isotopes. The generally accepted conclusion is that the high exposure ages determined with ^{40}K were caused by a flux of cosmic rays that has been ∼25–50% higher over the past $\sim 10^6$ years than over the past 10^8–10^9 years.

As discussed above, the cosmic-ray records of meteorites clearly show variations in the intensities of cosmic rays in the inner solar system over an 11-year solar cycle,

hint at an increased flux during the Maunder Minimum during 1645–1715, set limits of less than ~30% variations for $\sim 10^5$-year periods over the past 5×10^6 years, and strongly imply lower averaged intensities $\sim 10^8$–10^9 years ago. These cosmic-ray-intensity variations are consistent with what we know about the sources of the cosmic rays, their transport to the solar system, and their modulation by the Sun. The effects of the 11-year solar cycle and the Maunder Minimum are clearly seen in other records. The transit times of cosmic-ray particles from their sources to the solar system, $\sim 10^7$ years, the 6.6×10^7-year period of the Sun and solar system in its vertical motion relative to the galactic plane, and the distributions of supernovae in our galaxy (the probable energy source for most cosmic rays) are consistent with the meteorite results for longer time periods.

VI. HISTORIES OF METEORITES

A. Preatmospheric Geometries and Meteorite Orbits

The cosmic-ray record of a meteorite can be used to reconstruct its geometry in space while it was being irradiated by the cosmic rays. This reconstruction is needed because the meteorite recovered on Earth was ablated and possibly fragmented during its passage through the atmosphere. We can determine the preatmospheric geometry of a meteorite and the location of a given sample within the meteoroid because, as discussed above, the production rates of cosmogenic products vary with geometry and depth. Usually the products that have production profiles that vary the most with shielding are best for such studies. The density of heavy-nuclei tracks and concentrations of neutron-capture-produced nuclides are well suited for such studies. Ratios of the concentrations of certain cosmogenic products, such as the ^{3}He/^{21}Ne ratio shown in Fig. 6 or the ^{22}Ne/^{21}Ne ratio, are often used in estimating the preatmospheric radii of meteoritic samples. Evidence for SCR-produced products in meteoritic samples also is a sensitive indicator that a sample had very little preatmospheric shielding. The results of such shielding studies are often needed in the interpretations of other cosmogenic products.

The production rates of heavy-nuclei tracks vary rapidly with depth in a meteoroid (see Fig. 1), so, if the exposure age of a meteorite can be independently determined, the track density of a sample is a very good indicator of how far that mineral grain was from the nearest preatmospheric surface. (Care must be taken in track studies of small meteorites to account for the tracks made by nuclei arriving from all directions.) The usual methods involving radioactive or stable cosmogenic nuclides give exposure ages that are more than adequate for determining the rates at which the tracks were produced in a meteoritic sample. The experimental rate is compared with curves such as that in Fig. 1 to get the sample's preatmospheric depth. As the production curve is fairly steep, small errors in the track density or the exposure age do not much affect the determination of the depth. An uncertainty of $\pm 100\%$ in the radius only affects the inferred depth by about $\pm 10\%$. These preatmospheric depths are very useful in interpreting the record of cosmogenic nuclides. To get the meteorite's complete preatmospheric geometry requires measurements for samples that were at different locations. For several meteorites, such as St. Severin and Jilin, cores were drilled through the main mass and complete track profiles measured. For meteorites that fell as a shower of many fragments, it is harder to reconstruct the original geometry.

Several studies of preatmospheric depths used both tracks and cosmogenic nuclides. The ^{22}Ne/^{21}Ne ratio is a rough measure of a sample's shielding conditions. The concentrations of ^{21}Ne (and sometimes ^{3}He or ^{38}Ar) give the exposure age. The cosmogenic-nuclide data help to estimate the preatmospheric radius and narrow the ranges of depths consistent with the measured track production rate. Comparisons of the measured track production rates with a cosmogenic nuclide measurement, such as the ^{22}Ne/^{21}Ne ratio or the activity of a radionuclide such as ^{53}Mn, can be used to identify meteorites with complex exposure histories and to limit the sets of radius and depth combinations that are possible for that sample. For example, in Fig. 9, the production rates for tracks and ^{53}Mn measured in the St. Severin core samples are compared with calculated profiles. The results in Fig. 9 show that this core had a preatmospheric radius of between 20 and 25 cm, in good agreement with the length of the core plus the ablation losses determined from the track data. Similar plots of track production rates versus ^{22}Ne/^{21}Ne ratios have been used in such studies. The results for a single sample can give a fairly good estimate of the preatmospheric radius in many cases.

For stony meteorites, a major study by N. Bhandari and co-workers showed that the fraction of the preatmospheric mass lost by ablation ranged from about 27% to 99.9% with a median value of about 85%. Preatmospheric masses typically ranged from 10 to 1000 kg with most meteorites having been near the lower end of this range. The probability of a preatmospheric mass in this range varied inversely with the mass. The amount of ablation varies over the surface of a given meteorite, and there are often regions of relatively low (<7 cm) ablation on heavily ablated meteorites. Track results for ablation have shown that several meteorites, such as Bansur and Rosebud, were very lightly ablated in certain places and heavily ablated

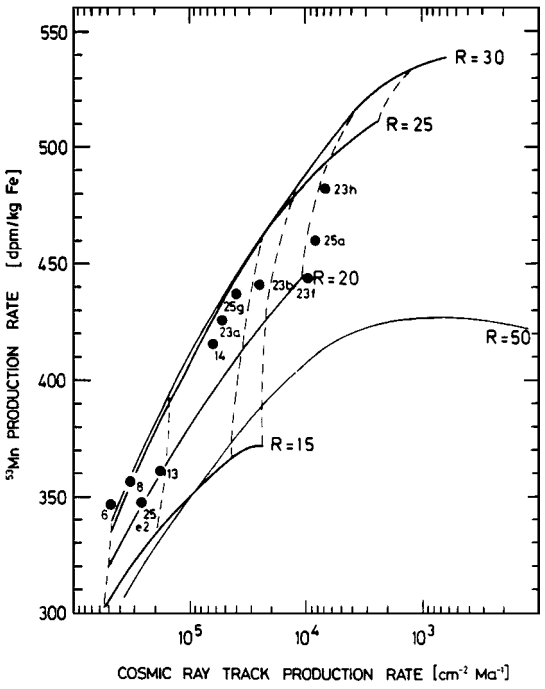


FIGURE 9 Production rates for ^{53}Mn versus track production rates in stony meteorites of various radii (in cm) are shown with measured data from core samples (dots with numbers) of the St. Severin LL-chondrite using an exposure age of 13×10^6 years. Figure and data for ^{53}Mn are from P. Englert; track densities from D. Lal and co-workers or Y. Cantelaube and co-workers; theoretical ^{53}Mn production rates from T. Kohman and M. Bender; and calculated track production rates from S. Bhattacharya and co-workers.

elsewhere. Studies that related ablation with the meteoroid's velocity before it reached the Earth have shown that the mass fraction removed is a steep function of velocity, but that many parameters besides entry speed, such as mass, shape, entry orientation, and meteorite type, also affect the amount of ablation. Meteoroids approaching the Earth with high velocities (>25 km/s) must have very large masses if they are not to be completely burned during entry; thus, meteorites with large aphelia (>3 AU) are not properly represented in our collections.

The concentration of a neutron-capture-produced nuclide, such as ^{60}Co, can frequently by used to determine or set limits on preatmospheric sizes and sample depths. An extensive study of the large Jilin chondrite by G. Heusser and co-workers used activities of ^{60}Co from documented locations of the main mass to establish that Jilin was spherical with a radius of about 85 cm just before it fell in 1976. Usually a ^{60}Co activity about 1 dpm/kg or less implies that the preatmospheric size of the meteorite was too small to thermalize neutrons, that is, it had a radius of less than about 25 cm. The production of ^{36}Cl, ^{80}Kr, ^{82}Kr, ^{128}Xe, and ^{131}Xe in meteorites by neutron-capture reactions with chlorine, bromine, iodine, and barium has also been used to set lower or upper limits on the preatmospheric radii of meteorites.

In several meteorites, unusual profiles or concentrations of cosmogenic products have been observed that appear to have been produced by solar-cosmic-ray particles. For example, the Antarctic meteorite ALHA77003 has a ^{53}Mn profile that decreases with distance from the surface and implies that ALHA77003 was probably a very small object in space. As the amount of SCR-production usually is fairly small and roughly compensates for the decreased GCR production near the surface, it is often hard to detect that SCR production has occurred. The concentrations of ^{21}Ne measured in the outermost 1 cm of a core from St. Severin was higher than in deeper samples, a clear sign of SCR production. A high activity of SCR-produced ^{56}Co was measured in the Salem meteorite, which means that Salem was very small in space. Meteorites that were very small objects in space seem to be very rare.

The ablation of iron meteorites has been estimated by mapping the concentrations of various cosmogenic noble-gas isotopes measured from slabs. Contours of equal concentrations for several cosmogenic nuclides that were measured by Signer and Nier in the Grant iron meteorite were slightly ellipsoidal, and comparisons with calculations showed that about 75% of the preatmospheric mass was lost by ablation. Noble-gas measurements for a single sample and an iron meteorite's exposure age can be used with several models to infer the sample's preatmospheric depth and the meteoroid's radius.

Although measurements for a single sample of a meteorite are adequate to get an estimate of its preatmospheric radius, a suite of samples are needed to get a meteorite's preatmospheric shape. The few meteorites that have had their shapes in space determined tend to be nonspherical, although Jilin's geometry just prior to hitting the Earth appears to have been very close to a sphere. Cosmogenic products have often been used to help show that two or more meteorites of the same type were part of the same parent body just prior to entry into the Earth's atmosphere. If the shielding-corrected exposure ages are similar and the cosmogenic products fit along the profile for a given radius (such as those in Fig. 9), then there is strong evidence that they came from a common parent object. A number of stony meteorites from Antarctica or iron meteorites have been paired with the aid of cosmogenic products. It appears that many of the thousands of meteorites found in Antarctica came from a much smaller set of meteoroids hitting the Earth.

The orbits of meteorites are hard to infer using most cosmogenic nuclides as the variation in GCR-particle fluxes in the inner solar system is low. However, there are two records that have been used to infer something about a

meteorite's orbit. The presence of nuclides made by SCR particles is an indication that a meteorite did not spend most of its time at a distance of several AU from the Sun. Thus, the observation of SCR-produced radionuclides, including ^{26}Al, in Salem suggests that Salem did not have a large aphelion recently and that it spent most of its time fairly near the Sun. The lack of a strong SCR component for ^{21}Ne in Salem could indicate an orbital change several million years ago.

The amount of thermoluminescence in some meteorites has been used to infer their orbits. The preservation of TL in a meteorite is a sensitive function of the highest temperature to which the meteorite was exposed in its orbit. Low values of TL in a few meteorites indicate that they could have had orbits that approached the Sun as closely as 0.6 AU, consistent with orbits determined from fireball and other data. Most ordinary chondrites had perihelia between 0.8 and 1.0 AU. One group of meteorites from Antarctica had very high values of TL that suggested that they had a decrease in their perihelia from 1.1 AU or more to 1 AU within the last $\sim 10^5$ years. Several recent falls, including Jilin with a most-recent exposure age of about 4×10^5 years, also have high TL values.

B. Terrestrial Ages

The lengths of time that meteorite finds have been on the Earth's surface can often be determined from their cosmic-ray records. The amount of weathering is a qualitative indication of terrestrial age but varies quite widely with location of falls and meteorite type. As the activities of radionuclides and the amount of thermoluminescence (TL) decays with time, these cosmogenic products can be used to determine or set limits to a meteorite's terrestrial age. Terrestrial ages are sometimes needed to correct the activities of various radionuclides for decay while the meteorite was on the Earth's surface. They also can be used to identify whether several meteorites are pairs, fragmented from a common object during passage through the Earth's atmosphere. Falls and finds also have been used to see if the distribution of meteorite types hitting the Earth has varied in the past. However, as terrestrial ages are usually much shorter than exposure ages, it seems unlikely that the present meteorites in our collections will be able to show much difference in the frequencies of meteorite types striking the Earth in the past. Meteorites from the moon, as found in Antarctica, are exceptions because of the short time that they are in space, and their terrestrial ages probably will date when they were ejected from the lunar surface.

The decay of TL depends on a meteorite's temperature record since it fell, but meteorites usually have been stored in museums for most of the time. The decay of TL also varies with the temperature at which the glow was measured, as the lower temperature TL decays faster than that at higher temperatures. The decay of TL is not a pure exponential, so calibration with samples of known terrestrial ages is needed. For the Antarctic meteorites, the mean life for the decay of TL has been measured to be $\sim 10^5$ years, much longer than for meteorites that fell in other locations.

Many meteorites from Antarctica have had their terrestrial ages determined from the decay of ^{14}C, ^{41}Ca, ^{81}Kr, ^{36}Cl, ^{26}Al, and ^{10}Be. The distribution of the number of cases versus terrestrial age is concentrated below about 3×10^5 years and decreases rapidly for higher ages. The greatest terrestrial age for an Antarctic meteorite is about 2×10^6 years. Such results are important in determining how the Antarctic meteorites were concentrated at the few ice fields on which they are found. The terrestrial ages of the Antarctic meteorites found by the Japanese near the Yamato Mountains generally are shorter than those recovered next to the Allan Hills by American expeditions. These meteorites found in Antarctica are believed to have been carried inside the ice to areas where the ice flow is stopped by an obstacle, such as a mountain range. Thus, terrestrial ages of these meteorites can be used to examine ice movement in Antarctica.

The cosmogenic radionuclide ^{14}C has been used to determine terrestrial ages of stony meteorites from areas other than Antarctica. Not many such stones had terrestrial ages greater than 10^4 years, although the Potter L6-chondrite, which was found in a dry climate, has a terrestrial age that is greater than 2×10^4 years. Iron meteorites can have much longer terrestrial ages than stones. The longest terrestrial age is about 3×10^6 years for the IIIA iron meteorite Tamarugal, found in the desert of northern Chile. The terrestrial ages for many of the 12 hexahedrites (type IIA) found in northern Chile helped to show that these iron meteorites represent at least six different falls. Many of the iron meteorites found with large terrestrial ages were found in mountain areas, probably because alluvial processes make it hard to find old meteorites in the lowlands.

C. Complex Histories

There are a number of indications that certain meteorites have received their cosmogenic products both before and after a major change in their geometry. Such a history for meteorites is called a complex one, as opposed to a simple or one-stage history where all the cosmogenic products were produced without any shielding changes. For a number of meteorites, the exposure ages determined by different methods gave a wide range of ages. For example, some stones have low exposure ages based on ^{26}Al activities and higher ages from their ^{21}Ne concentrations. In interpreting

the exposure-age records of meteorites, samples of meteorites with complex histories need to be identified because the age determined assuming a simple history would be incorrect. The fraction of meteorites with complex histories relative to those with simple histories can be used to infer how meteoroids evolved prior to hitting the Earth. The rapid erosion of meteoroids by dust and radiation in space, which could create histories that appear complex, has been shown by various arguments and experiments to be unimportant, and thus collisions among larger objects are the sources of geometry changes in space.

A number of methods have been used to detect whether a meteorite had a complex history. Neutron-capture-produced nuclides or a low $^{22}Ne/^{21}Ne$ ratio (below about 1.07 for chondrites) are indications of production in a large object, and their presence in a meteorite that does not appear to have been very large is a sign to check for a possible complex history. Often the concentration of a stable cosmogenic product, such as ^{21}Ne or tracks, is higher than that predicted from the undersaturation of a radionuclide. The activities of the radionuclides ^{26}Al and ^{53}Mn have been used with ^{21}Ne concentrations to search for complex histories. The experimentally determined trends for a radionuclide's production rate versus the track production rate (as in Fig. 9) or versus the $^{22}Ne/^{21}Ne$ ratio can identify complex histories if the data plot outside the allowable range. In some comparisons using radionuclide activities and noble-gas concentrations, the exposure age determined from the noble gas is lower than that determined from the radionuclide's activity, implying loss of the noble gas. It is possible that the excess amounts of a noble gas from an earlier exposure could be later lost, thus creating a record that seems to have been a simple one. The use of the $^{22}Ne/^{21}Ne$ ratio versus the inferred track production rate can be used to search for complex histories, as there is a range of values prohibited for simple histories.

A problem with these correlation methods is that a sample with a complex history could be shifted from one part to another part of the allowable range. This is especially true if only one sample from a meteorite has been analyzed, and there probably are some meteorites that have been incorrectly identified as having simple histories from such correlation trends. Measurements of a suite of cosmogenic products from a number of different locations can help to distinguish complex versus simple histories. In the Jilin chondrite, the ratios of ^{22}Na to ^{26}Al varied widely from sample to sample, showing that it had a complex history with a large geometry change fairly recently. Extensive analyses of Jilin samples have shown that it was first exposed to cosmic rays for $\sim 9 \times 10^6$ years as part of a very large object and then was changed to an 85-cm-radius sphere about 4×10^5 years ago. Such detailed unfolding of a meteorite's complex history, however, is rare.

In iron meteorites, cases of complex histories have been discovered when exposure ages determined with several radioactive/stable pairs have disagreed. For example, exposure ages determined with ^{39}Ar, ^{36}Cl, or ^{10}Be often disagree with those determined with ^{40}K after the factor of 1.45 for the variations in the cosmic-ray fluxes is considered. The exposure ages determined with the shorter-lived radionuclides are much less than the $^{40}K/^{41}K$ ages in $\sim$30% of the cases, indicating enhanced production of cosmogenic nuclides fairly recently in the meteorite's history because of a shielding change. The isotopic ratios in these cases usually indicate that the sample was heavily shielded for most of its history. For very large iron meteorites (such as Odessa, Canyon Diablo, or Sikhote-Alin), different fragments can have different exposure ages, probably as a consequence of fragmentation of only one part of the meteoroid.

A special case of a complex history is when a part of a meteorite received a cosmic-ray irradiation prior to its incorporation into the body in which it was found. Gas-rich grains in meteorites received an irradiation by solar-wind ions and energetic SCR heavy nuclei early in their history. Cosmogenic nuclei also have been observed in several track-rich grains from gas-rich meteorites. Xenoliths, or foreign inclusions, separated from the H4-chondrite Weston and the LL6-chondrite St. Mesmin had concentrations and ratios of cosmogenic noble gases that differed significantly from the trends for other inclusions in the same meteorite. These xenoliths appear to have been exposed to cosmic rays prior to the compaction of these brecciated meteorites.

There are not very good statistics on the numbers of meteorites with well-documented complex or simple histories, partially because it is hard to be certain which type of history a meteorite had. It appears that iron meteorites are more likely to have complex histories than stones, which would be expected because of their longer exposure ages. The types of histories that meteorites have had probably cover a continuum that ranges from simple to ones with significant production in two very different geometries. As mentioned above, one explanation for the high ^{21}Ne production rates inferred from undersaturation of ^{26}Al could be small amounts of ^{21}Ne made prior to the present geometry. A small excess of cosmogenic ^{21}Ne would normally be hard to detect, and many meteorites might have small amounts of cosmogenic products made prior to their most recent exposure geometry.

Complex histories are useful in examining another stage further back into the evolution of a meteorite than is studied with exposure ages of meteorites with simple histories. G. Wetherill has predicted that complex histories should be more common (on the order of half the cases) than observed if the collisional model for the origins of meteorites is correct and wonders whether this discrepancy is

a problem with the model or with interpreting the cosmic-ray records of meteorites. R. Greenberg and C. Chapman have argued that many types of meteorites could be delivered to the Earth directly by cratering on large main-belt asteroids. Many meteorites probably were put into Earth-crossing orbits by orbital resonances and other non-collisional processes. Such meteorites are less likely to have had complex histories. Complex histories for the very mechanically weak carbonaceous chondrites could be the result of irradiation in the surface layers of a comet prior to the release of the meteorite after the ice around it has been evaporated or sublimed.

Because the real histories of meteorites have not been well determined in most cases, it is hard to use complex histories in studies of the evolution of meteorites. Many more measurements, especially of products with a wide range of production profiles (including tracks and neutron-capture nuclides) in a number of different samples from each meteorite are needed. Detailed studies have shown that, like Jilin, the meteorites Bur Gheluai, Chico, and Torino had complex exposure histories.

D. Exposure Ages

In meteorites, the concentrations of certain cosmogenic products can be used to determine the length of time for which the meteoroid was exposed to cosmic rays, the meteorite's exposure age. An exposure age is the interval of time from when the meteorite was removed from a very heavily shielded location many meters deep in a parent body to when it hit the Earth. The cosmic-ray exposure ages for most meteorites are orders of magnitude shorter than the radiometric ages, usually 4.55×10^9 years, for their formation. Exposure ages for meteorites are important in studies of the sources of meteoroids and the mechanisms that caused them to hit the Earth. Except for terrestrial ages, cosmic-ray exposure ages are the youngest of the possible ages for a meteorite. Other ages used in studies of the evolution of meteorites since their initial formation include collisional shock ages, typically 3×10^7 to 7×10^8 years as inferred from losses of radiogenic gases or resetting of radiometric clocks, and ages for the brecciation of certain meteorites from the regoliths of parent bodies about 1.4×10^9 to 4.4×10^9 years ago.

The results from studies of some meteorites can be used in understanding how their parent bodies evolved early in the solar system, and it would be nice to know where in the solar system the meteorites formed. The orbits of the objects from which meteoroids were produced provide an understanding about the fluxes of interplanetary meteoroids and asteroids. Such objects in Earth-crossing orbits can produce craters, some of which are very large and which probably caused extinctions on Earth in the past, and there are considerable uncertainties in their fluxes.

An exposure age is usually determined from the concentration of a stable cosmogenic product, such as ^{21}Ne, and a production rate. Often corrections to the production rate because of the sample's shielding are used, such as those based on the ^{22}Ne/^{21}Ne ratio. Many radioactive/stable pairs, such as ^{39}Ar/^{38}Ar, do not require shielding corrections when used to determine exposure ages. Sometimes the activity of a radionuclide, such as ^{26}Al, below its equilibrium value can be used to calculate an exposure age. In calculating exposure ages, it is assumed that the production rate has remained constant, not having varied in the past. As discussed above, this assumption is generally valid. Even if the cosmic-ray flux has varied, the trends and groupings in exposure ages would still be preserved and useful for meteorite studies. The geometry of the meteoroid in space is assumed not to have changed during its exposure to the cosmic rays. Complex exposure histories, involving major changes in the meteoroid's geometry with respect to the cosmic rays due to collisions, are known to have happened to some meteorites. Thus, it is possible that some exposure ages are incorrect because the meteorite's exposure history was assumed to be simple, but actually was complex.

For stony meteorites, concentrations of ^{21}Ne and shielding corrections based on ^{22}Ne/^{21}Ne ratios have been used to determine exposure ages for hundreds of meteorites. As the production rates for ^{21}Ne were revised downward several decades ago, the exposure ages given here are occasionally higher than those originally reported. The stony meteorites have exposure ages that range from ~50,000 years for the L5-chondrite Farmington to about 8×10^7 years for the aubrite Norton County. The distributions of exposure ages for the three largest types of meteorites, the H-, L-, and LL-chondrites, are shown in Fig. 10. The H-chondrites have a major peak near 7×10^6 years that contains almost half of this type of meteorite. Its width (about ±15%) probably represents variations among the accuracies of the measurements by different laboratories, other measurement uncertainties, poor shielding corrections, diffusion losses of ^{21}Ne from some samples, and possibly small amounts of ^{21}Ne made prior to the present exposure geometry. This peak contains all petrologic types of H-chondrites, although type H5 is somewhat more frequent and H6 less common in this peak than on the average. The exposure ages of the LL-chondrites have a big peak near 15×10^6 years. The distribution of exposure ages in Fig. 10 for L-chondrites shows no large peaks but has clusters near 5, 20–30, and 40 million years.

There have been fewer exposure ages measured for stony meteorites of other types, so often the statistics are too poor to see distinct peaks in their exposure-age distributions. The CM2 type of carbonaceous chondrites are very young with exposure-age groups near $\sim 5 \times 10^5$ and $\sim 1.5 \times 10^6$ years. Eight diogenites (all those measured

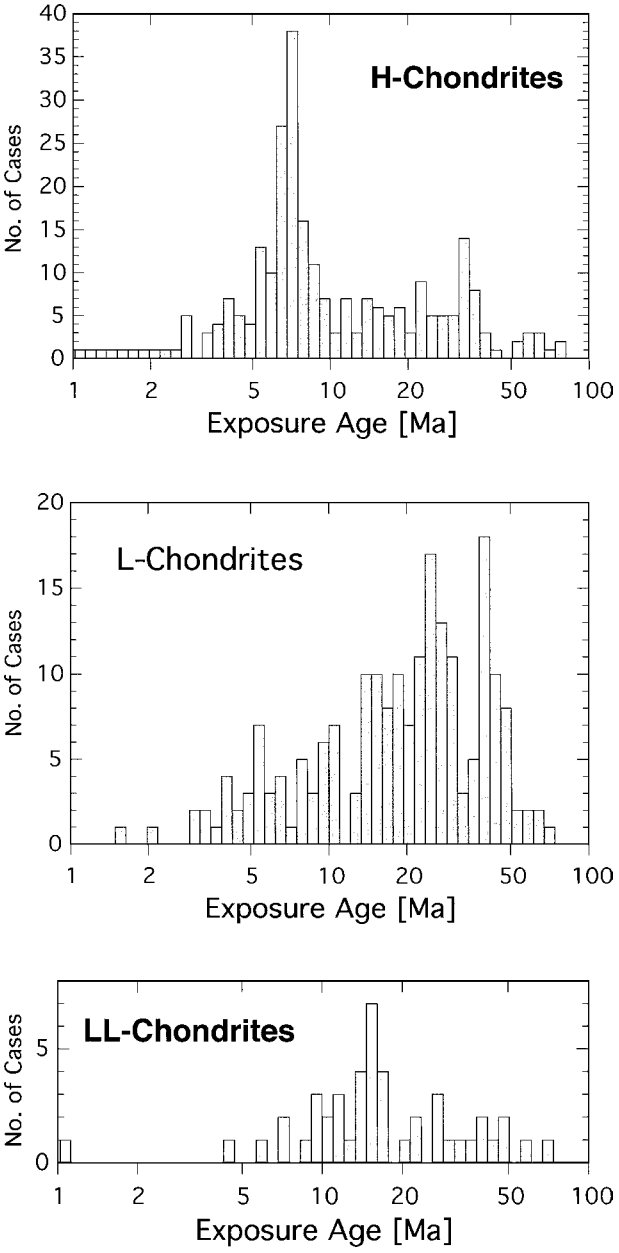


FIGURE 10 Cosmic-ray exposure ages (in 10^6 years) of the three major types of stony meteorites as calculated from measured ^{21}Ne concentrations and ^{22}Ne/^{21}Ne ratios by L. Schultz (private communication, 2000).

except Manegaon) cluster at exposure ages of $\sim 2.0 \times 10^7$ or $\sim 4.0 \times 10^7$ years. The howardites and eucrites have similar exposure-age distributions with several clusters between 1×10^7 and 4×10^7 years. Most of the aubrites have exposure ages near 5×10^7 years.

The exposure ages for three of the four Shergottites are near 2.5×10^6 years, and those of the three Nakhlites and of Chassigny are about 1.1×10^7 years. These relatively short exposure ages of the Shergottites (another Shergottite's exposure age is $\sim 1 \times 10^6$ years), the Nakhlites, and Chassigny, which are all collectively referred to as the "SNC" meteorites, are consistent with a hypothesis, based on a variety of other data, that they came from Mars. The lunar meteorites often have complex exposure records and appear to represent many ejection events on the Moon.

The carbonaceous chondrites tend to have shorter exposure ages than the other types of chondrites, with ages above 2×10^7 years being fairly rare. The ages generally are less for the lower petrological types of the carbonaceous chondrites. Achondrites are much more likely to have exposure ages above $\sim 3 \times 10^7$ years than are the chondrites. These trends, plus the fact that stony irons and the iron meteorites have much greater exposure ages than stones, indicates that a greater physical strength of a meteorite helps it to survive in space longer.

While about half of the H-chondrites seemed to have been produced over a very short period of time, probably by a single large event on a parent body, other meteorite types (such as the L-chondrites) have essentially a continuum for their exposure-age distributions. One question is whether the L-chondrites were really produced continuously or by a number of individual events. Given the width of the H-chondrite peak, it has been estimated that possibly as few as six events could account for most of the distribution shown for the L-chondrites in Fig. 10. The presence of all petrological types and also of gas-rich meteorites in the 6×10^6-year peak for the H-chondrites has been interpreted as implying that the parent body on which this event occurred was one that had earlier been nearly completely fragmented and then gravitationally reassembled into a megabreccia. The relatively large fraction (about 14%) of gas-rich H-chondrites in this peak suggests that this event mainly removed objects from near the surface of the parent body.

The exposure ages of iron meteorites are much greater than those of the stony meteorites and range from 4×10^6 years for the IB Pitts to 2.3×10^9 years for the anomalous ataxite Deep Springs. Very few iron meteorites have exposure ages greater than 1.0×10^9 years. The exposure-age frequencies of several types of iron meteorites are shown in Fig. 11. These ages are based on the ^{40}K/^{41}K ages as determined by H. Voshage using ^{4}He/^{21}Ne ratios for shielding corrections. (In Fig. 11, the scarcity of exposure ages less than $\sim 2 \times 10^8$ years is partially because of the difficulty of detecting the relatively small amount of ^{40}K. Other measurements indicate that roughly 10% of iron meteorites have exposure ages below 10^8 years.) The most prominent peak is near 6.75×10^8 years for the IIIAB iron meteorites (medium octahedrites) and has a width of about $\pm 1.0 \times 10^8$ years. The IVA irons (hexahedrites) have exposure ages that cluster near 4.5×10^8 years. The IIA irons have relatively low ages (most below 3×10^8 years). The exposure-age

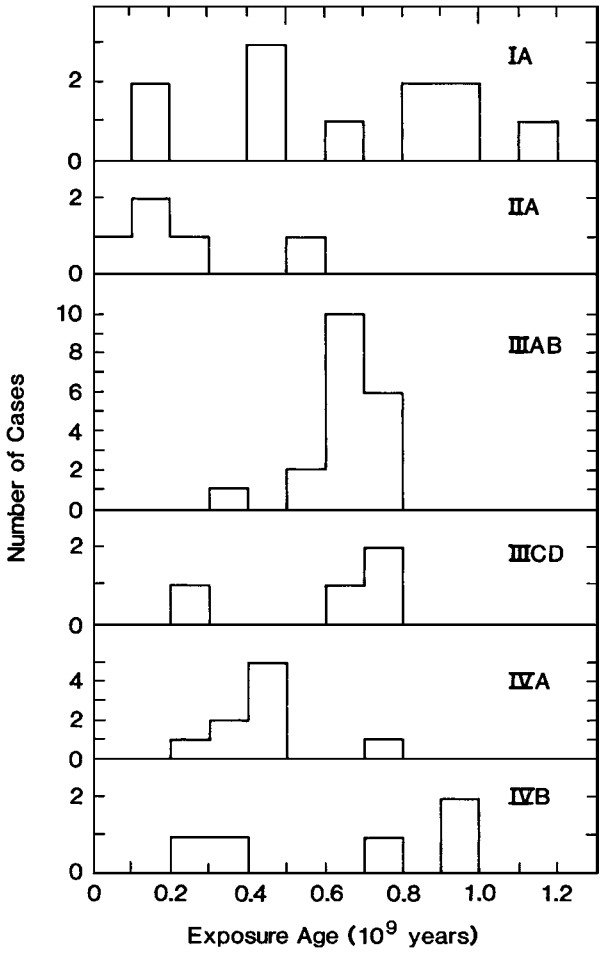


FIGURE 11 Cosmic-ray exposure ages of six types of iron meteorites as calculated from measured ratios of $^{40}K/^{41}K$ and $^{4}He/^{21}Ne$ by H. Voshage.

distributions vary widely among iron-meteorite types. Exposure ages for stony irons are intermediate between those for stones and those for irons. Stony irons with exposure ages greater than 2×10^8 years are fairly rare.

Noble-gas measurements have shown that major losses of the helium and argon made by the decay of the natural radioisotopes of uranium, thorium, and potassium have occurred occasionally in the past. For example, about a third of the L-chondrites lost most of their noble gases in a major collision of their parent body $\sim 5 \times 10^8$ years ago. The exposure ages of these L-chondrites with low gas-retention ages cover the whole range shown in Fig. 10. The more recent events dated by the exposure ages usually did not cause much loss of such radiogenic gases. The general lack of shock features, such as gas losses, associated with the events that initiated the cosmic-ray exposures of most meteorites indicates that most meteorites were probably ejected from their parent bodies with low velocities (about 1 km/s or less). Several mechanisms that can eject meteorite-sized fragments to planetary escape velocities with minimal shock effects (such as spalling or jets of volatiles) have been proposed to account for the lunar and martian meteorites, which require escape velocities of 2.4 and 5 km/s, respectively.

In addition to dating the collision events that produced these meteorites, the exposure ages also reflect the probability that these objects eventually hit the Earth before being destroyed in space. The long exposure ages for Norton County and the other relatively mechanically weak aubrites could indicate that they had orbits with low probabilities for collisions with other objects in space. While the much longer exposure ages for iron meteorites relative to the stones probably is due to their mechanical strength, it also could be partially caused by their places of origin. The short exposure ages of the lower petrological types of carbonaceous chondrites (CI and CM) have been interpreted as being possibly due to their origin from comets, which have orbital lifetimes of 10^6 to 10^7 years. Meteoroids in Earth-crossing orbits have been calculated to have lifetimes of 10^7 to 10^8 years, similar to exposure ages of stony meteorites. The exposure ages of the iron meteorites, 10^8–10^9 years, are similar to the lifetimes for Mars-crossing asteroids. Thus, exposure ages of meteorites are consistent with the generally accepted hypothesis that most meteorites come from asteroids.

SEE ALSO THE FOLLOWING ARTICLES

ASTEROID IMPACTS AND EXTINCTIONS • COSMIC RADIATION • METEORITES • NOBLE–GAS CHEMISTRY • NUCLEAR CHEMISTRY • RADIOCARBON DATING • SOLAR PHYSICS • STABLE ISOTOPES AS TRACERS OF GLOBAL CYCLES • THERMOLUMINESCENCE DATING

BIBLIOGRAPHY

Benoit, P. H., and Chen, Y. (1996). *Radiat. Meas*. **26,** 281.
Benoit, P. H., and Sears, D. W. G. (1997). *Icarus* **125,** 281.
Fleischer, R. L., Price, P. B., and Walker, R. M. (1975). "Nuclear Tracks in Solids," University of California Press, Berkeley.
Goswami, J. N. (1991). Solar flare Heavy-Ion Tracks in Extraterrestrial Objects. *In* "The Sun in Time" (C. P. Sonett, M. S. Giampapa, and M. S. Matthews, eds.), pp. 426–444. University Arizona Press, Tucson.
Herzog, G. F. (1994). *Nucl. Instrum. Methods Phys. Res*. **B92,** 492.
Lal, D. (1972). *Space Sci. Rev*. **14,** 3.
Lavielle, B., Marti, K., Jeannot, J.-P., Nishiizumi, K., and Caffee, M. (1999). *Earth Planet. Sci. Lett*. **170,** 93.
Marti, K., and Graf, T. (1992). *Annu. Rev. Earth Planet. Sci*. **20,** 221.
Reedy, R. C., Arnold, J. R., and Lal, D. (1983). *Science* **219,** 127; *Annu. Rev. Nucl. Part. Sci*. **33,** 505.
Vogt, S., Herzog, G. F., and Reedy, R. C. (1990). *Rev. Geophys*. **28,** 253.

Moon (Astronomy)

Bonnie J. Buratti
Jet Propulsion Laboratory,
California Institute of Technology

GLOSSARY

Apogee Farthest distance from the earth in the moon's orbit.
Barycenter Center of mass of two bodies, around which they revolve.
Bond albedo Fraction of the total incident radiation reflected by a celestial body.
Geometric albedo Ratio of the brightness at full moon compared with a diffuse, perfectly reflecting disk.
Librations Wobbles in the moon's orbit, which cause more than half of it to be visible from earth.
Mare (pl. maria) Low albedo plains of the moon.
Nodes Two points in the moon's orbit, at which it intersects the ecliptic.
Perigee Closest distance to the earth in the moon's orbit.
Planetesimals Bodies up to kilometers in size, which populated the solar system before the planets accreted.
Primary Celestial body around which a satellite, or secondary, orbits.
Regolith Layer of debris and dust, which covers the moon.
Saros cycle Period during which the lunar and solar eclipses repeat (18.03 years).
Sidereal month Period of revolution of the moon about the earth with respect to the stars (27.32 days).
Synchronous rotation Dynamical state caused by tidal interactions in which the moon presents the same face toward the earth.
Synodic month Period of revolution of the moon about the earth with respect to the sun (29.53 days).
Uplands Heavily cratered, higher albedo portions of the lunar surface.

THE MOON is the planet earth's only natural satellite. It orbits the earth approximately once every month as the two

TABLE I Gross Characteristics of the Moon

Mean distance from earth (semi-major axis)	Rotation period	Revolution period	Radius	Mass
384.4×10^3 km	27.32 days	27.32 days	1738 km	7.35×10^{22} kg 0.012 of earth's

Density	Visual magnitude	Visual geometric albedo	Bond albedo
3.34 gm/cc	−12.5	0.11	0.12

bodies journey together about the sun. The moon keeps the same face turned toward the earth because it is tidally evolved. The quest to understand the phases of the moon, the cycle of lunar and solar eclipses, the ocean tides, and the motion of the moon in the sky served as a basis for early scientific investigation. The moon has also figured prominently in the legends and lore of the world's peoples.

I. INTRODUCTION

Table I summarizes the gross characteristics of the moon. The earth-moon system is somewhat unusual in that the mass of the moon in comparison to its primary is large (0.012). Charon, the only known companion of Pluto, is the only satellite with a larger relative mass (about 0.1).

In 1610 Galileo first observed the moon through a telescope. He perceived the bright, rough lunar highlands, which were named after terrestrial mountain chains, and the darker lunar plains, or maria (Latin for seas), which were givelhhhiful names such as the Sea of Tranquility, Sea of Humors, and Sea of Clouds. Lunar craters are generally named after famous scientists. Figure 1 in the article within this encyclopedia, "Planetary Satellites, Natural" shows the near side of the moon as seen through a large telescope.

Although the moon is the only celestial body to have been visited by human beings, many questions, including that of its origin, and the existence of water ice on its surface or on its mantle, remain unanswered.

II. PROPERTIES OF THE MOON

A. Size, Shape, and Distance

Aristarchus of Samos, a Greek philosopher of the third century B.C., who ascribed to the heliocentric view of the universe, attempted to measure the relative distances between the earth and the sun and moon by noting the angular distance between the latter two bodies when the moon as seen from the earth was half illuminated, that is, during quarter phase. Although his estimate was grossly

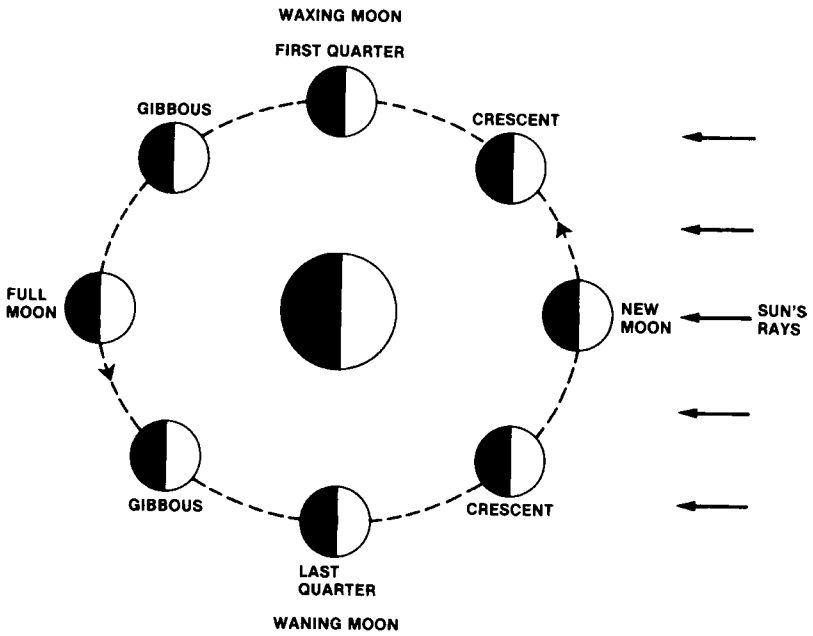


FIGURE 1 A schematic diagram of the phases of the moon as seen from earth (center).

inaccurate due to the difficulty of measuring this angular separation, he correctly deduced that the sun was much farther away from the earth than the moon. In the next century, Hipparchus made estimates accurate to within a few percent of both the distance between the earth and moon and the moon's diameter based on a measurement of the moon's parallax.

The experimental techniques developed by the Greeks were lost in the course of Europe's dark ages and the rise of the more abstract Aristotelean school during the Middle Ages (although Arab and Chinese scientists continued to conduct astronomical observations). It was not until the emergence of modern astronomy and the heliocentric world view in the sixteenth and seventeenth centuries that significant progress was made in understanding the nature of the cosmos. The distance of the moon from the earth and its size continued to be estimated by measuring its angular size and parallax until the twentieth century.

The development of radar during World War II led to accurate measurements of the distances to the moon and nearby planets. By measuring the delay required for a radio signal to be returned to earth, scientists calculated the average distance between the earth and the moon to be 384,403 km. It is possible for present laser ranging techniques to measure distances to the moon to within an error of 3 cm. Radar ranging measurements give topographic profiles of lunar craters and mountains with typical errors of tens of meters. Accurate measurements of the angular size of the moon yield a value of 3476 km for the lunar diameter.

The mass of the moon relative to the earth is inversely proportional to the ratio of their distances from the barycenter of the system (the barycenter is the center of mass around which the two bodies revolve). More accurate determinations of the lunar mass can be obtained by measuring the perturbations the moon exerts on spacecraft. The moon is about $\frac{1}{18}$ as massive as the earth, or 7.35×10^{25} gm. The mean density of the moon is 3.34 gm/cm^3, less than the earth's value of 5.52. The lower value for the moon is due to the absence of a large metallic core, and is consistent with a predominantly rocky composition.

The moon is not a perfect sphere. Its earth-facing radius is 1.08 km larger than its polar radius, and 0.2 km larger than its radius in the direction of motion.

B. The Motion of the Moon in the Sky

The moon executes a combination of complex motions, both real and apparent, as it orbits the earth. The period of revolution of the moon about the earth with respect to the stars is 27.32 days and is known as a sidereal month. Because the earth–moon system moves $\frac{1}{13}$ of the way around the sun in one sidereal month, the moon must travel another 27° in its orbit about the earth to be in the same position in the sky with respect to the sun. This period is known as the synodic month and takes 29.53 days to complete.

The moon, like other celestial objects, rises in the east and sets in the west. But each night, it is seen to move approximately 13° to the east with respect to the background stars. This apparent motion is due to the moon's revolution about the earth once every lunar month. The revolution of the earth–moon system about the sun at the rate of about 1° per day causes the path of the moon to be displaced only 12° eastward with respect to the sun. This eastward motion of the moon causes it to cross the local celestial meridian an average of 50 min later each day. Because the velocity of the moon in its orbit about the earth varies in accordance with Kepler's second law (which states that a celestial body will sweep out equal areas of its elliptical orbit in equal times), this successive delay ranges from 38 to 66 min. However, for moonrise and moonset, this delay varies (depending on the geographical latitude of the observer) from only several minutes to over an hour. This difference arises because the path of the moon is in general not perpendicular to the horizon. Thus, on successive nights the moon does not have as far to travel in its apparent motion from below to above the horizon. In mid or far northern latitudes, the extreme example of this phenomenon is the harvest moon, which occurs near the vernal equinox, when the angle subtended by the path of the moon with the horizon is a minimum. On the several nights occurring near the full moon closest to the vernal equinox, the brightly lit moon continues to rise shortly after sunset to provide extra time to farmers reaping their fall crops.

The height of the moon above the horizon varies markedly throughout the year. Like the sun, the moon is higher in the sky during the summer than in the winter, due to the 23.5° inclination of the earth's equator to the plane of its orbit about the sun (the ecliptic). In addition, the moon's orbit about the earth is inclined 5° to the ecliptic. The moon's height also depends on the geographical latitude of the observer. At extreme northern or southern latitudes, the moon appears lower in the sky than near the earth's equator.

Although the moon keeps the same face turned toward the earth, 59% of the moon's surface is visible from the earth because of librations, or wobbles, in the lunar orbit. Geometric librations are caused by three factors in the orbital relations between the earth and moon. A libration in latitude is a consequence of the inclination of 6.7° between the lunar spin axis and orbital plane. A libration of longitude, which causes an additional 7.6° to be seen, is due to the fact the law of inertia constrains the rotational velocity of the moon to be uniform, whereas its orbital velocity varies in accordance with Kepler's second law. The third geometrical libration, known asthe diurnal libration, is

TABLE II Characteristics of the Lunar Orbit

Apogee	Perigee	Mean eccentricity	Inclination to earth's equator	Obliquity
405,500 km	363,300 km	0.055	18 to 29°	6.7 degrees

attributable to the fact the radius of the earth subtends nearly a degree of arc at the distance of the moon. Therefore, the daily rotation of the earth on its axis causes an observer to view a slightly different aspect of the lunar surface. Additional smaller librations, known as physical librations, are caused by pendulumlike oscillations induced by the earth in the moon's motions due to the latter's nonspherical shape.

C. Orbital Motions

The major characteristics of the moon's orbit are summarized in Table II. The theory that describes the complete motion of the moon is complex: only the basics are outlined here. The earth and moon revolve about their center of mass, or barycenter. As a first approximation, the path of the moon about the earth is an ellipse with the earth at one of the foci. The eccentricity of the moon's orbit averages 0.0549, but varies from 0.044 to 0.067 due to perturbations by the earth and sun. The closest approach of the moon to the earth (perigee) is 363,000 km and the maximum distance (apogee) is 405,500 km. Tidal forces exerted by the sun cause the semi-major axis, or the line of apsides, to rotate eastward every 8.85 yr.

The plane defined by the moon's orbit intersects the ecliptic at two points known as the nodes. The ascending node is defined by the moon's motion northward as it rises above the ecliptic as seen by a northern observer, and the descending node occurs when the moon moves below the ecliptic. Perturbations of the moon by the earth and sun cause the line of nodes, defined by the intersection of the orbits of the earth and moon, to move westward along the ecliptic every 18.6 yr in a motion called the regression of the nodes. This 18.6-yr cycle is known as the nutation period. Two other cycles are the draconitic month, which is the time required for the moon to return to the same node, and the anomalistic month, or the period between two perigee passages.

The regression of the nodes causes the earth's axis to wobble as much as 9 sec. Additional perturbations cause the inclination of the moon's orbit to the ecliptic to vary between 4° 57 min and 5° 20 min.

D. Phases of the Moon

In the course of the lunar month, the moon presents a different fraction of its illuminated face to an observer on earth. Figure 1 is a schematic diagram of the resulting phases of the moon and their names. The moon is said to be waxing when the illuminated face is growing, and waning when it is progressing toward new moon. The quest for an understanding of the moon's phases and eclipses was perhaps the primary impetus for the development of ancient astronomy. By the fourth century B.C., Aristotle presented a clear explanation of these phenomena, although Anaxagoras a century earlier was probably the first western astronomer to offer correct explanations.

The various phases of the moon rise and set at specific times of the night (or day). The full moon always rises shortly after sunset. A waning gibbous moon rises later and later in the evening and the moon at last quarter rises around midnight. The new moon rises at dawn and sets shortly after sunset. The first quarter rises in midday and the waxing gibbous moon rises later and later in the afternoon until it is finally full.

E. Eclipses

Solar and lunar eclipses were the root of much lore and superstition in the ancient world. Both ancient eastern and western astronomers realized that eclipses occurred in cycles and accurately calculated their dates.

When the earth passes between the sun and the full moon and all three objects are in a straight line, the Earth's shadow is cast on the face of the moon to cause a lunar eclipse (Fig. 2). The earth's shadow has both a full zone (the umbra) and a partial zone (the penumbra). If the moon is totally immersed in the full terrestrial shadow, a total lunar eclipse results. If the moon is in the penumbra, the eclipse is said to be penumbral, and is in fact barely visible. If the moon is partly illuminated, the eclipse is said to be partial. A lunar eclipse can last as long as 3 hr, 40 min with the duration of totality as much as 1 hr 40 min. A lunar eclipse is visible to an observer on any point of the earth's nightside portion. During an eclipse, the moon is bathed in a reddish glow, which is due to light refracted by the earth's atmosphere into the earth's shadow.

The more spectacular solar eclipse occurs when the new moon passes in front of the sun (Fig. 2). Because the angular sizes of the moon and sun are both about a half of a degree, the lunar disk barely obscures the sun to reveal the spectacular sight of the solar corona. The moon's full shadow on the earth is at most 267 km wide, and even the area of partial eclipse, where the observer is in the moon's penumbra, is about 6500 km wide. The duration of totality is always less than 7.5 min. Lunar eclipses are thus more frequently observed than solar eclipses at a specific terrestrial location. If the moon is close enough to the earth so that its angular size is less than that of the sun, the result is an annular eclipse, in which a ring of the sun's disk encircles the moon.

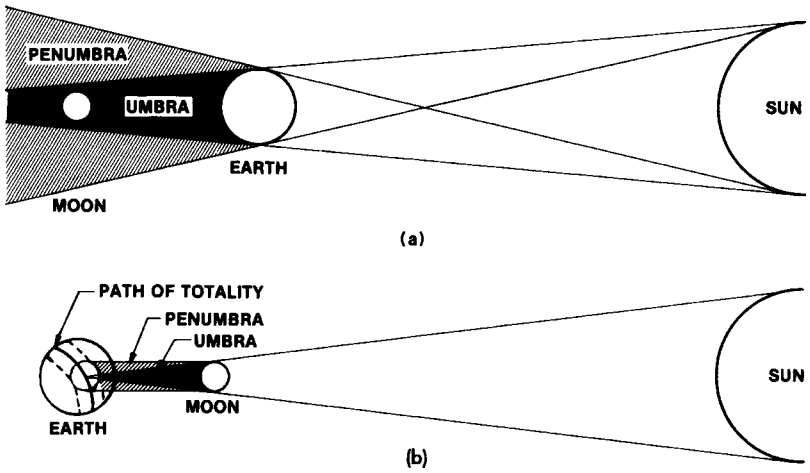


FIGURE 2 (a) A lunar eclipse occurs when the earth's shadow covers the face of the full moon. (b) A solar eclipse occurs during new moon, when the moon's shadow is cast on the earth.

Eclipses occur in cycles of 18.03 yr because they are possible only when both the moon and sun are close to the nodes of the lunar orbit. One eclipse period, also known as the Saros cycle, consists of a whole number of both synodic months and eclipse years (the time required for the sun to return to a node in the moon's orbit, or 346.62 days).

F. The Tides

Gravitational interactions between the earth and the moon cause the well-known phenomenon of the oceanic tides (Fig. 3). The gravitational forces experienced by the side of the earth facing the moon is greater than that between the moon and the center of the earth, which in turn is greater than that exerted on the far side of the earth. Thus the differential gravitational forces of the moon on the earth act to flatten the earth with the long axis pointing toward the moon. The liquid seas are influenced more by tidal forces than the solid earth. The sun also exerts tidal forces on the earth (although they are not as great because they depend on the inverse cube of the distance between the two bodies). At full and new moon, when all three bodies are in a line, spring, or maximum tides result. Similarly, during first and last quarter, the minimum neap tides occur (see Fig. 3).

G. The Temperature of the Moon

The surface of the moon undergoes greater extremes of temperature than the earth, because it has no significant atmosphere and no ocean to temper the effects of the sun, and because it rotates on its axis only once every month. The temperatures on the moon range from a daytime high of 380 K to nighttime lows of 100 K.

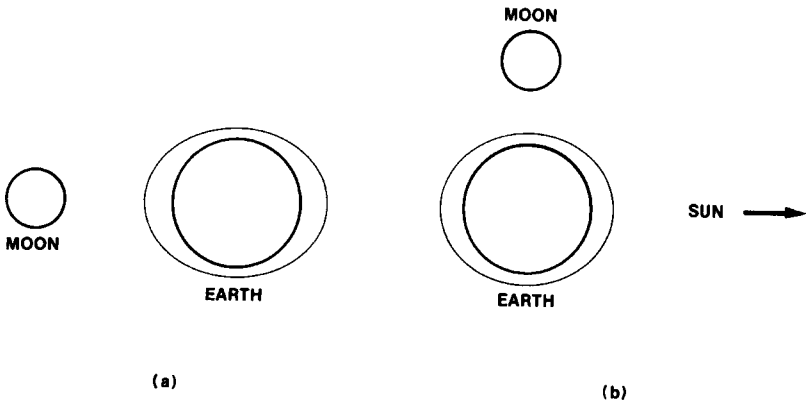


FIGURE 3 The tidal bulges induced on the earth by the moon and sun. Spring tides (a) occur when all three bodies are in a straight line. The minimum or neap tide (b) is seen during the quarter phases of the moon. The earth's rotation causes high tides to appear twice each day.

H. Photometric Properties

The moon is the second brightest celestial object in the sky, yet it is only little more than 2 millionths of the brightness of the sun. The fraction of visible light that the full moon reflects back to an observer (the geometric albedo) is about 0.1. The intensity of the moon varies markedly over its surface: the lunar maria reflect only 6.5–9% of visible light, whereas the lunar uplands and crater bottoms reflect about 11%, and the bright ray craters 13–17%. The Bond albedo (the disk-integrated amount of total radiation emitted compared with that received) is 0.12. The dark portion of the moon is illuminated slightly by earthshine, which is light reflected from the surface of the earth back to the moon.

The moon is associated with a number of striking optical phenomena. On clear dry days in the spring, a ring is often seen around the moon appearing 22° from its disk. The ring is due to the refraction of moonlight by small hexagonal ice crystals high in the earth's atmosphere. Popular belief states that the ring foretells rainy weather: there is a modicum of truth in this belief because cirro-stratus clouds, which are composed partly of ice crystals, are associated with low atmospheric pressure. The deep orange color of the moon when it is near the horizon is caused by the greater degree of scattering of blue light by the atmosphere, which has a much higher optical depth when the observer looks toward the horizon. The rare blue moon is a contrast phenomenon caused by the position of the moon next to pink clouds. ("Blue moon" also refers to a second full moon in a single month.)

III. SUMMARY OF OBSERVATIONS

A. Earth-Based

Because of its proximity and its influence on the earth, the moon is perhaps the most studied celestial body. Detailed studies of the moon's motion, its appearance in the sky, its role in causing the tides and eclipses, has occupied the time of astronomers since ancient times.

In 1610, Galileo was the first to observe the moon through a telescope: he discerned the delineation of the moon into the bright uplands (which he called terrae, or land) and the darker plains (which he called maria or seas). The moon was first photographed in 1840. However, most of the detailed mapping of lunar features from the earth has been done visually, because instabilities in the earth's atmosphere constrain the resolution in photographic exposures to be only a fourth of that possible with the human eye. From the earth, a resolution of about 2 km on the moon's surface is possible with the best telescopes and ideal seeing conditions.

Observations from the earth of the visible and near infrared spectrum of the moon suggested its surface is composed of basaltic minerals. Analysis of the lunar phase curve, which is the variation in disk-integrated brightness with changing phase, suggested the existence of mutual shadowing among the particles in the optically active surface, which is consistent with a fluffy soil. Infrared measurements of how fast the surface loses heat as it comes out of eclipse also pointed to a porous surface. Infrared measurements demonstrated that the moon does not have an internal heat source.

B. Spacecraft Reconnaissance

A series of reconnaissance craft to the moon sent by United States and the Russia began in 1959, when the Soviet Luna 3 photographed the far side of the moon (see Table II in the article within this encyclopedia "Planetary Satellites, Natural"). The Ranger program, which was a series of impact probes sent to the moon by the United States between 1961–65, returned many thousands of live television images of the moon's surface, some with 1-m resolution. Five Lunar Orbiter missions launched by the United States between 1966 and 1968 successfully returned high resolution images of 95% of the lunar surface. These spacecraft also mapped gravity anomalies below the moon's surface.

The first spacecraft to land successfully on the moon was the Russian Luna 9 in 1966. The United States successfully landed five spacecraft on the moon as part of its Surveyor program. They sent back over 87,000 images, and studied the composition of the lunar soil with an alpha-particle spectrometer. Between 1970 and 1976, three of the Russian Luna spacecraft sent back to the earth small samples of lunar rock.

A renewed program of remote exploration of the moon was initiated in 1994 by the Clementine Mission, a joint venture of NASA and the Department of Defense. The payload included five instruments that obtained images covering the ultraviolet, visual, and infrared regions of the spectrum, as well as a LIDAR unit and a particle detector. During a two-month mapping period that covered the entire lunar surface, the spacecraft obtained over two million images. In 1998, Lunar Prospector obtained a global compositional map of the moon and mapped its gravity and magnetic fields. Both Clementine and Prospector obtained evidence that water ice exists under the surface of the moon.

C. Manned Exploration

The United States successfully launched six manned Apollo missions to the moon between 1969 and 1972. The major scientific goals of the Apollo mission were to return rock samples for geochemical and morphological analysis, to obtain photographic coverage of the moon for geologic study, to place seismometers, magnetometers, heat

flow instruments, gravimeters, and other experiments on the surface, and to employ an array of spectrometers for compositional studies.

Extensive photographic coverage and manned exploration of the moon has transformed it into a tangible geologic world. Although the vast amount of data has shed light on the basic problems of lunar morphology, mineralogy, and geophysical evolution, some questions, such as the structure of the moon's interior and the origin of the moon, remain unanswered.

IV. THEORIES ON THE MOON'S ORIGIN

A. Introduction

The important observational constraints that must be accounted for by a model for the origin of the earth–moon system include: an anomalously large ratio of satellite to primary mass; a lower mean density for the moon than the earth, which indicates the absence of a significant iron–nickel core for the moon; important chemical differences in the lunar crust such as enrichment of refractory elements (especially aluminum, calcium, and titanium); and depletion of volatile materials such as water, sulfur, sodium, and potassium.

The four standard theories for the origin of the moon are (1) capture, in which the moon forms elsewhere in the solar system and is gravitationally captured by the earth; (2) fission, which asserts that the moon broke off from the earth early in its formation; (3) coaccretion,which asserts that the earth and moon formed independently but nearly simultaneously near their present locations; and (4) impact and reaccretion, in which a second body collides with the young earth to create a debris disk that rapidly accretes to form the moon.

B. Capture

The capture hypothesis explains the chemical differences between the earth and moon by having the moon form in an area of the solar system where refractory elements were rich and volatile elements rare (presumably nearer the sun). However, the likelihood of gravitationally binding such a body in a closed orbit about the earth is small, particularly if the body is not in the earth's immediate neighborhood. In a more recent variation of the capture theory, planetesimals were captured by the earth and later accreted to form the moon.

C. Fission

In the fission model, the proto-earth had so much angular momentum after the iron core was formed that it was able to fling off a mass of material that became the moon. The model was first developed by George Darwin in 1898, who discovered that the moon was once closer to the earth and has been gradually receding due to tidal interaction (see Section IX.A). The fission model explains the chemical differences between the two bodies by having fissioned material consist entirely of the primordial earth's mantle. The problems with the model include explaining how the earth acquired enough angular momentum to fling off a mass as large as the moon, and why the moon revolves in a plane inclined to the earth's own spin equator. In one variation of the fission model, which overcomes these difficulties, a large body, perhaps $\frac{1}{10}$ as massive as the earth, impacted the planet at a grazing angle and blew material into earth's orbit, which then accreted to form the moon.

D. Coaccretion

In the coaccretion model, the moon formed separately from the earth, most likely from debris, which was in orbit around the proto-earth. The large relative mass of the moon is obtained by having a body being captured and subsequently sweeping up material. The major problem with this model is that it is difficult to explain the differences in chemical compositions between the two bodies if they were formed from the same association of materials. One way of overcoming this difficulty is to have heavier iron-rich debris preferentially accreted onto the earth.

E. Impact and Reaccretion

In this scenario for the moon's origin, a Mars-sized body collided with the proto-earth at the end of the accretionary phase of planetary formation. The resulting debris disk rapidly (within approximately one day) reaccreted to form the moon. This model of the formation of the moon is currently favored because it best accounts for the observational constraints. The formation of the impacting body on a higher temperature region of the proto-solar nebula causes the enrichment of the moon's refractory materials.

V. THE LUNAR SURFACE

A. Surface Features and Morphology

1. The Lunar Uplands

The basic fact of lunar surface morphology is that the moon is divided into two major terrains: the older, brighter heavily cratered lunar uplands and the darker, younger maria. This dichotomy is responsible for the appearance of the "man in the moon" (other cultures have seen a woman

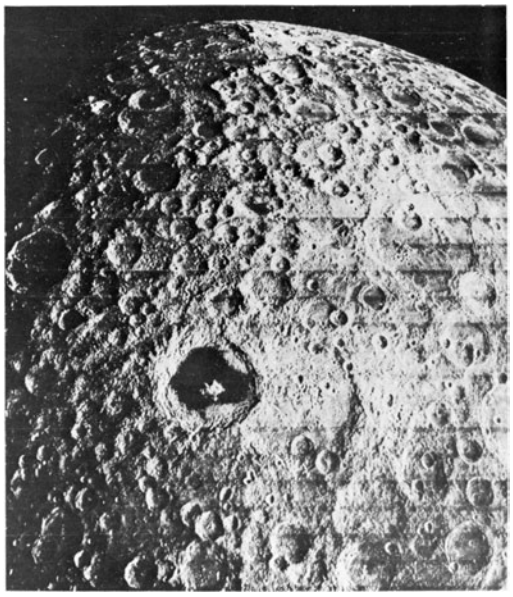


FIGURE 4 A Lunar Orbiter picture of the southwest portion of the moon's far side. Heavily cratered highlands dominate this hemisphere. The crater Tsiolkovsky near the center of the picture is filled with dark mare material.

in the moon, or a hare). The uplands cover about 80% of the lunar surface as a whole and nearly 100% of the far side.

Figure 4 depicts the rugged, cratered appearance typical of the uplands. Some areas are saturated with craters. The oldest rocks date from the period of the formation and differentiation of the moon (4.3–4.6 billion years ago). Most of the uplands consists of anorthositic gabbro, (a coarse grained, sodium–calcium rich silicate) which has been pulverized by impacts and fused together by shock metamorphism into a rock known as breccia. When compared to the maria, the uplands are richer in so-called KREEP basalt, which is higher in potassium (K), rare-earth elements (REE), and phosphorus (P). The highlands are also enriched in aluminum and depleted in magnesium and titanium.

A large area of the lunar uplands consists of the Cayley formation, a smooth higher albedo unit interspersed among the rugged cratered terrain. Although the Cayley formation shows some evidence of having been created by vulcanism, the highly brecciated samples returned by Apollo 16b suggest that it was formed by sheets of material ejected from a large impact event such as the formation of the Orientale Basin.

2. Maria and Basins

The lunar maria (or plains), which were formed between 3.1 and 3.9 billion years ago, are the youngest geologic units on the lunar surface, except for more recent impact craters. The release of heat from large impacts caused extensive melting and extrusion of basaltic lavas on the moon. In some cases the extrusions may have occurred in two stages: first from the impact melt itself and later from eruptions caused by subsequent heating. Several maria are large filled-in impact features, or basins, with clearly circular delineations. Some basins, such as Mare Orientale, are multiring structures.

Large mountain chains at the edges of the large basins were uplifted at the time of impact. Wrinkle ridges, which are believed to be compressional features formed near the end of the period of vulcanism, are seen in many of the maria.

The lunar maria are found primarily on the earth side of the moon. One possible physical explanation for the unequal distribution is that the maria formed in accordance with hydrostatic equilibrium: the moon's center of mass is closer to the earth's side and the crust on the farside is thus thicker.

3. Craters

Even when observed with a pair of binoculars or a small telescope, the moon appears covered with craters: about 30,000 are visible from earth. According to a tradition began by Riccioli in 1651, craters are named after well-known scientists. Lunar craters range in size from small millimeter sized pits caused by micrometeorites to large structures hundreds of kilometers in diameter (see Fig. 5).

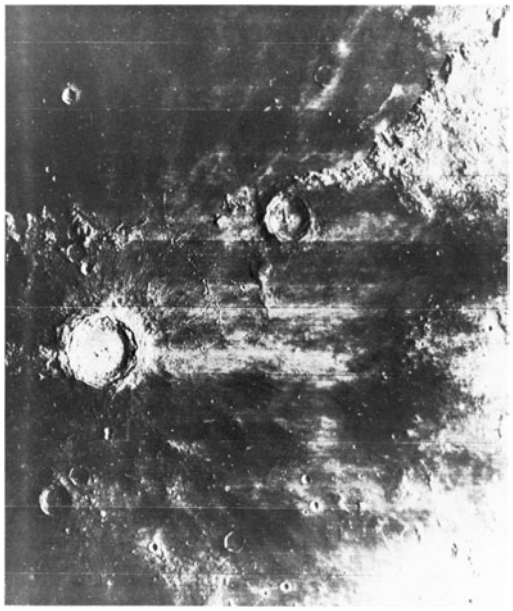


FIGURE 5 The rayed cratered Copernicus, showing an extensive ejecta blanket, secondary impacts, and central peak. The crater Erastosthenes is to the upper right of center. Several older craters in the picture have been filled in by lava.

Before the period of lunar exploration in the 1960s, one of the major questions of lunar science was whether the craters were of volcanic origin or caused by impacts. Based on morphological evidence, most geologists have determined that craters were caused by impacts. This evidence includes the existence of ejected material (known as ejecta blankets) around crater rims, the creation of secondary impacts in the vicinity of large craters, and the existence of central peaks, which are due to the rebound of material after a large impact.

Throughout its history, the moon has been bombarded by meteoroids ranging in size from a few microns to several kilometers. Most craters, particularly the large ones, date from the period of heavy bombardment, which ended its intense phase about 3.9 billion years ago and tapered off until 3.1 billion years ago. Since then, the moon has experienced only an occasional impact. Because geological processes on the moon do not include wind or water erosion or plate tectonics, which have oblitered nearly all of earth's craters, the moon's face bears the scars of its past history. Extending from the most recent craters, such as Tycho (0.27 billion years old), are bright rays caused by the disturbance and subsequent exposure of fresher, brighter subsurface material.

Geologists use crater-counting methods to date lunar surfaces. Although the techniques in practice are complex and involve many assumptions, the principle behind them is simple: if we assume a specific flux (not necessarily constant) of impacting objects throughout the moon's history, the age of a surface (i.e., when it solidified) is proportional to the number of craters on it.

Meter-sized lunar craters are often simple bowl-shaped depressions. As the size of the crater increases, a number of complex phenomena are seen. These include raised crater rims, the formation of central peaks, secondary impact craters caused by large ejected chunks, slumping of material down the crater walls, and raised blankets of ejected material encircling the crater (Fig. 5). In general, larger craters have a smaller depth-to-diameter ratio than small craters. The largest impacts form the basins, some of which have been filled in by lava flows.

One type of intriguing lunar feature is the dark halo crater, which is a small low-rimmed structure with areas of darker material extending from the rim. Evidence that the dark material may have been deposited during an eruption is given by their tendency to be associated with fissures.

4. Volcanic Features

A number of geologic formations on the moon have volcanic origins. One common feature is the rille, which is a collapsed lava tube or channel (Fig. 6). The Hadley Rille is a sinuous formation 125 km in length. Crater chains are often associated with rilles or fissures, and may be tens of kilometers long.

Domes or even cones are often seen in maria areas and are probably extinct volcanoes.

B. The Regolith

The upper few meters of the moon consists of dust and rubble, which is known collectively as the lunar regolith. This material is the debris from eons of impact events, which have "gardened" the surface. The regolith is fluffy—about 60–80% of its volume is void. It is rich in glassy spherules, which were formed by melting during impacts by micrometeorites.

Another erosional process occurring on the lunar surface is due to interactions by the solar wind, which tends to darken the upper regolith.

C. Water Ice on the Moon

In the 1960s calculations of the surface temperatures of the moon suggested that water frost would be stable in permanently shadowed craters at the lunar poles. Comets impacting the surface of the moon would provide a constant supply of water ice to the lunar surface, and the burial of this material by impact-ejecta would lock it up in the upper regions of the lunar crust. Both the Clementine LIDAR experiment and the Prospector's measurement of neutrons characteristic of hydrogen suggested that water ice is buried at the lunar poles.

VI. THE LUNAR INTERIOR

Knowledge of the lunar interior comes primarily from analysis of seismic waves detected by instruments placed on the lunar surface by Apollo astronauts. Interpretations of the data must in addition be consistent with the parameters defining the gross structure of the moon such as its density (3.34 g/cc) and its coefficient of moment of inertia (0.395). The moon's density is lower than that of the earth (5.52), which implies a deficiency of metallic components and a smaller, less dense core, and its moment of inertia is consistent with the differentiation of the moon into crust, mantle, and core. Lunar samples returned by the Apollo astronauts were indeed depleted in metallic elements and enriched in refractory materials.

Because the moon is smaller than the earth and thus cooled more rapidly during its formation, its crust is

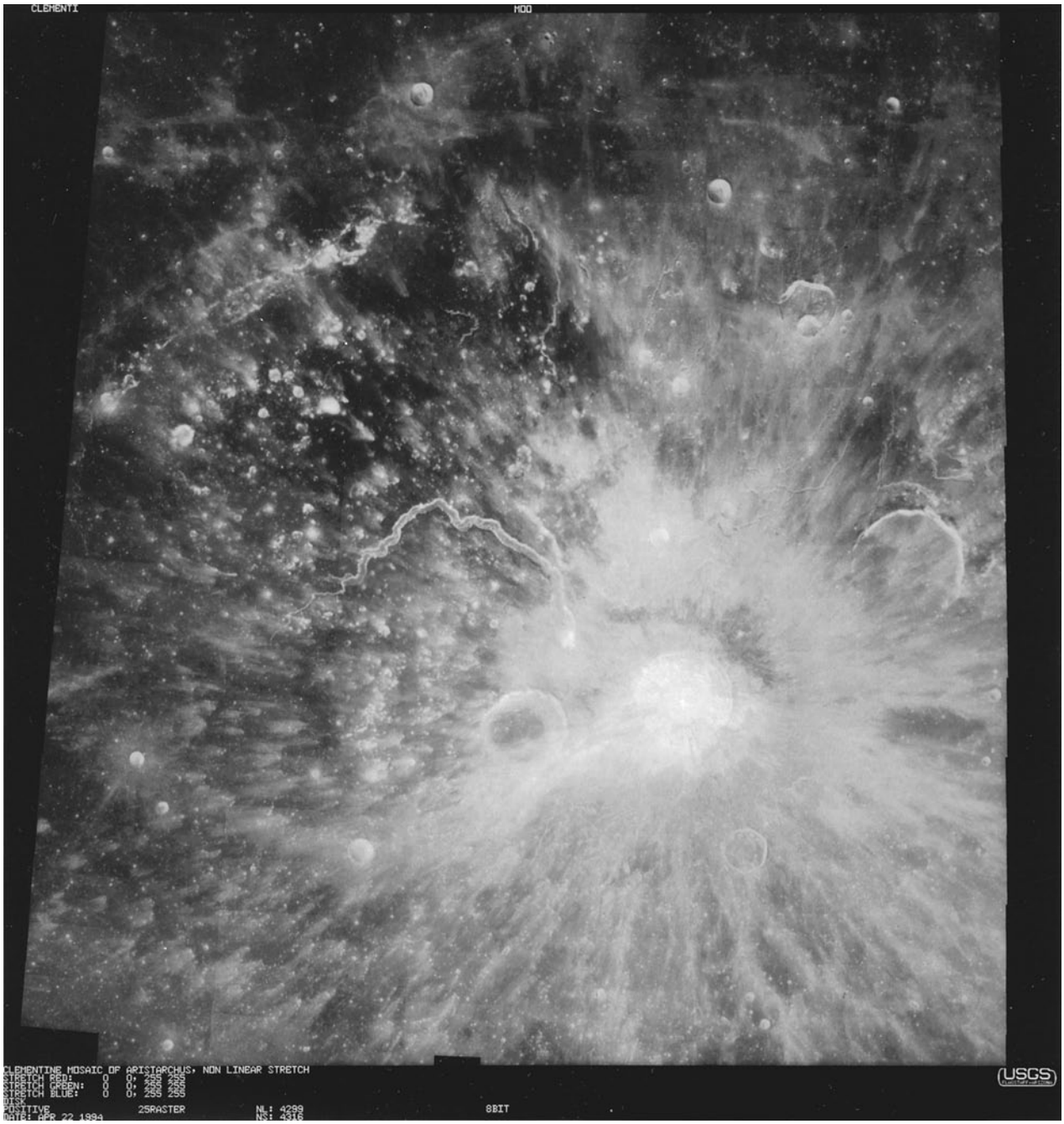


FIGURE 6 This Clementine mosaic of images shows the Aristarchus Plateau, one of the most geologically recent areas on the Moon. The cobrahead, a collapsed lava tube, appears in the middle left of the image. [Image processing by A. McEwen and the United States Geologic Survey.]

60 km—about 4 times as thick as that of the earth. The crust consists largely of anorthositic gabbro and mare basalts. The mantle is composed of ultramafic rocks (minerals enriched in heavy elements and depleted in silica). Seismic evidence indicates partial melting in the lower mantle. Moonquakes detected during the Apollo mission generally originated in the middle mantle (800–1000 km deep). In the core, which has a diameter between 300 and 700 km, attenuation of seismic s-waves, which do not propogate through a liquid, suggests that this part of the moon is molten. The existence of iron and perhaps nickel would give the core its required higher density.

During the Lunar Orbiter mission, the spacecraft experienced a number of anomalous accelerations in its orbit. These perturbations, which were observed only on the nearside, were attributed to the implacement of

higher density lava in large (>200 km) filled craters and maria. Geologists called these anomalies mascons (mass concentrations).

Instruments placed on the surface of the moon by Apollo 15 and 17 measured the heat flow from the lunar interior, which originates primarily from the decay of radioactive isotopes. The measurements indicate that the moon has a heat flow of only $\frac{1}{5}$ that of earth.

Clementine measurements of the lunar crust showed that the surface exhibits a 16-km range in elevation. The crust is significantly thinner under the maria.

VII. THE MAGNETIC FIELD

Planetary magnetic fields are believed to be caused by strong convective currents in planetary cores. These currents require for their production complete melting over of a large region of the planet's interior and rapid rotation of the planet. Because the moon has neither a large core or rotates rapidly, it has no appreciable magnetic field. However, analysis of remnant magnetism in lunar rock indicates that the moon had a weak magnetic field ($\frac{1}{20}$ that of the earth's current field) around the time of the formation of the maria (3.9–3.1 billion years ago).

VIII. THE LUNAR ATMOSPHERE

In 1987, scientists observed an extremely tenuous lunar atmosphere of sodium and potassium vapor. The surface density is 67 atoms per cubic centimeter for sodium and 15 atoms per cubic centimeter for potassium. The scale height of the atmosphere (the point at which the density falls to $1/e$ of its surface value) is about 100 km. The atmosphere is probably formed by the vaporization of lunar minerals.

IX. EVOLUTION OF THE MOON

A. Dynamical History

The moon is slowly receding from the earth. The cause for this recession is tidal interactions in the earth–moon system. Because the earth is spinning on its axis, the earth's tidal bulge closest to the moon (see Fig. 3) leads the line connecting the two bodies' center of masses. The moon exerts a greater force on this bulge, with the result that the spin rate of the earth, and thus its angular momentum, decreases. Because angular momentum in the system must be conserved, the moon is moving further away from the earth. Evidence that the lunar month and terrestrial day were shorter in the past is found in the fossilized shells of some species of sea corals, which grow in cycles following the lunar tides.

B. Geologic Evolution

1. Accretion and Early Bombardment

Radiometric dating of lunar samples indicates that the moon solidified about 4.6 billion years ago, at the same time the solar system formed. Accretional heating due to the release of gravitational energy during formation caused the moon to melt and differentiate into core, mantle, and crust. Toward the end of the accretional phase, the solid moon was bombarded by remnant planetesimals. The largest of these formed the large ringed impact basins. The heavily cratered portions of the lunar uplands date from the tail end of the bombardment phase.

2. Formation of Maria

Mechanical energy released from the impacting objects during the bombardment phase was sufficient to partially melt the moon's upper mantle. Between 3.1 and 3.9 billion years ago, the large impact basins were flooded with mare basalts. Some amount of localized mare vulcanism perhaps occurred as late as 2.5 billion years ago.

3. Later Evolution

The major geologic processes occurring on the moon today are infrequent cratering events and "gardening" of the regolith by micrometeorites and the solar wind (Section V.B.). Although the present cratering rate on the moon is a small fraction of what it was in the past, several well-known craters are of recent origin, such as Tycho (0.27 billion years old), and Copernicus (0.9 billion years old).

The small seismic events (1–2 on the Richter scale) recorded by the Apollo instruments are due to tidal stresses exerted on the moon by the earth or to continued contraction of the moon as it cools.

Although the moon is essentially geologically dead, there have been documented observations of possible small-scale current activity. In 1178, five Canterbury monks reported seeing a brightening near the upper horn of the crescent moon. This event may have been the impact that caused the formation of the crater Giordano Bruno. More recently there have been documented lunar transient events, including an observation in 1958 by a Russian astronomer of a gas emission spectrum from the crater Alphonsus, and an observation in 1963 by astronomers in Arizona of glowing red spots near the crater Aristarchus. These two events may have been due to outgassing caused by tidal stress or other factors. Another hypothesis is that thermal or seismic activity causes rock fracturing and associated electrodynamic effects, including pulses of

visible light. A bright flash which appeared to be on the surface of the moon was photographed by Greek astronomers in 1985, although subsequent calculations demonstrated the probability that the flash was light reflected from an earth orbiting satellite.

SEE ALSO THE FOLLOWING ARTICLES

CELESTIAL MECHANICS • LUNAR ROCKS • PLANETARY SATELLITES, NATURAL • SPACEFLIGHT PHYSIOLOGY • SPACE TRANSPORTATION SYSTEMS, ADVANCED • TIDAL POWER SYSTEMS

BIBLIOGRAPHY

Abell, G. (1969). "Exploration of the Universe," 2nd ed., Holt, Rinehart and Winston, New York.

Beatty, J. K., Petersen, C. C., and Chaikin, A., eds. (1988). "New Solar System," 4th ed., Cambridge Univ. Press, Cambridge, U.K.

Binder, A., and the Prospector Science Team (1994). *Science* **281,** 1475–1500.

Hartmann, W. K. (1988). "Moons and Planets," 4th ed., Brooks/Cole Publishing.

Heiken, G., Vaniman, D. T., and French, B. M. (1993). "The Lunar Sourcebook," Cambridge Univ. Press, Cambridge, U.K.

Nozette, S., and the Clementine Science Team (1994). *Science* **266,** 1835–1862.

Planetary Atmospheres

Joel S. Levine
NASA Langley Research Center

GLOSSARY

Atmospheric pressure Weight of the atmosphere in a vertical column, 1 cm^2 in cross section, above the surface of a planet. On earth, the average value of atmospheric pressure at sea level is 1.013×10^6 dyne cm^{-2}, or 1013 mbar, which is equivalent to a pressure of 1 atmosphere.

Cosmic abundance of the elements Relative proportion of the elements in the cosmos based on abundances deduced from astronomical spectroscopy of the sun, the stars, and interstellar gas clouds and chemical analyses of meteorites, rocks, and minerals.

Gravitational escape Loss of atmospheric gases from a planetary atmosphere to space. If an upward-moving atmospheric atom or molecule is to escape the gravitational field of a planet, its kinetic energy must exceed its gravitational potential energy. The two lightest atmospheric gases, hydrogen and helium, usually possess enough kinetic energy to escape from the atmospheres of the terrestrial planets. In photochemical escape, some heavier atmospheric species, such as atomic nitrogen and atomic oxygen, are imparted with sufficient kinetic energy from certain photochemical and chemical reactions to escape from planetary gravitational fields. Over geological time, gravitational escape has been an important process in the evolution of the atmospheres of the terrestrial planets.

Greenhouse effect Increase in the infrared opacity of an atmosphere which leads to an increase in the lower atmospheric and surface temperature. For example, water vapor and carbon dioxide, the two most abundant outgassed volatiles, increase the infrared opacity of an atmosphere by absorbing outgoing infrared radiation emitted by the surface and lower atmosphere. The absorbed infrared radiation is then re-emitted by the absorbing molecule. The downward directed component of the re-emitted radiation heats the surface and lower atmosphere.

Magnetosphere Region in upper atmosphere of planet possessing magnetic field where ions and electrons are contained by magnetic lines of force. The earth and Jupiter are surrounded by magnetospheres.

Mantle One of the three major subdivisions of the earth's interior (the core and the crust being the other two). The mantle contains about 70% of the mass of the earth and is iron deficient. The mantle surrounds the core, which is believed to consist mainly of iron. Surrounding the

mantle is the relatively thin-layered crust. The core, mantle, and crust are composed of refractory elements and their compounds.

Mixing ratio Ratio of the number of atoms or molecules of a particular species per cm^3 to the total number of atmospheric atoms or molecules per cm^3. At the earth's surface, at standard temperature and pressure, there are about 2.55×10^{19} molecules per cm^{-3}. The mixing ratio is a dimensionless quantity, usually expressed in parts per million by volume (ppmv $= 10^{-6}$), parts per billion by volume (ppbv $= 10^{-9}$), or parts per trillion by volume (pptv $= 10^{-12}$).

Photodissociation Absorption of incoming solar radiation, usually radiation of visible wavelengths or shorter, that leads to the dissociation of atmospheric molecules to their constituent molecules, atoms, or radicals. For example, the photodissociation of water vapor leads to the formation of atomic hydrogen (H) and the hydroxyl radical (OH).

Primary or primordial atmosphere Atmosphere resulting from capture of the gaseous material in the primordial solar nebula from which the solar system condensed about 4.6 billion years ago. The atmospheres of Jupiter, Saturn, Uranus, and Neptune are believed to be remnants of the primordial solar nebula and, hence, contain atoms of hydrogen, helium, nitrogen, oxygen, carbon, and so on in the same elemental proportion as the sun. The atmospheres of these planets are composed of molecular hydrogen and helium, with smaller amounts of methane, ammonia, and water vapor, and their photodissociation products.

Primordial solar nebula Interstellar cloud of gas, dust, and ice of a few solar masses, at a temperature of about 10 K, that collapsed under its own gravitational attraction to form the sun, the planets, and the rest of the solar system about 4.6 billion years ago. Compression caused the temperature of the contracting cloud to increase to several thousand degrees, vaporizing all but the most refractory compounds, while conservation of angular momentum flattened the cloud into a disk. The refractory elements in the equatorial plane began to accumulate into large bodies, eventually forming the planets by accretion and coalescence. The bulk of the mass of the primordial solar nebula, composed primarily of hydrogen and helium, formed the sun.

Refractory elements Elements or their compounds that volatilize only at very high temperatures, such as silicon, magnesium, and aluminum. Refractory elements and their compounds formed the terrestrial planets through the processes of accretion and coalescence.

Secondary atmosphere Atmosphere resulting from the outgassing of trapped volatiles, that is, the atmospheres surrounding earth, Venus, and Mars.

Troposphere Lowest region of the earth's atmosphere, which extends from the surface to about 15 km in the tropics and to 10 km at high latitudes. About 80% of the total mass of the atmosphere is found in the troposphere (the rest of the total mass of the atmosphere is found in the stratosphere, which extends to about 50 km, with only a fraction of a percent of the total mass of the atmosphere found in the atmospheric regions above the troposphere and stratosphere: the mesosphere, thermosphere, exosphere, ionosphere, and magnetosphere).

Volatile elements Elements that are either gaseous or form gaseous compounds at relatively low temperatures.

Volatile outgassing Release of volatiles trapped in the solid earth during the planetary formation process. The release of the trapped volatiles led to the formation of the atmosphere and ocean.

GRAVITATIONALLY BOUND to the planets are atmospheres, gaseous envelopes of widely differing masses and chemical compositions. The origin of the atmospheres of the planets is directly related to the origin of the planets some 4.6 billion years ago. Much of our knowledge and understanding of the origin, evolution, structure, composition, and meteorology of planetary atmospheres has resulted from the exploration of the planets and their atmospheres by a series of planetary fly-bys, orbiters, and landers. The atmospheres of the terrestrial planets (earth, Venus, and Mars) most probably resulted from the release of gases originally trapped in the solid planet during the planetary formation process. Water vapor, carbon dioxide, and molecular nitrogen outgassed from the terrestrial planets to form their atmospheres. By contrast, it is generally thought that the very dense hydrogen and helium atmospheres of the outer planets (Jupiter, Saturn, Uranus, and Neptune) are the gaseous remnants of the primordial solar nebula that condensed to form the sun and the planets. Of all of the planets in the solar system, the atmosphere of the earth has probably changed the most over geological time in response to both the geochemical cycling of a geologically active planet and the biochemical cycling of a biologically active planet.

I. FORMATION OF THE PLANETS AND THEIR ATMOSPHERES

The sun, earth, and the other planets condensed out of the primordial solar nebula, an interstellar cloud of gas and dust, some 4.6 billion years ago (orbital information and the physical characteristics of the planets are summarized

TABLE I The Planets: Orbital Information and Physical Characteristics

	Mercury	Venus	Earth	Mars	Jupiter	Saturn	Uranus	Neptune	Pluto
Mean distance from sun (millions of km)	57.9	108.2	149.6	227.9	778.3	1427	2869	4496	5900
Period of revolution	88 days	224.7 days	365.26 days	687 days	11.86 yr	29.46 yr	84.01 yr	164.1	247.7 yr
Rotation period	59 days	−243 days Retrograde	23 hr 56 min 4 sec	24 hr 37 min 23 sec	9 hr 55 min 30 sec	10 hr 39 min 20 sec	17.24 hr	22 hr or less	−6 days 9 hr 18 min Retrograde
Inclination of axis	2°	3°	23°27′	25°12′	3°5′	26°44′	97°55′	28°48′	60°?
Inclination of orbit to ecliptic	7°	3.4°	0°	1.9°	1.3°	2.5°	.8°	1.8°	17.2°
Eccentricity of orbit	0.206	0.007	0.017	0.093	0.048	0.056	0.047	0.009	0.25
Equatorial diameter (km)	4880	12,104	12,756	6787	142,800	120,400	51,800	49,500	3500
Atmosphere (main components)	Virtually none	Carbon dioxide	Nitrogen dioxide Oxygen	Carbon	Hydrogen Helium	Hydrogen Helium	Hydrogen Helium	Hydrogen Helium	Methane (?)
Known satellites	0	0	1	2	16	18	15	8	1
Rings	—	—	—	—	Yes (1)	Yes (8)	Yes (11)	Yes (4)	—

in Table I). The chemical composition of the primordial solar nebula most probably reflected the cosmic abundance of the elements (see Table II). Volatiles, elements that were either gaseous or that formed gaseous compounds at the relatively low temperature of the solar nebula, were the major constituents. The overwhelmingly prevalent volatile element was hydrogen, followed by helium, oxygen, nitrogen, and carbon (see Table II). Considerably less abundant in the solar nebula, but key elements in the formation of the solid planets, were the nonvolatile refractory elements, such as silicon, iron, magnesium, nickel, and aluminum, which formed solid elements and compounds at the relatively low temperature of the solar nebula. The terrestrial planets (Mercury, Venus, earth, and Mars) formed through the processes of coalescence and accretion of the refractory elements and their compounds, beginning with grains the size of dust, to boulder-sized "planetesimals," to planetary-sized bodies. The terrestrial planets may have grown to their full size and mass in as little as 10 million years. Volatiles incorporated in a late-accreting, low-temperature condensate may have formed as a veneer surrounding the newly formed terrestrial planets. The chemical composition of this volatile-rich veneer resembled that of carbonaceous chondritic meteorites, which contain relatively large amounts of water (H_2O) and other volatiles. The collisional impact of the refractory material during the coalescence and accretion phase caused widespread heating in the forming planets. The heating was accompanied by the release of the trapped volatiles through a process termed volatile outgassing. The oxidation state and, hence, the chemical composition of the outgassed volatiles depended on the structure and composition of the solid planet and, in particular, on the presence or absence of free iron in the upper layers of the solid planet. If the terrestrial planets formed as geologically differentiated bodies, i.e., with free iron having already migrated to the core (as a result of the heating and high temperature accompanying planetary accretion), surrounded by an iron-free mantle of silicates, the outgassed volatiles would have been composed of water vapor, carbon dioxide (CO_2), and molecular nitrogen (N_2), not unlike the chemical composition of present-day volcanic emissions. Current theories of planetary formation suggest that the earth, Venus, and Mars formed as geologically differentiated objects. Some volatile outgassing may have also been associated with the impact heating during the final stages of planetary formation. This outgassing would have resulted in an almost instantaneous formation of the atmosphere, coincident with the final stages of planetary formation. As a result of planetary accretion and volatile outgassing, the terrestrial planets are characterized by iron–silicate interiors with atmospheres composed primarily of carbon dioxide (Venus and Mars) or molecular nitrogen (earth), with surface pressures that range from about 1/200 atm (Mars) to about 90 atm (Venus) (the surface pressure of the earth's atmosphere is 1 atmosphere).

Since Mercury does not possess an appreciable atmosphere, it is not discussed in any detail in this article, which concentrates on the chemical composition of planetary atmospheres. Measurements obtained by Mariner 10, which encountered Mercury three times in 1974–1975 after a 1974 Venus fly-by, indicated that the surface pressure of the atmosphere of Mercury is less than a thousandth of a trillionth of the earth's, with helium resulting from

TABLE II Cosmic Abundance of the Elements[a]

Element	Abundance[b]	Element	Abundance[b]
$_{1}$H	2.6×10^{10}	$_{44}$Ru	1.6
$_{2}$He	2.1×10^{9}	$_{45}$Rh	0.33
$_{3}$Li	45	$_{46}$Pd	1.5
$_{4}$Be	0.69	$_{47}$Ag	0.5
$_{5}$B	6.2	$_{48}$Cd	2.12
$_{6}$C	1.35×10^{7}	$_{49}$In	2.217
$_{7}$N	2.44×10^{6}	$_{50}$Sn	4.22
$_{8}$O	2.36×10^{7}	$_{51}$Sb	0.381
$_{9}$F	3630	$_{52}$Te	6.76
$_{10}$Ne	2.36×10^{6}	$_{53}$I	1.41
$_{11}$Na	6.32×10^{4}	$_{54}$Xe	7.10
$_{12}$Mg	1.050×10^{6}	$_{55}$Cs	0.367
$_{13}$Al	8.51×10^{4}	$_{56}$Ba	4.7
$_{14}$Si	1.00×10^{6}	$_{57}$La	0.36
$_{15}$P	1.27×10^{4}	$_{58}$Ce	1.17
$_{16}$S	5.06×10^{5}	$_{59}$Pr	0.17
$_{17}$Cl	1970	$_{60}$Nd	0.77
$_{18}$Ar	2.28×10^{5}	$_{62}$Sm	0.23
$_{19}$K	3240	$_{63}$Eu	0.091
$_{20}$Ca	7.36×10^{4}	$_{64}$Gd	0.34
$_{21}$Sc	33	$_{65}$Tb	0.052
$_{22}$Ti	2300	$_{66}$Dy	0.36
$_{23}$V	900	$_{67}$Ho	0.090
$_{24}$Cr	1.24×10^{4}	$_{68}$Er	0.22
$_{25}$Mn	8800	$_{69}$Tm	0.035
$_{26}$Fe	8.90×10^{5}	$_{70}$Yb	0.21
$_{27}$Co	2300	$_{71}$Lu	0.035
$_{28}$Ni	4.57×10^{4}	$_{72}$Hf	0.16
$_{29}$Cu	919	$_{73}$Ta	0.022
$_{30}$Zn	1500	$_{74}$W	0.16
$_{31}$Ga	45.5	$_{75}$Re	0.055
$_{32}$Ge	126	$_{76}$Os	0.71
$_{33}$As	7.2	$_{77}$Ir	0.43
$_{34}$Se	70.1	$_{78}$Pt	1.13
$_{35}$Br	20.6	$_{79}$Au	0.20
$_{36}$Kr	64.4	$_{80}$Hg	0.75
$_{37}$Rb	5.95	$_{81}$Tl	0.182
$_{38}$Sr	58.4	$_{82}$Pb	2.90
$_{39}$Y	4.6	$_{83}$Bi	0.164
$_{40}$Zr	30	$_{90}$Th	0.034
$_{41}$Nb	1.15	$_{92}$U	0.0234
$_{42}$Mo	2.52		

[a] From Cameron, A. G. W. (1968). *In* "Origin and Distribution of the Elements" (L. H. Ahrens, ed.), Pergamon, New York. Copyright 1968 Pergamon Press.

[b] Abundance normalized to silicon (Si) = 1.00×10^{6}.

radiogenic decay and subsequent outgassing as a possible constituent. Mercury was found to possess an internal magnetic field, similar to but weaker than the earth's. Mariner 10 photographs indicated that the surface of Mercury is very heavily cratered, resembling the highlands on the moon (Fig. 1). A large impact basin (Caloris), about 1300 km in diameter, was discovered. Long scarps of cliffs, apparently produced by crustal compression, were also found.

In direct contrast to the terrestrial planets, the outer planets (Jupiter, Saturn, Uranus, and Neptune) are more massive (15–318 earth masses), larger (4–11 earth radii), and possess multiple satellites and ring systems (see Table I). The atmospheres of the outer planets are very dense and contain thick clouds and haze layers. These atmospheres are composed primarily (85–95% by volume) of molecular hydrogen (H_2) and helium (He) (5–15%) with smaller amounts of compounds of carbon, nitrogen, and oxygen, primarily present in the form of saturated hydrides [methane (CH_4), ammonia (NH_3), and water vapor] at approximately the solar ratio of carbon, nitrogen, and oxygen. The composition of the atmospheres of the outer planets suggests that they are captured remnants of the primordial solar nebula that condensed to form the solar system, as opposed to having formed as a result of the outgassing of volatiles trapped in the interior, as did the atmospheres of the terrestrial planets. It has been suggested that a thick atmosphere of molecular hydrogen and helium, the overwhelming constituents of the primordial solar nebula, may have surrounded the terrestrial planets very early in their history (during the final stages of planetary accretion). However, such a primordial solar nebula remnant atmosphere surrounding the terrestrial planets would have dissipiated very quickly, due to the low mass of these planets and, hence, their weak gravitational attraction, coupled with the rapid gravitational escape of hydrogen and helium, the two lightest gases, from the "warm" terrestrial planets. Therefore, an early atmosphere composed of hydrogen and helium surrounding the terrestrial planets would have been extremely short lived, if it ever existed at all. The large masses of the outer planets and their great distances from the sun (and colder temperatures) have enabled them to gravitationally retain their primordial solar nebula remnant atmospheres. The colder temperatures resulted in a "freezing out" or condensation of several atmospheric gases, such as water vapor, ammonia, and methane forming cloud and haze layers in the atmospheres of the outer planets.

The most distant planet in the solar system, Pluto, has a very eccentric orbit, which at times brings it closer to the sun than Neptune's orbit. By virtue of its great orbital eccentricity and small mass, it is suspected that Pluto may have originally been a satellite of another planet. Methane has been detected on Pluto. Very little is known about Pluto. Most of what we know about Pluto is summarized in Table I. Of the numerous smaller bodies in the solar system, including satellites and asteroids, only Saturn's satellite, Titan, has an appreciable atmosphere (surface

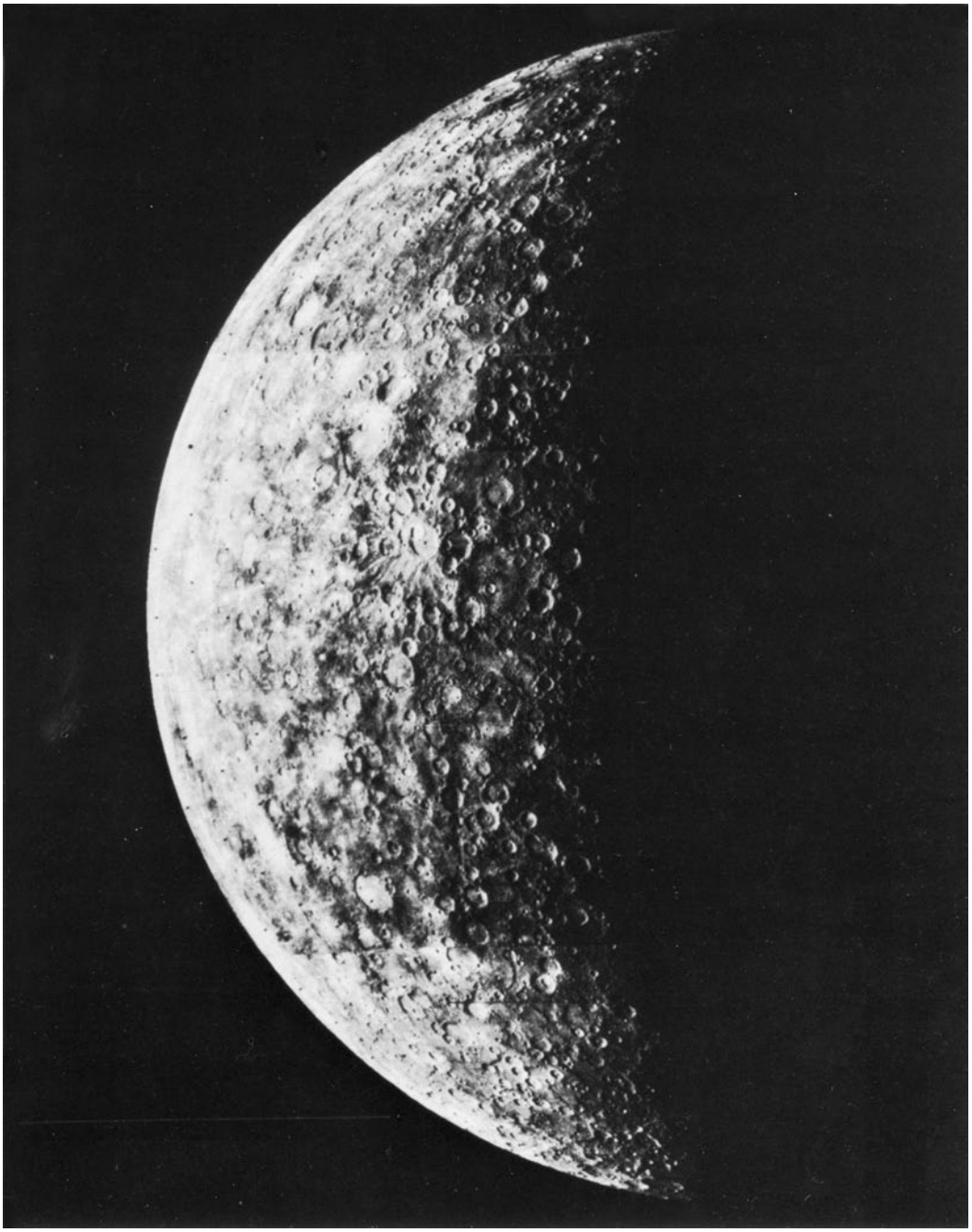


FIGURE 1 Mosaic of Mariner 10 photographs of Mercury. With no appreciable atmosphere, we can see right down to the cratered surface of Mercury, which is similar to the cratered highlands on the moon. The largest craters in the photograph are about 200 km in diameter.

pressure about 1.5 atm), composed of molecular nitrogen and a small amount of methane.

II. EARTH

The atmospheres of the earth, Venus, and Mars resulted from the outgassing of volatiles originally trapped in their interiors. The chemical composition of the outgassed volatiles was not unlike that of present-day volcanic emissions: water vapor = 79.31% by volume; carbon dioxide = 11.61%; sulfur dioxide (SO_2) = 6.48%; and molecular nitrogen = 1.29%. On earth, the bulk of the outgassed water vapor condensed out of the atmosphere, forming the earth's vast ocean. Only small amounts of water vapor remained in the atmosphere, with almost all

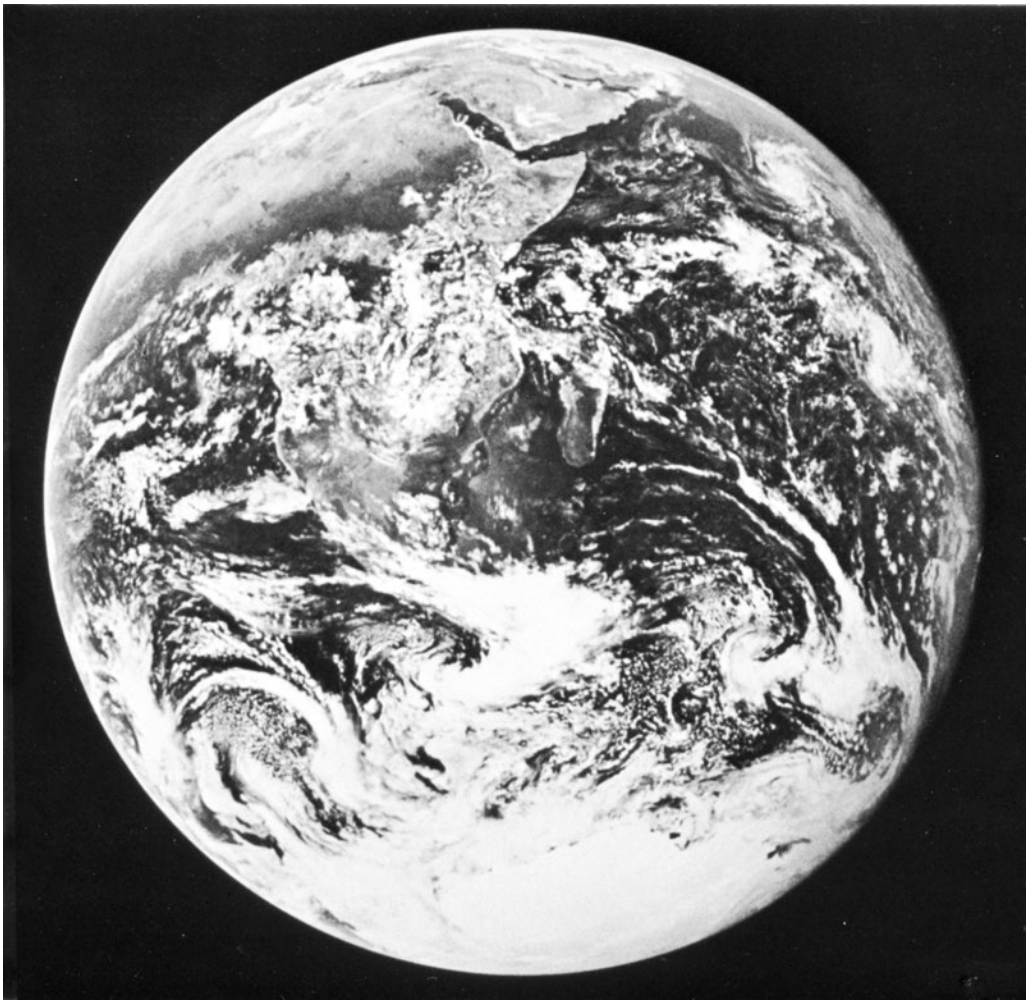


FIGURE 2 Planet earth as photographed by Apollo 17 astronauts on their journey to the moon. Scattered clouds, which cover only about 50% of the earth at any given time, permit viewing the surface of the earth. Earth is a unique planet in many respects, including the presence of life, the existence of liquid water on the surface, and the presence of large amounts of oxygen in the atmosphere.

of it confined to the troposphere. Some atmospheric water is in the condensed state, found in the form of cloud droplets. Water clouds cover about 50% of the earth's surface at any given time and are a regular feature of the atmosphere (Fig. 2). Near the ground, the water vapor concentration is variable, ranging from a fraction of a percent to a maximum of several percent by volume. Once the ocean formed, outgassed carbon dioxide, the second most abundant volatile, which is very water soluble, dissolved into the ocean. Once dissolved in the ocean, carbon dioxide chemically reacted with ions of calcium and magnesium, also in the ocean, and precipitated out in the form of sedimentary carbonate rocks such as calcite ($CaCO_3$), and dolomite [$CaMg(CO_3)_2$].

The concentration of carbon dioxide in the atmosphere is about 0.036% by volume, which is equivalent to 360 parts per million by volume (ppmv). It has been estimated that the preindustrial (ca. 1860) level of atmospheric carbon dioxide was about 280 ppmv, with the increase to the present level attributable to the burning of fossil fuels, notably coal. For each carbon dioxide molecule in the present-day atmosphere, there are approximately 50 carbon dioxide molecules physically dissolved in the ocean and almost 30,000 carbon dioxide molecules incorporated in sedimentary carbonate rocks. All of the carbon dioxide presently incorporated in carbonate rocks originally outgassed from the interior and was at one time in the atmosphere. Hence, the early atmosphere may have contained 100 to 1000 times or more carbon dioxide than it presently contains. Sulfur dioxide, the third most abundant gas in volcanic emissions, is chemically unstable in the atmosphere. Sulfur dioxide is rapidly chemically transformed

to water-soluble sulfuric acid (H_2SO_4), which readily rains out of the atmosphere. Hence, the atmospheric lifetime of sulfur dioxide is very short.

Molecular nitrogen is the fourth most abundant gas in volcanic emissions. Nitrogen does not condense out of the atmosphere (as does water vapor), is not water soluble (as is carbon dioxide), and is not chemically active (as is sulfur dioxide). As a result, the bulk of the outgassed molecular nitrogen accumulated in the atmosphere, and over geological time became the most abundant atmospheric constituent (about 78% by volume). Molecular oxygen (O_2), produced as a by-product of photosynthetic activity, built up in the atmosphere to become the second most abundant species (about 21% by volume). Argon (isotope 40), the third most abundant atmospheric species (about 1% by volume), is a chemically inert gas resulting from the radiogenic decay of potassium (isotope 40) in the crust. Hence, the bulk chemical composition of the earth's atmosphere can be explained in terms of volatile outgassing and the ultimate sinks of the outgassed volatiles, including condensation/precipitation, dissolution and carbonate formation in the ocean, biogenic activity, and radiogenic decay. These and other processes, including photochemical and chemical reactions, have resulted in a myriad of other trace atmospheric gases in the atmosphere, which are listed in Table III.

III. VENUS

Venus has been described as the earth's twin because of its similar mass (0.81 earth masses), radius (0.95 earth radii), mean density (95% that of earth), and gravity (90% that of earth) (see Table I). However, in terms of atmospheric

TABLE III Composition of the Earth's Atmosphere[a]

	Surface concentration[b]	Source
Major and minor Gases		
Nitrogen (N_2)	78.08%	Volcanic, biogenic
Oxygen (O_2)	20.95%	Biogenic
Argon (Ar)	0.93%	Radiogenic
Water vapor (H_2O)	Variable, up to 4%	Volcanic, evaporation
Carbon dioxide (CO_2)	0.036%	Volcanic, biogenic, anthropogenic
Trace gases		
Oxygen species		
Ozone (O_3)	10–100 ppbv	Photochemical
Atomic oxygen (O) (ground state)	10^3 cm^{-3}	Photochemical
Atomic oxygen [O(^{1}D)] (excited state)	10^{-2} cm^{-3}	Photochemical
Hydrogen species		
Hydrogen (H_2)	0.5 ppmv	Photochemical, biogenic
Hydrogen peroxide (H_2O_2)	10^9 cm^{-3}	Photochemical
Hydroperoxyl radical (HO_2)	10^8 cm^{-3}	Photochemical
Hydroxyl radical (OH)	10^6 cm^{-3}	Photochemical
Atomic hydrogen (H)	1 cm^{-3}	Photochemical
Nitrogen species		
Nitrous oxide (N_2O)	330 ppbv	Biogenic, anthropogenic
Ammonia (NH_3)	0.1–1 ppbv	Biogenic, anthropogenic
Nitric acid (HNO_3)	50–1000 pptv	Photochemical
Hydrogen cyanide (HCN)	~200 pptv	Anthropogenic(?)
Nitrogen dioxide (NO_2)	10–300 pptv	Photochemical
Nitric oxide (NO)	5–100 pptv	Anthropogenic, biogenic, lightning, photochemical
Nitrogen trioxide (NO_3)	100 pptv	Photochemical
Peroxyacetylnitrate ($CH_3CO_3NO_2$)	50 pptv	Photochemical
Dinitrogen pentoxide (N_2O_5)	1 pptv	Photochemical
Pernitric acid (HO_2NO_2)	0.5 pptv	Photochemical
Nitrous acid (HNO_3)	0.1 pptv	Photochemical

continues

TABLE III (*Continued*)

	Surface concentration[b]	Source
Nitrogen aerosols		
Ammonium nitrate (NH_4NO_3)	~100 pptv	Photochemical
Ammonium chloride (NH_4Cl)	~0.1 pptv	Photochemical
Ammonium sulfate [$(NH_4)_2SO_4$]	~0.1 pptv(?)	Photochemical
Carbon species		
Methane (CH_4)	1.7 ppmv	Biogenic, anthropogenic
Carbon monoxide (CO)	70–200 ppbv (N hemis.) 40–60 ppbv (S hemis.)	Anthropogenic, biogenic, photochemical
Formaldehyde (H_2CO)	0.1 ppbv	Photochemical
Methylhydroperoxide (CH_3OOH)	10^{11} cm^{-3}	Photochemical
Methylperoxyl radical (CH_3O_2)	10^{8} cm^{-3}	Photochemical
Methyl radical (CH_3)	10^{-1} cm^{-3}	Photochemical
Sulfur species		
Carbonyl sulfide (COS)	0.5 ppbv	Volcanic, anthropogenic
Dimethyl sulfide[$(CH_3)_2S$]	0.4 ppbv	Biogenic
Hydrogen sulfide (H_2S)	0.2 ppbv	Biogenic, anthropogenic
Sulfur dioxide (SO_2)	0.2 ppbv	Volcanic, anthropogenic, photochemical
Dimethyl disulfide [$(CH_3)_2S_2$]	100 pptv	Biogenic
Carbon disulfide (CS_2)	50 pptv	Volcanic, anthropogenic
Sulfuric acid (H_2SO_4)	20 pptv	Photochemical
Sulfurous acid (H_2SO_3)	20 pptv	Photochemical
Sulfoxyl radical (SO)	10^{3} cm^{-3}	Photochemical
Thiohydroxyl radical (HS)	1 cm^{-3}	Photochemical
Sulfur trioxide (SO_3)	10^{-2} cm^{-3}	Photochemical
Halogen species		
Hydrogen chloride (HCl)	1 ppbv	Sea salt, volcanic
Methyl chloride (CH_3Cl)	0.5 ppbv	Biogenic, anthropogenic
Methyl bromide (CH_3Br)	10 pptv	Biogenic, anthropogenic
Methyl iodide (CH_3I)	1 pptv	Biogenic, anthropogenic
Noble gases (chemically inert)		
Neon (Ne)	18 ppmv	Volcanic
Helium (He)	5.2 ppmv	Radiogenic
Krypton (Kr)	1 ppmv	Radiogenic
Xenon (Xe)	90 ppbv	Radiogenic

[a] From Levine, J. S. (1985). *In* "The Photochemistry of Atmospheres: Earth, the Other Planets, and Comets" (J. S. Levine, ed.), Academic, Orlando, FL.

[b] Species concentrations are given in percentage by volume, in terms of surface mixing ratio, parts per million by volume (ppmv $\equiv 10^{-6}$), parts per billion by volume (ppbv $\equiv 10^{-9}$), parts per trillion by volume (pptv $\equiv 10^{-12}$), or in terms of surface number density (cm^{-3}). The species mixing ratio is defined as the ratio of the number density of the species to the total atmospheric number density (2.55×10^{19} molec cm^{-3}). There is some uncertainty in the concentrations of species at the ppbv level or less. The species concentrations given in molec cm^{-3} are generally based on photochemical calculations, and species concentrations in mixing ratios are generally based on measurements.

structure and chemical composition, Venus is anything but a twin of earth. The mean planetary surface temperature of Venus is about 750 K, compared with about 300 K for earth; the surface pressure on Venus is about 90 atm, compared with 1 atm for earth; carbon dioxide at 96% by volume is the overwhelming constituent in the atmosphere of Venus, while it is only a trace constituent in the earth's atmosphere (0.034% by volume). In addition, Venus does not have an ocean or a biosphere and is completely covered by thick clouds, probably composed of sulfuric acid (Fig. 3). Hence, the atmosphere is inhospitable and very unlike the earth's atmosphere.

Much of our information on the structure and composition of the atmosphere of Venus has been obtained through

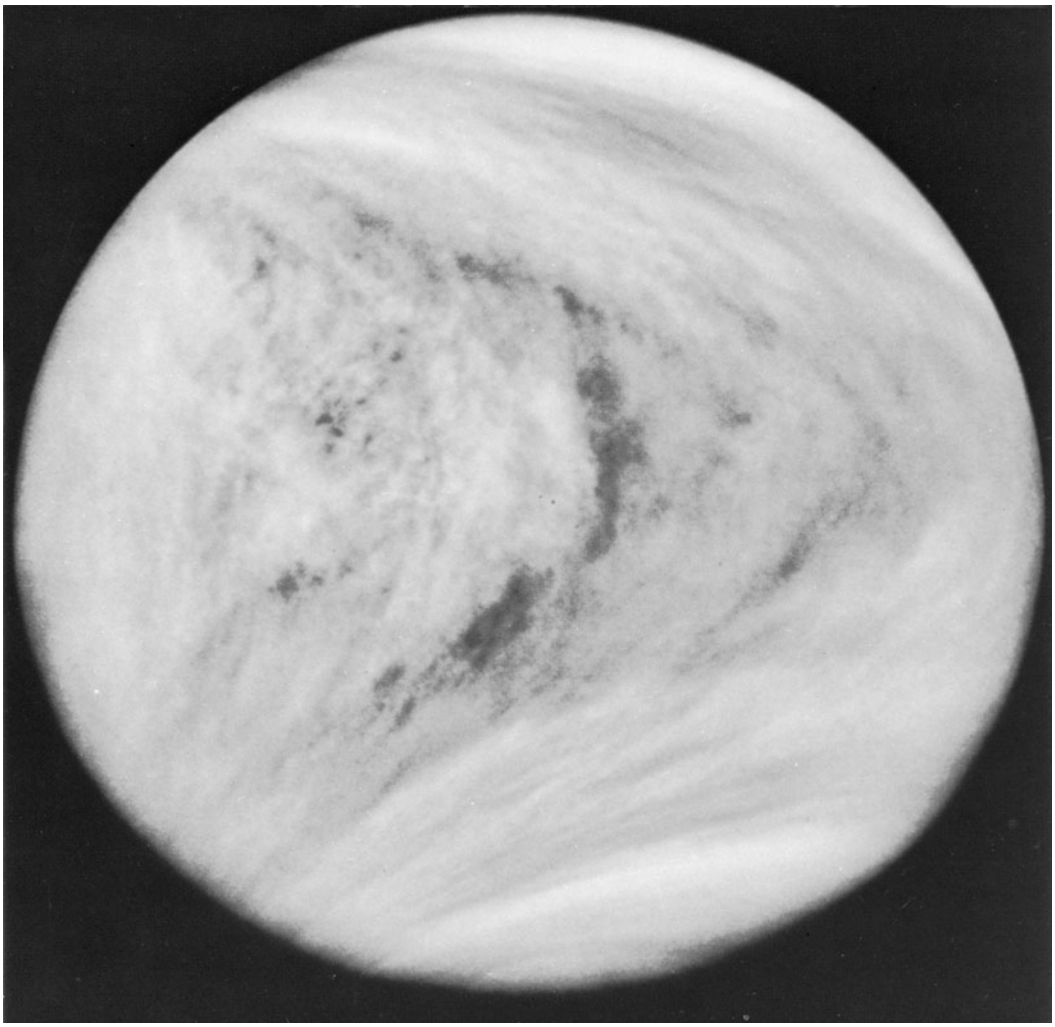


FIGURE 3 Venus as photographed by the Pioneer Venus Orbiter. The thick, cloud-covered atmosphere continually and completely hides the surface of Venus, which at very high temperature and atmospheric pressure is not a hospitable environment.

a series of U.S. and U.S.S.R. Venus fly-bys, orbiters, and landers, which are summarized in Table IV.

The clouds on Venus are thick and contain no holes, hence, we have never directly observed the surface of Venus from earth. These clouds resemble a stratified low-density haze extending from about 45 to about 65 km. The total extinction optical depth of the clouds in visible light is about 29. The extinction of visible light is due almost totally to scattering. The lower clouds are found between 45 and 50 km; the middle clouds from 50 to 55 km; and the upper clouds from 55 to 65 km. The tops of the upper clouds, which are the ones visible from earth, appear to be composed of concentrated sulfuric acid droplets (see Fig. 3).

As already noted, carbon dioxide at 96% by volume, is the overwhelming constituent of the atmosphere of Venus. The next most abundant atmospheric gas is molecular nitrogen at 4% by volume. The relative proportion by volume of carbon dioxide and molecular nitrogen in the atmospheres of Venus and Mars is almost identical. The chemical composition of the atmosphere of Venus is summarized in Table V. At the surface of Venus, the partial pressure of carbon dioxide is about 90 bar, molecular nitrogen is about 3.2 bar, and water vapor is only about 0.01 bar (more about water on Venus later). For comparison, if the earth were heated to the surface temperature of Venus (about 750 K), we would have a massive atmosphere composed of water vapor at a surface partial pressure of about 300 bar (resulting from the evaporation of the ocean), a carbon dioxide partial pressure of about 55 bar (resulting from the thermal composition of crustal carbonates), and a molecular nitrogen pressure of about 1 to 3 bar (resulting

TABLE IV Missions to Venus

Name	Designator (U.S.S.R.)	Launch date	Mission remarks
Venera 1	61-Gamma 1	Feb. 12, 1961	Passed Venus at 100,000 km May 19–21, 1961; contact lost Feb. 27. 1961.
Mariner II	—	Aug. 27, 1962	Planetary exploration: First successful interplanetary probe. Found no magnetic field; high surface temperatures of approximately 800°F. Passed Venus Dec. 14, 1962 at 21,600 miles, 109 days after launch.
Zond 1	64-16D	April 2, 1964	Passed Venus at 100,000 km July 19, 1964; communications failed after May 14, 1964.
Venera 2	65-91A	Nov. 12, 1965	Passed Venus at 24,000 km, Feb. 27, 1966; communications failed.
Venera 3	65-92A	Nov. 16, 1965	Struck Venus March 1, 1966; communications failed earlier.
Venera 4	67-58A	June 12, 1967	Probed atmosphere.
Venera 5	69-1A	Jan. 5, 1969	Entered Venus atmosphere May 16, 1969.
Mariner V	—	June 14, 1967	Planetary exploration: All science and engineering subsystems nominal through encounter with Venus; data indicate Venus has a moonlike effect on solar plasma and strong H_2 corona comparable to Earth's, 72 to 87% CO_2 atmosphere with balance probably nitrogen and O_2. Closest approach, 3,900 km on Oct. 19, 1967.
Venera 6	69-2A	Jan. 10, 1969	Entered Venus atmosphere May 17, 1969.
Venera 7	70-60A	Aug. 17, 1970	Soft landed on Venus; signal from surface.
Venera 8	72-21A	Mar. 27, 1972	Soft landed on Venus; sent data from surface.
Mariner 10 Venus/Mercury	—	Nov. 3, 1973	Conducted exploratory investigations of planet Mercury during three fly-bys by obtaining measurements of its environment, atmosphere, surface, and body characteristics and conducted similar investigations of Venus. Mariner 10 encountered Venus on Feb. 5, 1974 and Mercury on Mar. 29 and Sept. 21, 1974 and Mar. 16. 1975. Resolution of the photographs was 100 m, 7000 times greater than that achieved by earth-based telescopes.
Venera 9-orbiter	75-50A	June 8, 1975	Orbited Venus Oct. 22, 1975. Orbiter and lander launched from single D-class vehicle (Proton), 4650 kg thrust.
Venera 9-lander	75-50D	June, 1975	Soft landed; returned picture.
Venera 10-orbiter	75-54A	June 14, 1975	Orbited Venus Oct. 25, 1975. Orbiter and lander launched from single D-class vehicle (Proton), 4659 kg thrust.
Venera 10-lander	75-54D	June 14, 1975	Soft landed; returned picture.
Pioneer 12	—	May 20, 1978	Orbiter launched in May studied interaction of atmosphere and solar wind and made radar and gravity maps of the planet. The multiprobe spacecraft launched in August returned information on Venus's wind and circulation patterns as well as atmospheric composition, temperature and pressure readings. Pioneer 12 entered Venus orbit Dec. 4, 1978; Pioneer 13 encountered Venus Dec. 9, 1978.
Pioneer 13 Pioneer Venus		Aug. 8, 1978	
Venera 11-orbiter	78-84A	Sept. 9, 1978	Passed Venus at 35,000 km Dec. 25, 1978; served as relay station. Orbiter and lander launched from single D-class vehicle (Proton), 4650 kg thrust.
Venera 11-lander	78-84E	Sept. 9, 1978	Soft landed on Venus.
Venera 12-orbiter	78-86A	Sept. 14, 1978	Passed Venus at 35,000 km Dec. 21, 1978; served as relay station. Orbiter and lander launched from single D-class vehicle (Proton), 4650 kg thrust.
Venera 12-lander	78-86E	Sept. 14, 1978	Soft landed on Venus.
Venera 13-orbiter	1981-106A	Oct. 30, 1981	Both orbiter and lander launched from single D-class vehicle (Proton), 4650 kg thrust.
Venera 13-lander	None	Oct. 30, 1981	Soft landed on Venus Mar. 3, 1982; returned color picture.
Venera 14-orbiter	1981-110A	Nov. 4, 1981	Both orbiter and lander launched from single D-class vehicle (Proton), 4650 kg thrust.
Venera 14-lander	None	Nov. 4, 1981	Soft landed on Venus Mar. 5, 1982; returned color picture.
Magellan		May 4, 1989	Magellan was the first planetary spacecraft to be launched from a space shuttle. It arrived at Venus on August 10, 1990, and orbited Venus in a near-polar, elliptical orbit, with a minimum altitude of 243 km. It made detailed maps of 98% of the planet's surface using synthetic aperture radar.

from the present atmosphere plus the outgassing of crustal nitrogen).

A major puzzle concerning the chemical composition of the atmosphere of Venus (as well as the atmosphere of Mars) is the stability of carbon dioxide and the very low atmospheric concentrations of carbon monoxide (CO) and oxygen [atomic (O) and molecular], which are the photodissociation products of carbon dioxide. In the daytime

TABLE V Composition of the Atmosphere of Venus[a]

Gas	Volume mixing ratio: Troposphere (below clouds)	Volume mixing ratio: Stratosphere (above clouds)
CO_2	9.6×10^{-1}	9.6×10^{-1}
N_2	4×10^{-2}	4×10^{-2}
H_2O	10^{-4}–10^{-3}	10^{-6}–10^{-5}
CO	$(2–3) \times 10^{-5}$	5×10^{-5}–10^{-3}
HCl	$<10^{-5}$	10^{-6}
HF	?	10^{-8}
SO_2	1.5×10^{-4}	5×10^{-8}–8×10^{-7}
S_3	$\sim 10^{-10}$[b]	?
H_2S	$(1–3) \times 10^{-6}$[b]	?
COS	$<2 \times 10^{-6}$	?
O_2	$(2–4) \times 10^{-5}$[b]	$<10^{-6}$[b]
H_2	?	2×10^{-5}[b]
^{4}He	10^{-5}	10^{-5}
^{20,22}Ne	$(5–13) \times 10^{-6}$	$(5–13) \times 10^{-6}$
^{30,38,40}Ar	$(5–12) \times 10^{-5}$	$(5–12) \times 10^{-5}$
^{84}Kr	$<2 \times 10^{-8}$–4×10^{-7}	$<2 \times 10^{-8}$–4×10^{-7}

[a] From Lewis, J. S., and Prinn, R. G. (1986). "Planets and Their Atmospheres," Academic, New York. Copyright 1984 Academic Press.

[b] Single experiment; corroboration required.

upper atmosphere (above 100 km), carbon dioxide is readily photodissociated with a photochemical atmospheric lifetime of only about one week. The recombination of carbon monoxide and atomic oxygen in the presence of a third body to reform carbon dioxide is efficient only at the higher atmospheric pressures occurring at and below 100 km. However, at these lower altitudes, atomic oxygen recombines with itself in the presence of a third body to form molecular oxygen considerably faster than the three-body reaction that leads to the recombination of carbon dioxide. Thus, essentially all of the photolyzed carbon dioxide produces carbon monoxide and molecular oxygen. Yet, the observed upper-limit atmospheric concentration of molecular oxygen above the cloud tops could be produced in only about one day, and the observed abundance of carbon monoxide could be produced in only about three months. Photodissociation could easily convert the entire concentration of carbon dioxide in the atmosphere to carbon monoxide and molecular oxygen in only about 4 million years, geologically a short time period.

This dilemma also applies to carbon dioxide on Mars. Considerable research has centered around the recombination of carbon monoxide and molecular oxygen back to carbon dioxide. It became apparent that the only way to maintain low carbon monoxide and oxygen concentrations and high carbon dioxide concentrations in the 100- to 150-km region is by the rapid downward transport of carbon monoxide and oxygen, balanced by the upward transport of carbon dioxide. It is believed that carbon dioxide is reformed from carbon monoxide and oxygen at an altitude of about 70 km through various chemical reactions and catalytic cycles involving chemically active compounds of hydrogen and chlorine.

If Venus and the earth contained comparable levels of volatiles and outgassed them at comparable rates, then Venus must have somehow lost about 300 bar of water vapor. This may have been accomplished by a runaway greenhouse. In the runaway greenhouse on Venus, outgassed water vapor and carbon dioxide entered the atmosphere, contributing to steadily increasing atmospheric opacity and thus to increasing surface and atmospheric temperatures via the greenhouse effect. On earth, water vapor condensed out of the atmosphere forming the ocean, and the oceans then removed atmospheric carbon dioxide via dissolution and subsequent incorporation into carbonates. The greater proximity of Venus to the sun and its higher initial surface temperature appear to be the simple explanation for the divergent fates of water vapor and carbon dioxide on Venus and earth. In the runaway greenhouse scenario, the photodissociation of massive amounts of outgassed water vapor in the atmosphere of Venus would have led to the production of large amounts of hydrogen and oxygen. Hydrogen could have gravitationally escaped from Venus, and oxygen could have reacted with crustal material. The runaway greenhouse and the accompanying high surface and atmospheric temperatures, too hot for the condensation of outgassed water on Venus, would explain the present water vapor-deficient and carbon dioxide-rich atmosphere of Venus. An alternative suggestion is that Venus may have originally accreted without the levels of water that the earth contained, resulting in a much drier Venus.

To a large extent, the earlier Venus missions concentrated on the atmosphere of Venus. The Magellan mission concentrated on the surface of Venus, which cannot be observed from Earth due to the thick and extensive clouds in the atmosphere of Venus. The primary objective of the Magellan mission was to map the surface of Venus using synthetic aperture radar. The surface of Venus is covered by about 20% lowland plains, 70% rolling uplands, and 10% highlands. The surface has been shaped by volcanism, impacts, and deformation of the crust. Volcanoes have left their mark on about 85% of the surface of Venus, with lava plains, lava domes, large shield volcanoes, and extremely long lava channels. More than 100,000 small shield volcanoes dot the surface along with hundreds of larger volcanoes. Giant calderas, more than 100 km in diameter, were found. Calderas are basin-like depressions

in the surface that form after the collapse of the center of a volcano. In the north, an elevated region, Ishtar Terra, is a lava-filled basin larger than the continental United States. Near the equator, the Aphrodite Terra highlands, more than half the size of Africa, extend for almost 10,000 km. Volcanic flows have produced long, sinuous channels extending for hundreds of miles. The rest of Venus is covered with ranges of deformed mountains. Maxwell Montes, a mountain taller than Mount Everest, sits at one end of Ishtar Terra. No direct evidence of active volcanoes has been found, although large variations in atmospheric sulfur dioxide suggest that volcanoes may indeed be active. Impact cratering is affected by the thick atmosphere: craters smaller than 1.5–2 km across do not exist on Venus, largely because small meteors burn up in the dense atmosphere before they can reach the surface.

Magellan also made detailed gravity maps of Venus, finding that the gravity field is highly correlated with surface topography.

IV. MARS

The atmosphere of Mars is very thin (mean surface pressure only about 6.36 mbar), cold (mean surface temperature about 220 K, with the temperature varying from about 290 K in the southern summer to about 150 K in the polar winter), and cloud-free, making the surface of Mars readily visible from the earth (Fig. 4). Much of our information on the structure and composition of the atmosphere of Mars has been obtained through a series of fly-bys, orbiters, and landers, which are summarized in Table VI. As already noted, the composition of the atmosphere of Mars is comparable to that of Venus. Carbon dioxide is the overwhelming constituent (95.3% by volume), with smaller amounts of molecular nitrogen (2.7%) and argon (1.6%), and trace amounts of molecular oxygen (0.13%) and carbon monoxide (0.08%), resulting from the photodissociation of carbon dioxide (the composition of the atmosphere of Mars is summarized in Table VII). Water vapor and ozone (O_3) are also present, although their abundances vary with season and latitude. The annual sublimation and precipitation of carbon dioxide out of and into the polar cap produce a planet-wide pressure change of 2.4 mbar, or 37% of the mean atmospheric pressure of 6.36 mbar.

The amount and location of water vapor in the atmosphere of Mars are controlled by the temperature of the surface and the atmosphere. The northern polar cap is a source of water vapor during the northern summer. The surface of Mars is also a source of water vapor, depending on the location and season. The total amount of water vapor in the atmosphere varies seasonally between the equivalent of 1 and 2 km^3 of liquid water, with the maximum occurring in the northern summer and the minimum in the northern winter. Ozone is also a highly variable constituent of the atmosphere of Mars. Ozone is present only when the atmosphere is cold and dry.

There is evidence to suggest that significant quantities of outgassed carbon dioxide and water vapor may reside on the surface and in the subsurface of Mars. In addition to the polar caps, which contain large concentrations of frozen carbon dioxide and, in the case of the northern polar cap, of frozen water, considerable quantities of these gases may be physically adsorbed to the surface and subsurface material. It has been estimated that if the equilibrium temperature of the winter polar cap would increase from its present value of about 150 K to 160 K, sublimation of frozen carbon dioxide would increase the atmospheric pressure to more than 50 mbar. This in turn would cause more water vapor to leave the polar cap and enter the atmosphere. Mariner and Viking photographs indicate the existence of channels widely distributed over the Martian surface. These photographs show runoff channels, tributary networks, and streamlined islands, all very suggestive of widespread fluid erosion. Yet, there is no evidence for the existence of liquid water on the surface of Mars today. In addition, a significant quantity of water vapor may have escaped from Mars in the form of hydrogen and oxygen atoms, resulting from the photolysis of water vapor in the atmosphere of Mars. If the present gravitational escape rate of atoms of hydrogen and oxygen has been operating over the history of Mars, then an amount of liquid water covering the entire planet about 2.5 m high may have escaped from Mars. Viking measurements of argon and neon in the atmosphere of Mars suggest that Mars may have formed with a lower volatile content than either earth or Venus. This is consistent with ideas concerning the capture and incorporation of volatiles in accreting material and how volatile incorporation varies with temperature, which is a function of the distance of the accreting terrestrial planets from the sun.

Unlike the very thick atmosphere of Venus, where the photolysis of carbon dioxide occurs only in the upper atmosphere (above 100 km), on Mars the photodissociation of carbon dioxide occurs throughout the entire atmosphere, right down to the surface. For comparison, the 6.36 mbar surface pressure of the atmosphere of Mars corresponds to an atmospheric pressure at an altitude of about 33 km in the earth's atmosphere. On Mars, carbon dioxide is reformed from its photodissociation products, carbon monoxide and oxygen, by reactions involving atomic hydrogen (H) and the oxides of hydrogen.

Viking photographs indicate that the surface rocks on Mars resemble basalt lava (see Fig. 5). The red color of the surface is probably due to oxidized iron. The soil

FIGURE 4 Mars as photographed by the Viking 2 orbiter. The thin, cloud-free atmosphere of Mars permits direct observation of the Martian surface from space.

TABLE VI Missions to Mars

Name	Designator (U.S.S.R.)	Launch date	Mission remarks
Mars 1	62-Beta Nu 3	Nov. 1, 1962	Passed Mars June 19, 1963 at 193,000 km; communications failed March 21, 1963.
Mariner IV	—	Nov. 28, 1964	Planetary and interplanetary exploration: Encounter occurred July 14, 1965 with closest approach 6100 miles. Twenty-two pictures taken.
Zond 2	64-78C	Nov. 30, 1964	Passed Mars at 1500 km Aug. 6, 1965; communications failed earlier.
Mariner VI	—	Feb. 25, 1969	Planetary exploration: Mid-course correction successfully executed to achieve a Mars flyby within 3330 km on July 31, 1969. Designed to perform investigations of atmospheric structures and compositions and to return TV photos of surface topography.
Mariner VII	—	Mar. 27, 1969	Planetary exploration: Spacecraft identical to Mariner VI. Mid-course correction successful for 3518 km flyby on Aug. 5, 1969.
Kosmos 419	71-42A	May 10, 1971	Failed to separate.
Mars 2-orbiter	71-45A	May 19, 1971	Orbited Mars Nov. 27, 1971. Mars 2 orbiter and lander launched from single D-class vehicle (Proton), 4650 kg thrust.
Mars-lander	71-45E	May 19, 1971	Landed 47°E.
Mars 3-orbiter	71-49A	May 28, 1971	Orbited Mars Dec. 2, 1971. Mars 3 orbiter and lander launched from single D-class vehicle (Proton), 4650 kg thrust.
Mars 3-lander	71-49F	May 28, 1971	Landed 45°S, 158°W.
Mariner IX	—	May 30, 1971	Entered Mars orbit on Nov. 13, 1971. Spacecraft responded to 38,000 commands and transmitted 6900 pictures of the Martian surface. All scientific instruments operated successfully. Mission terminated on Oct. 27, 1972.
Mars 4	73-47A	July 21, 1973	Passed Mars at 2200 km Feb. 10, 1974, but failed to enter Mars' orbit as planned.
Mars 5	73-49A	July 25, 1973	Orbited Mars Feb. 2, 1974 to gather Mars data and to serve as relay station.
Mars 6-orbiter	73-52A	Aug. 5, 1973	Mars 6 orbiter and lander launched from single D-class vehicle (Proton), 4650 kg thrust.
Mars 6-lander	73-52E	Aug. 5, 1973	Soft landed at 24°S, 25°W; returned atmospheric data during descent.
Mars 7-orbiter	73-53A	Aug. 9, 1973	Mars 7 orbiter and lander launched from single D-class vehicle (Proton), 4650 kg thrust.
Mars 7-lander	73-53E	Aug. 9, 1973	Missed Mars by 1300 km (aimed at 50°S, 28°W).
Viking 1 Lander and orbiter		Aug. 20, 1975	Scientific investigation of Mars. United States' first attempt to soft land a spacecraft on another planet. Successfully soft landed on July 20, 1976. First in situ analysis of surface material on another planet.
Viking 2 Lander and orbiter		Sept. 9, 1975	Scientific investigation of Mars. United States' second attempt to soft land on Mars. Successfully soft landed on Sept. 3, 1976 and returned scientific data. Orbiter from both missions returned over 40,000 high resolution photographs showing surface details as small as 10 m in diameter. Orbiter also collected gravity field data, monitored atmospheric water levels, thermally mapped selected surface sites.
Mars Observer		Sept. 25, 1992	Designed to study the atmosphere, surface, interior, and magnetic field from orbit. Communication with the spacecraft was lost on Aug. 22, 1993, before going into orbit around Mars.
Mars Pathfinder		Dec. 2, 1996	A microrover, named Sojourner, was encased in a self-righting tetrahedral lander, which, in turn, was encapsulated in an aeroshell designed to withstand planetary entry. The Mars Pathfinder arrived at the surface of Mars on July 4, 1997, and characterized surface features and analyzed the composition of rocks and soil at the landing site.
Mars Global Surveyor		Nov. 7, 1996	MGS arrived in orbit around Mars on Sept. 11, 1997, and mapped the entire surface at high resolution and gathered data on surface morphology, topography, gravity, weather and climate, surface and atmospheric composition, and the magnetic field.
Nozomi (Japanese for Hope)		July 4, 1998	Originally called Planet B and renamed Nozomi after launch, this Japanese spacecraft is scheduled to orbit Mars in December 2003 and investigate the atmosphere and ionosphere of Mars. Science payload includes a neutral mass spectrometer.
Mars Climate Orbiter		Dec. 11, 1998	On Sept. 23, 1999, as it prepared to enter orbit around Mars, the spacecraft was targeted too close to the surface, and either burned up in the atmosphere or continued past the planet into space.
Mars Polar Lander		Jan. 3, 1999	The first-ever landing in the polar regions of Mars, near the southern polar cap was lost on Dec. 3, 1999, during it entry into Mars.
Deep Space 2		Jan. 3, 1999	Deep Space 2 or the Mars Microphone Mission, hitched a ride to Mars Polar Lander, Deep Space 2 aboard the Mars Polar Lander. Like the Mars was lost on Dec. 3, 1999, during Mars entry.

TABLE VII Composition of the Atmosphere of Mars[a]

Species	Abundance (mole fraction)
CO_2	0.953
N_2	0.027
^{40}Ar	0.016
O_2	0.13%
CO	0.08%
	0.27%
H_2O	0.03%)[b]
Ne	2.5 ppm
^{36}Ar	0.5 ppm
Kr	0.3 ppm
Xe	0.08 ppm
O_3	(0.03 ppm)[b]
	(0.003 ppm)[b]

Species	Upper limit (ppm)
H_2S	<400
C_2H_2, HCN, PH_3, etc.	50
N_2O	18
C_2H_4, CS_2, C_2H_6, etc.	6
CH_4	3.7
N_2O_4	3.3
SF_6, SiF_4, etc.	1.0
HCOOH	0.9
CH_2O	0.7
NO	0.7
COS	0.6
SO_2	0.5
C_3O_2	0.4
NH_3	0.4
NO_2	0.2
HCl	0.1
NO_2	0.1

[a] From Lewis, J. S., and Prinn, R. G. (1984). "Planets and Their Atmospheres," Academic, New York. Copyright 1984 Academic Press, New York.

[b] Very variable.

is fine-grained and cohesive, like firm sand or soil on earth. Viking experiments gave no evidence for organic molecules or for biological activity in the Martian soil, despite unusual chemical reactions produced by the soil and measured by the life detection experiments.

After the Viking Landers touched down on the surface of Mars in the summer of 1976, two decades elapsed before the surface of Mars was revisited by robotic spacecraft from the earth. On July 4, 1996, the Mars Pathfinder landed on the surface of Mars. The primary objective of the Pathfinder mission was to demonstrate a low-cost method of delivering a set of science instruments and a rover to the surface of Mars. Science objectives were to characterize surface features and to determine the chemical composition of the rocks and soil at the landing site and to monitor atmospheric and weather conditions during the course of the mission.

The Mars Pathfinder painted an interesting picture of the red planet. Mars appeared more and more like a planet that was very earth-like, with a warmer and wetter atmosphere in its early history. Flowing water and weathering processes created a variety of rock types and surface features. Pathfinder photography and measurements indicated water-worn rock conglomerates and sand and surface features that were created by running liquid water. Closeup images of sand dunes around the Pathfinder landing site exhibited clear evidence that weathering processes, such as erosion, winds, and flowing water, contributed to the present landscape of Mars. Flowing water was found to be a dominant agent in forming the surface of Mars.

At a press conference at NASA Headquarters in Washington, D.C., on August 7, 1996, just a few weeks after the surface of Mars was visited by the Mars Pathfinder and Sojourner Rover, scientists announced that a meteorite (ALH84001) found in the Antarctic and believed to have originated on Mars may contain ancient, fossilized, microscopic life. The meteorite was catapulted away from Mars 15 million years ago when a huge comet or asteroid impacted the surface of Mars. The meteorite traveled through space for millions of years and then encountered the earth. The meteorite entered the earth's atmosphere about 13,000 years ago and landed in the Antarctica. The meteorite lay there until 1984, when it was found by NASA scientists in the Allan Hills ice field (hence, the meteorite is named ALH84001—ALH for Allan Hills and 84 for its discovery in 1984), and brought back to the NASA Johnson Space Center. In 1993, it was identified as a Mars meteorite, one of only 12 "SNC" meteorites, which match the unique chemical signature of Mars. At the press conference, the scientists presented four independent lines of evidence which, when taken together as a whole, suggest that ancient life on Mars is a logical conclusion: (1) Carbonate globules: The carbonate patterns found in the meteorite form a unique signature of life, and the density and composition of the carbonate patterns is consistent with how terrestrial operate. (2) Polycyclic aromatic hydrocarbons (PAHs): Organic compounds usually created by bacteria were found in the meteorite. (3) Magnetite globules: These globules are created by bacteria on earth, as well as by some chemical processes. (4) The most dramatic evidence of all, pictures were shown at the press conference of worm-like structures present in the meteorite. While these features are much smaller than terrestrial bacteria, they look very similar, but could also be mineral structures.

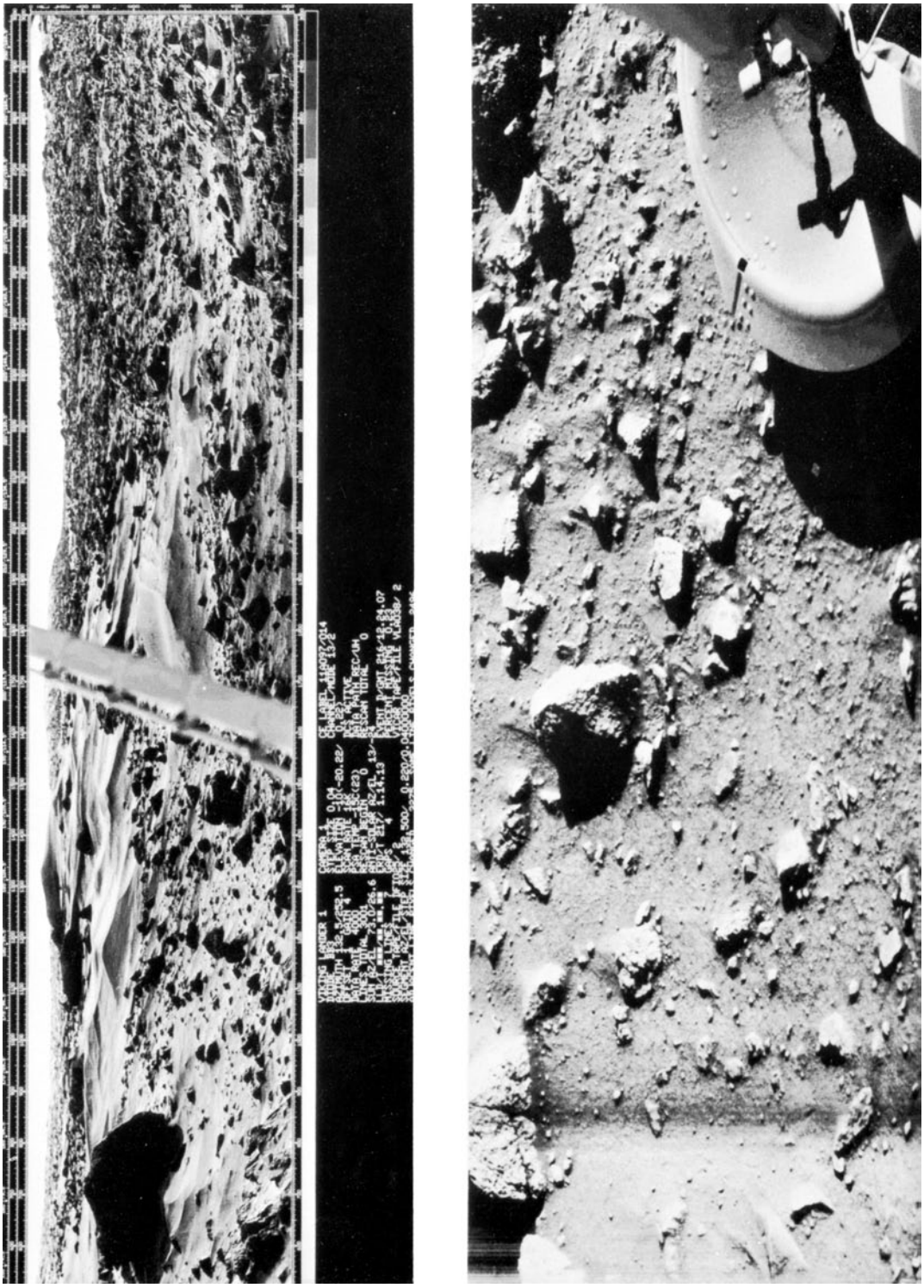


FIGURE 5 Surface of Mars as photographed by the Viking 1 lander. Upper photograph shows the Martian landscape with features very similar to those seen in the deserts of earth. Lower photograph, the Ærstever taken on the surface of Mars, was obtained just minutes after the Viking 1 landed on July 20, 1976.

TABLE VIII The Large Satellites of the Outer Planets: Physical Characteristics[a]

Planet	Satellite	Distance from planet (10^3 km)	Sidereal period (days)	Radius (km)	Density (g cm^{-3})	Surface gravity (cm sec^{-2})
Earth	Moon	384	27.3	1738	3.34	162
Jupiter	Io	422	1.77	1815	3.55	181
	Europa	671	3.55	1569	3.04	132
	Ganymede	1070	7.15	2631	1.93	144
	Callisto	1883	16.7	2400	1.83	125
Saturn	Titan	1222	16	2575	1.89	136
Neptune	Triton	355	6	∼2500	∼2.1	∼150

[a] From Strobel, D. F. (1985). *In* "The Photochemistry of Atmospheres: Earth, the Other Planets, and Comets" (J. S. Levine, ed.), Academic, Orlando, FL. Copyright 1985 Academic Press.

It is interesting to note that in 1976, the Viking Landers searched for organic materials in the soil of Mars. None were found, but we learned of a surface where organic material is absent, apparently as a result of intense solar ultraviolet radiation striking the surface (in the absence of a protective ozone layer) and strongly oxidizing compounds on the surface. Life in the Mars meteorite is consistent with the conclusions of Mars Pathfinder that early Mars may have been very different than present-day Mars. Early Mars may have been warmer and wetter and more hospitable for life to form and evolve. If indeed, early Mars was a warm, wet planet, what process or processes were responsible for transforming it to its present cold, dry state? To try to answer this question and the question concerning the possibility of life on early Mars and present-day Mars, NASA developed a new strategy to investigate Mars. Instead of waiting for two decades to pass as we did from Viking to Pathfinder,

TABLE IX Missions to Jupiter and Saturn[a]

Name	Launch date	Mission remarks
Pioneer 10	Mar. 3, 1972	Investigation of the interplanetary medium, the asteroid belt, and the exploration of Jupiter and its environment. Closest approach to Jupiter 130,000 km on Dec. 3, 1973. Exited solar system June 14, 1983; still active.
Pioneer 11 Jupiter/Saturn	Apr. 6, 1973	Obtained scientific information beyond the orbit of Mars with the following emphasis: (a) investigation of the interplanetary medium; (b) investigation of the nature of the asteroid belt; (c) exploration of Jupiter and its environment. Closest approach to Jupiter on Dec. 2, 1974; Saturn encounter: Sept. 1, 1979.
Voyager II	Aug. 20, 1977	Voyager II encountered Jupiter July 9, 1979, Saturn Aug. 26, 1981.
Voyager I	Sept. 5, 1977	Uranus Jan. 24, 1986, and Neptune, Aug. 24–25, 1989. Voyager I encountered Jupiter Mar. 5, 1979, and Saturn Nov. 13, 1980. Both returned a wealth of information about these two giant planets and their satellites including documentation of active volcanism on Io, one of the Galilean satellites.
Ulysses	Oct. 6, 1990	Ulysses was the first spacecraft to explore the sun's influence on interstellar space. A trip around Jupiter was required to give the spacecraft the gravity assist that it needed to accomplish its solar mission. Ulysses passed by Jupiter on Feb. 8, 1992, and obtained measurements of the magnetosphere and radiation environment around Jupiter.
Galileo Orbiter/Probe	Oct. 18, 1989	The Galileo Orbiter was the first spacecraft to orbit an outer planet, and the Galileo Probe was the first spacecraft to sample the atmosphere of an outer planet. The Orbiter and Probe arrived at Jupiter on Dec. 7, 1995, and studied Jupiter's atmosphere and magnetosphere and the four largest moons for 2 years (1995–1997). During an extended mission (1997–1999), Galileo investigated the moons Europa and Io in further detail.
Cassini Orbiter/ Huygens Probe	Oct. 15, 1997	Cassini is scheduled to go into orbit around Saturn on July 1, 2004. The Huygens Probe, built by the European Space Agency, is scheduled to enter Saturn's large moon, Titan, on Nov. 27, 2004. On route to Saturn, the Cassini spacecraft passed Venus twice (Apr. 26, 1998, and June 24, 1999), earth (Aug. 18, 1999), the asteroid Masursky (Jan. 23, 2000) and Jupiter (Dec. 30, 2000).

FIGURE 6 Jupiter as photographed by Voyager 1. The Great Red Spot can be seen in the lower center of the photograph. The atmosphere of Jupiter is massive and completely covered with clouds and aerosol haze layers.

NASA planned a Mars exploration strategy of sending a mission at every launch opportunity, which occurs every 267 months. Unfortunately, this new scientific assault on the Red planet resulted in significant failures. Out of four Mars missions [Mars Global Surveyor (1996), Mars Climate Orbiter (1998), and Mars Polar Lander and Deep Space 2 (1999)], only the Mars Global Surveyor was successful (see Table VI).

The Mars Global Surveyor (MGS) was launched on Nov. 7, 1996, and arrived at Mars on Sept. 11, 1997. The objective of the mission was to map the entire planet at high resolution and gather data on surface morphology, topography, gravity, weather and climate, surface and atmospheric composition, and magnetic field. The MGS carried wide-angle (surface resolution: 280 m per pixel at nadir and 2 km per pixel at the limb) and narrow-angle cameras (surface resolution of 1.4 m per pixel), a thermal emission spectrometer (TES), the Mars Orbiter Laser Altimeter (MOLA), and a magnetometer. TES obtained infrared scans of the planet to determine the general mineral composition of the surface (surface resolution: 9 km^2). MOLA transmitted a 1.06-μm laser beam to the surface of Mars to map the surface topography of the surface of Mars (horizontal resolution: 160 m, vertical resolution: local: 2 m, global: 30 m). The magnetometer mapped the spatial variability of the magnetic field of Mars.

The evidence gathered by the Viking Orbiters and Landers, the Mars Partfinder/Sojourner, and Mars Surveyor Missions painted a new picture of Mars. Early Mars was very different from present-day Mars. Early Mars was

warm and wet, with liquid water flowing on its surface. Early Mars appears to have been hospitable for the origin and life. Mars today is cold and very dry and not at all hospitable for life! The next three missions to Mars, the Mars Climate Orbiter (MCO, launched Dec. 11, 1998), the Mars Polar Lander (MPL, launched Jan. 3, 1999), and the Deep Space 2 or Mars Microprobe Mission (MMM, launched with the Mars Climate Orbiter on Jan. 3, 1999), were to investigate the possibility of water on Mars, past and present, and to search for clues or reasons for global climate change on Mars over its history. Unfortunately, all three missions were lost within a 3-month period as they entered the vicinity of Mars (MCO was lost on Sept. 23, 1999; MPL and MMM were lost on Dec. 3, 1999). The loss of three missions over a short period of time resulted in a reexamination of NASA's Mars exploration strategy. Once this reexamination is completed, NASA is committed to continue an active and viable program of Mars exploration, involving Mars orbiters, landers, robotic airplanes, sample-return missions, and eventually human exploration of Mars.

V. THE OUTER PLANETS

Jupiter, Saturn, Uranus, and Neptune are giant gas planets —great globes of dense gas, mostly molecular hydrogen and helium, with smaller amounts of methane, ammonia, water vapor, and various hydrocarbons produced from the photochemical and chemical reactions of these gases. They formed in the cooler parts of the primordial solar nebula, so gases and ices were preserved. These gas giants have ring systems and numerous satellites orbiting them. As already noted, the outer planets are more massive and larger and have very dense atmospheres that contain thick clouds and layers of aerosol and haze. The solid surfaces of the outer planets have never been observed, and we have only observed the top of the cloud and haze layers.

In many ways, Jupiter and Saturn are a matched pair, as are Uranus and Neptune. Jupiter and Saturn appear to have cores of silicate rocks and other heavy compounds comprising about 25 earth masses, surrounded by thick atmospheres of molecular hydrogen and helium. The total mass of Jupiter and Saturn are 318 and 95 earth masses, respectively. Uranus and Neptune appear to possess much less massive hydrogen/helium atmospheres relative to their cores. The total mass of Uranus and Neptune are only 14.5 and 17 earth masses, respectively.

The large satellites of Jupiter, Saturn, and Neptune are all larger than the earth's moon, with several comparable to the size of Mercury (see Table VIII for a summary of the orbital information and physical characteristics of these satellites). One of these satellites, Titan, the largest satellite of Saturn, has an appreciable atmosphere. Much of the new information about the atmospheres of Jupiter, Saturn, and Uranus their rings and satellites was obtained by the Voyager encounters of these planets (see Table IX for a summary of the missions to Jupiter and Saturn).

The Voyager spacecraft obtained high-resolution images of Jupiter, Saturn, and Uranus and their rings and satellites. Voyager instrumentation gathered new information on the chemical composition of their atmospheres. Jupiter, with its colorful banded, turbulent atmosphere, photographed by Voyager 1, is shown in Fig. 6. The Great Red Spot of Jupiter can be clearly seen in this photograph. The helium abundance in the atmosphere of Jupiter was found to be 11% by volume (with molecular hydrogen at 89% by volume), very close to that of the sun. The presence of methane, ammonia, water vapor, ethylene (C_2H_4), ethane (C_2H_6), acetylene (C_2H_2), benzene (C_6H_6), phosphine (PH_3), hydrogen cyanide (HCN), and germanium tetrahydride (GeH_4) in the atmosphere of Jupiter was confirmed (see Table X). The magnetosphere of Jupiter was found to be the largest object in the solar system, about 15 million km across (10 times the diameter of the sun). In addition to hydrogen ions, the magnetospherewas found to contain ions of oxygen and sulfur. A much denser region of ions was found in a torus surrounding the orbit of Jupiter's satellite, Io. The Io torus

TABLE X Composition of the Atmosphere of Jupiter[a]

Constituent	Volume mixing ratio[b]
H_2	0.89
He	0.11
CH_4	0.00175
C_2H_2	0.02 ppm
C_2H_4[c]	7 ppb
C_2H_6	5 ppm
CH_3C_2H[c]	2.5 ppb
C_6H_6[c]	2 ppb
CH_3D	0.35 ppm
NH_3[d]	180 ppm
PH_3	0.6 ppm
H_2O[d]	1–30 ppm
GeH_4	0.7 ppb
CO	1–10 ppb
HCN	2 ppb

[a] From Strobel, D. F. (1985). *In* "The Photochemistry of Atmospheres: Earth, the Other Planets, and Comets" (J. S. Levine, ed.), Academic, Orlando, FL. Copyright 1985 Academic Press.
[b] ppm ≡ parts per million; ppb ≡ parts per billion.
[c] Tentative identification, polar region.
[d] Value at 1–4 bar.

FIGURE 7 Ganymede, the largest satellite of Jupiter, as photographed by Voyager 2. The photograph shows a large, dark circular feature about 3200 km in diameter. The bright spots dotting the surface are relatively recent impact craters, while lighter circular areas may be older impact areas.

emits intense ultraviolet radiation and also generates aurora at high latitudes on Jupiter. In addition to a Jovian aurora, huge lightning flashes and meteors were photographed by Voyager on the nightside of Jupiter. A thin ring surrounding Jupiter, much narrower than Saturn's, was discovered by Voyager. The four large Galilean satellites (Ganymede, Callisto, Europa, and Io) were studied in detail (see Figs. 7–10 for Voyager photographs of these geologically varied satellites of Jupiter).

Jupiter's four largest moons are particularly intriguing since each has its own unique characteristics. Ganymede (Fig. 7) is the largest moon in the solar system and is the first moon known to have its own magnetic field. Callisto (Fig. 8) is extremely heavily cratered. Europa's surface (Fig. 9) is mostly water ice, and there is strong evidence that it may be covering an ocean of water or slushy ice. Io (Fig. 10) is the most active volcanic body in the solar system. Io's surface is mainly sulfur in different forms. Ganymede, Io, and Europa all appear to have a layered interior structure, as does the earth. Ganymede and Europa have a core, a rock envelope around the core, a thick soft ice layer (which on Europa could be liquid), and a thin crust water ice. Io appears to have a core and a mantle of at least partially molten rock, topped by a crust of solid rock coated with sulfur compounds. Callisto appears to be an ice–rock mixture both inside and outside. The four Galilean moons are very intriguing worlds.

The Galileo Orbiter and Probe was launched on Oct. 18, 1989, and reached Jupiter on Dec. 7, 1995. On December 7, 1995, the Galileo Probe parachuted about 150 km through the atmosphere of Jupiter and measured temperature, pressure, chemical composition, cloud characteristics, winds, sunlight, atmospheric lightning, and energy internal to the planet, during its brief (approximately 1-hr) life before it was destroyed by the high pressure

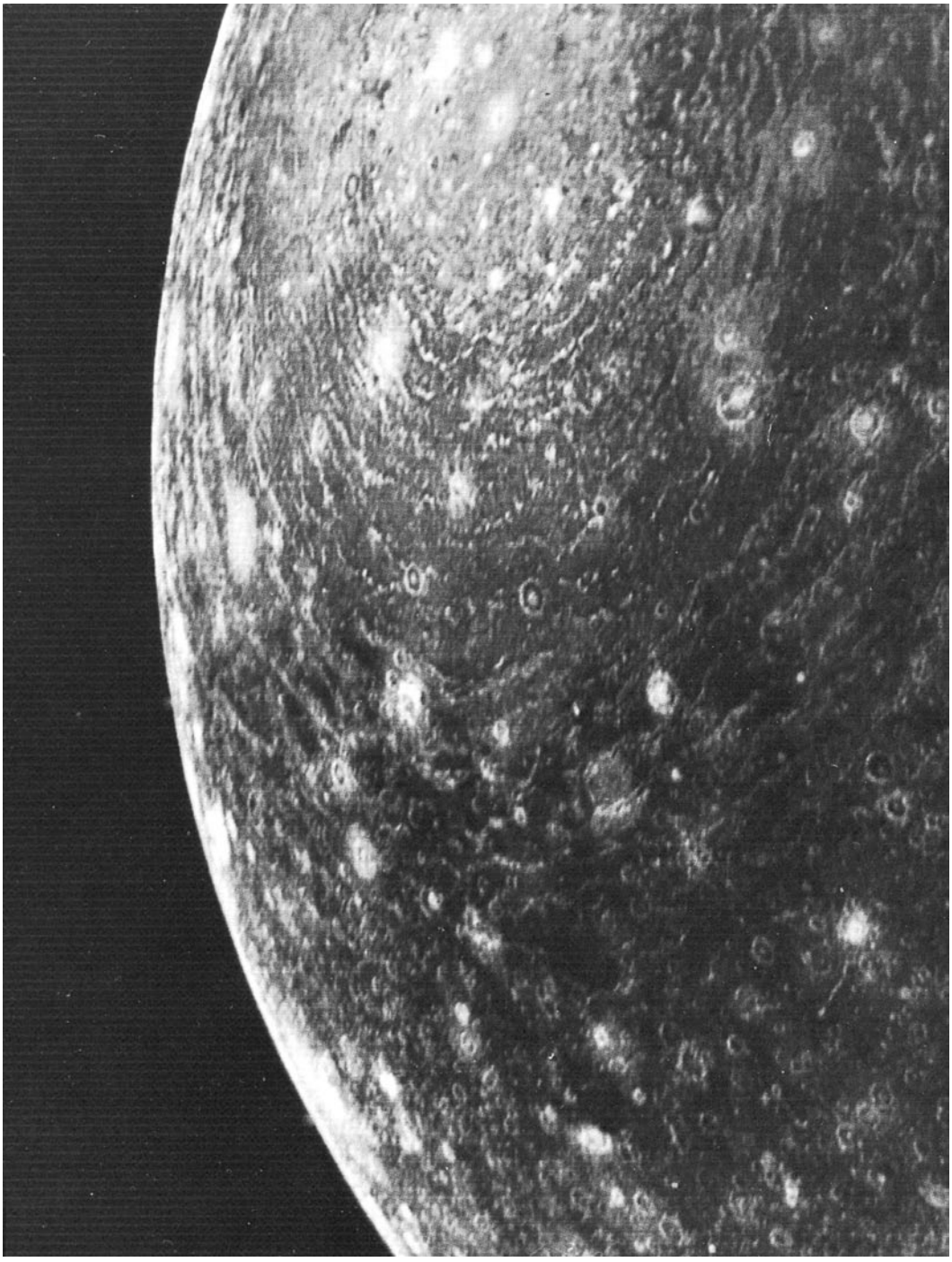


FIGURE 8 Callisto, the second largest satellite of Jupiter, as photographed by Voyager 1. Far more craters appear on the surface of Callisto than on the surface of Ganymede, suggesting that Callisto may be the oldest satellite of Jupiter.

and/or high temperature of Jupiter. The Galileo Orbiter investigated the atmosphere and magnetosphere of Jupiter, and the four large Galilean moons of Jupiter—Ganymede, Callisto, Europa, and Io. The Galileo Orbiter had two very close encounters with Io: 612 km on October 10, 1999, and 300 km on Nov. 25, 1999, and a close encounter with Europa of 350 km on January 3, 2000. The encounter of October 10 showed a lava field near the center of an

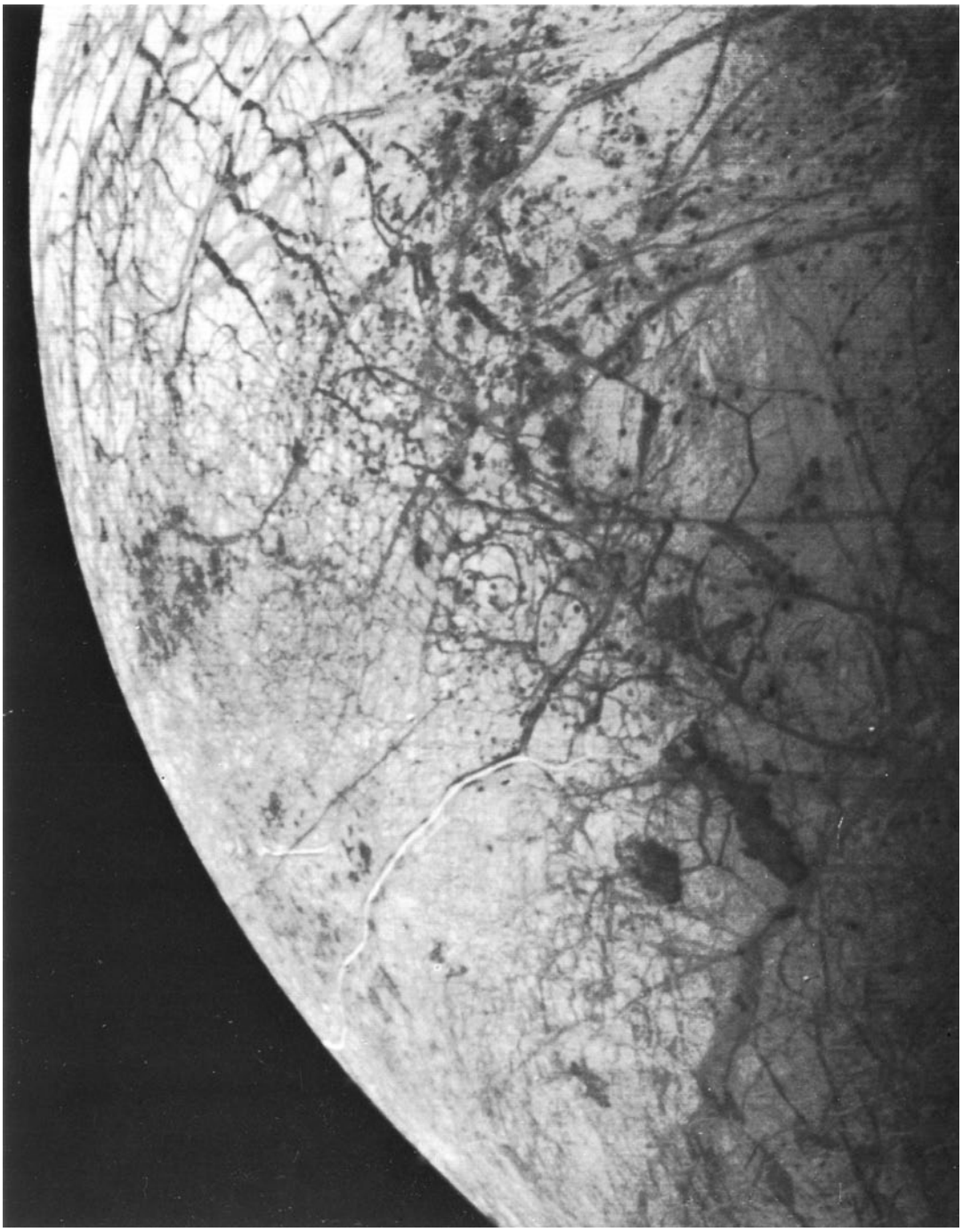


FIGURE 9 Europa, smallest of Jupiter's Galilean satellites, as photographed by Voyager 2. It is believed that Europa has a reasonable quantity of water in the form of a mantle of ice with interior slush, perhaps 100 km thick.

erupting volcano on Io. During the encounter of Nov. 25, 1999, a fiery lava fountain shooting more than than a mile above Io's surface was photographed. (Lava fountains on earth rarely exceed a few hundred yards in height.) Io is the most volcanically active body in the solar system. The close encounter with Europa on January 3, 2000, indicated that a liquid ocean lies beneath Europa's icy crust. Measurements suggest that the ocean lies beneath the surface somewhere in the outer 100 km, the approximate thickness of the ice–water layer.

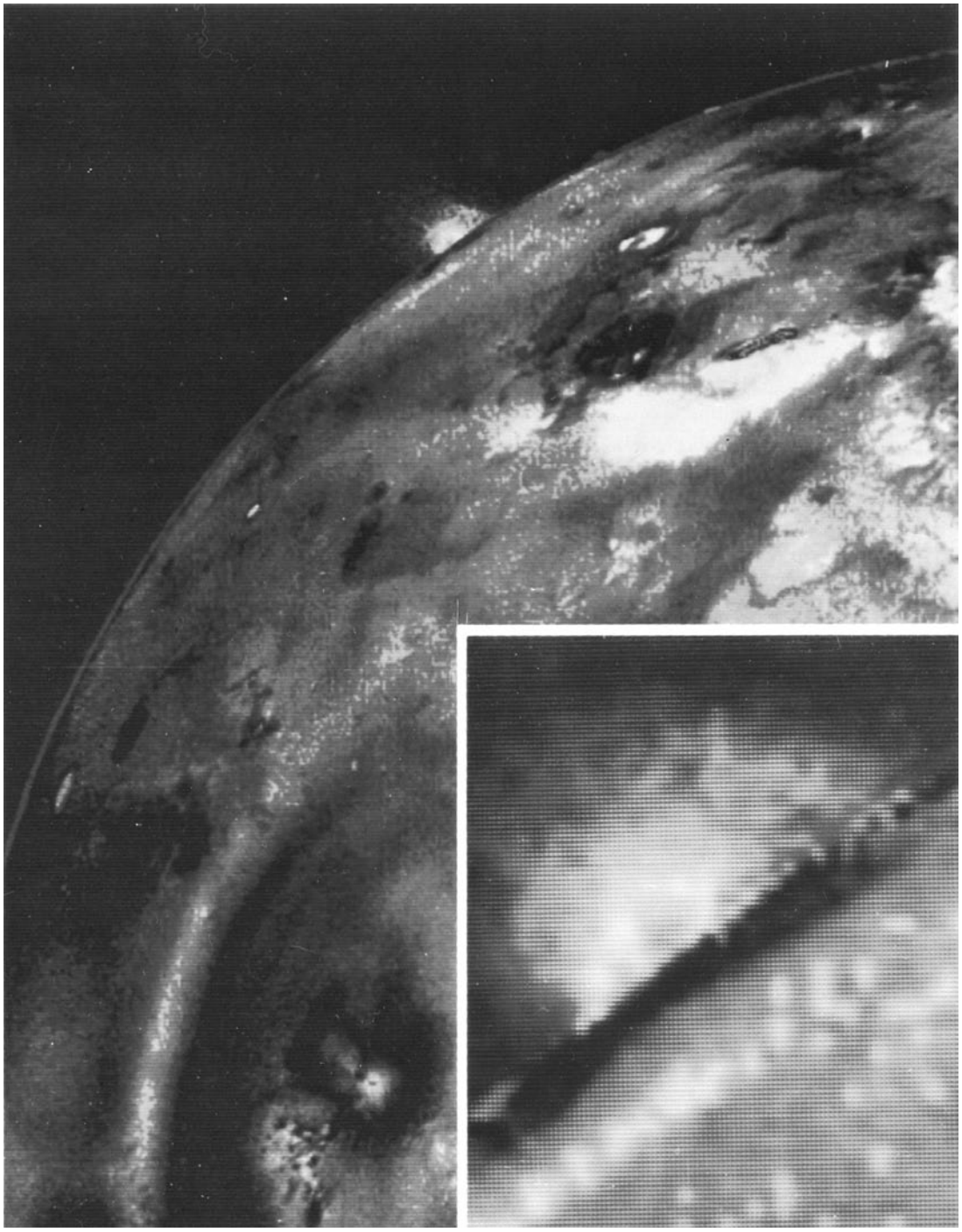


FIGURE 10 Io, satellite of Jupiter photographed by Voyager 1. An enormous volcanic eruption can be seen silhouetted against space over Io's bright limb. Solid volcanic material has been ejected up to an altitude of about 160 km.

Some of the key findings made by the Galileo mission include (1) the discovery of an intense radiation belt above the cloud tops on Jupiter, (2) the finding that helium concentration in the atmosphere of Jupiter is about the same concentration as in the sun, (3) extensive and rapid resurfacing of the moon Io due to widespread and frequent volcanism, (4) evidence that Io has a giant iron core that takes up half its diameter, (5) the presence of a magnetic field

TABLE XI Composition of the Atmosphere of Saturn[a]

Constituent	Volume mixing ratio
H_2	0.94
He	0.06
CH_4	0.0045
C_2H_2	0.11 ppm
C_2H_6	4.8 ppm
$CH_3C_2H^b$	No estimate
$C_3H_8^b$	No estimate
CH_3D	0.23 ppm
PH_3	2 ppm

[a] From Strobel, D. F. (1985). *In* "The Photochemistry of Atmospheres: Earth, the Other Planets, and Comets" (J. S. Levine, ed.), Academic, Orlando, FL. Copyright 1985 Academic Press.
[b] Tentative identification.

around Ganymede, (6) evidence that the heavily cratered Callisto may have a subsurface ocean, deep enough inside the moon that it does not affect the surface, and (7) evidence for liquid water oceans under the moon Europa's icy surface.

After encountering Jupiter and its satellites, both Voyager spacecraft visited Saturn and its satellite system (see Fig. 11). The six previously known rings were found to be composed of innumerable, individual ringlets with very few gaps observed anywhere in the ring system. Complex dynamical effects were photographed in the ring system, including spiral density waves similar to those believed to generate spiral structure in galaxies. The helium content of the atmosphere of Saturn was found to be about 6% by volume (with molecular hydrogen at about 94% by volume), compared with about 11% for Jupiter. The trace gases in the atmosphere of Saturn are similar to those in the atmosphere of Jupiter and include methane, acetylene, ethane, phosphene, and propane (C_3H_8) (see Table XI).

Titan, the largest satellite of Saturn, was found to have a diameter slightly smaller than that of Jupiter's largest satellite, Ganymede (see Table VIII). The atmosphere of Titan is covered by clouds and layers of aerosols and haze and has a surface pressure of about 1.5 bar, which makes it about 50% more massive than the earth's atmosphere. The surface temperature of Titan is a cold 100 K. The cloud- and haze-covered Titan is shown in Figs. 12 and 13, obtained by Voyager. Titan's atmosphere is mostly molecular nitrogen, with smaller amounts of methane and trace

TABLE XII Composition of the Atmosphere of Titan[a]

Constituent	Volume mixing ratio		
N_2	0.76–0.98[b]		
	Surface	**Stratosphere**	**Thermosphere (3900 km)**
CH_4	0.02–0.08	$\leq$0.026	0.08 ± 0.03
Ar	<0.16		<0.06
Ne	<0.002		<0.01
CO	60 ppm		<0.05
H_2	0.002 ± 0.001		
C_2H_6		20 ppm	
C_3H_8		1–5 ppm	
C_2H_2		3 ppm	~0.0015 (3400 km)
C_2H_4		0.4 ppm	
HCN		0.2 ppm	<0.0005 (3500 km)
C_2N_2		0.01–0.1 ppm	
HC_3N		0.01–0.1 ppm	
C_4H_2		0.01–0.1 ppm	
CH_3C_2H		0.03 ppm	
CO_2		1–5 ppb	

[a] From Strobel, D. F. (1985). *In* "The Photochemistry of Atmospheres: Earth, the Other Planets, and Comets" (J. S. Levine, ed.), Academic, Orlando, FL. Copyright 1985 Academic Press.
[b] Preferred value.

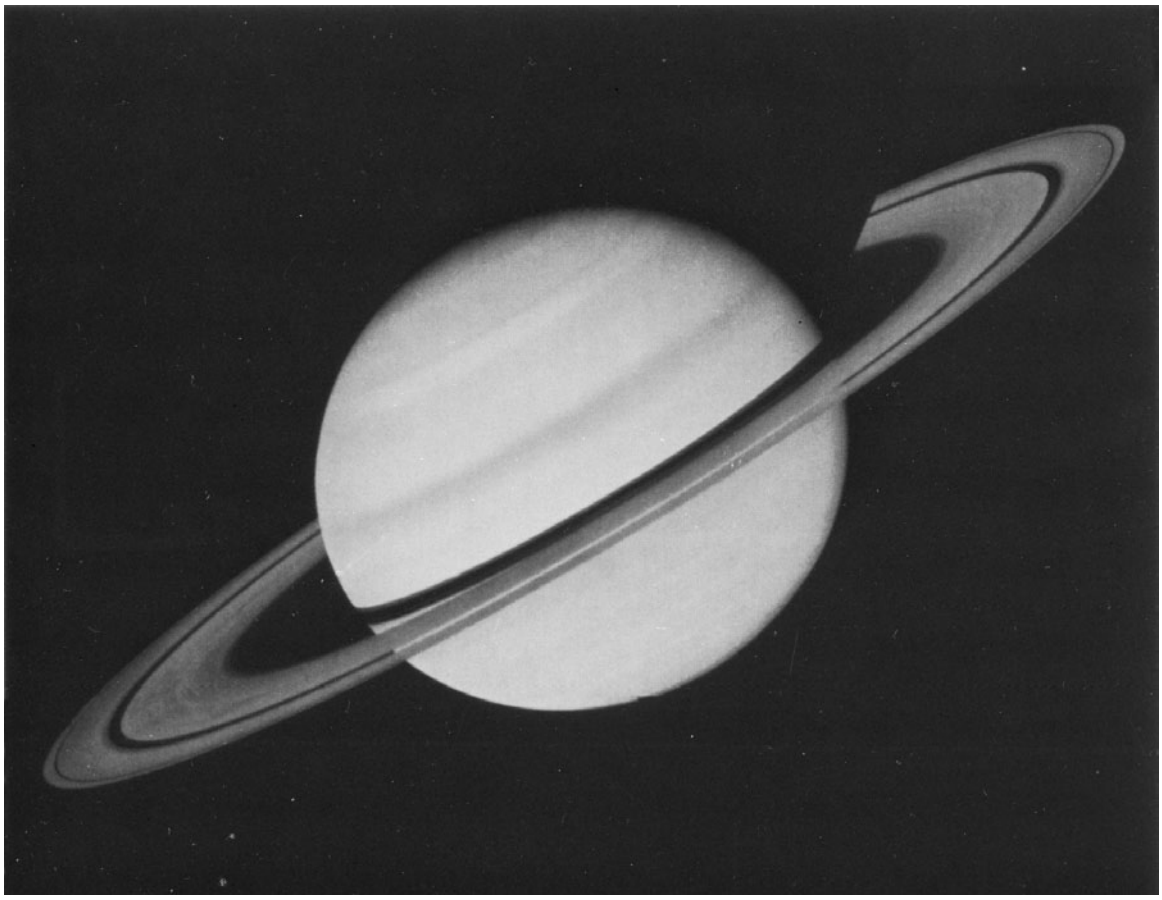


FIGURE 11 Saturn photographed by Voyager 1. Like Jupiter, the atmosphere of Saturn is massive and completely covered with clouds and aerosol and haze layers. The projected width of the rings at the center of the disk is 10,000 km, which provides a scale for estimating feature sizes on the image.

amounts of carbon monoxide, carbon dioxide, and various hydrocarbons (see Table XII for the chemical composition of Titan in different regions of its atmosphere). The surface of Titan may hold a large accumulation of liquid methane.

After encountering Saturn, Voyager 2 was targeted for Uranus. On January 24, 1986, Voyager 2 had its closest approach to Uranus. Prior to this encounter, very little was known about Uranus, one of the three planets (Neptune and Pluto, the other two) not known to the ancients. Uranus was discovered accidentally by Sir William Herschel in March 13, 1787. Uranus is so far away from the sun (2,869.6 million km) it only receives about 1/400 of the incident solar radiation that the earth receives. Voyager's radio signals took 2 hr and 45 min to reach the earth from Uranus.

As Voyager approached Uranus, its cameras indicated that Uranus did not exhibit the colorful and very turbulent cloud structure of Jupiter or the more subdued cloud banding and blending of Saturn. The very low contrast face of Uranus exhibited virtually no detail (see Fig. 14). The atmosphere of Uranus, like those of Jupiter and Saturn, is composed primarily of molecular hydrogen (about 85%) and helium ($15 \pm 5\%$). Methane is present in the upper atmosphere and is also frozen out in the form of ice in the cloud layer. The methane in the upper atmosphere selectively absorbs the red portion of the spectrum and gives Uranus its blue-green appearance. The volume percentage of methane may be as much as 2% deep in the atmosphere. Acetylene (C_2H_2) with a mixing ratio of about 2×10^{-7} was also detected in the atmosphere of Uranus. The temperature of the atmosphere was found to drop to a minimum of about 52 K (at the 100-mbar pressure level) before increasing to about 750 K in the extreme upper atmosphere.

Uranus has a ring system (as do Jupiter and Saturn). Two new rings (designated 1986 U1R and 1986 U2R) were discovered in Voyager 2 images of Uranus. The ring system of Uranus consists of 11 distinct rings that range in distance from about 37,000 to 51,000 from the center of Uranus.

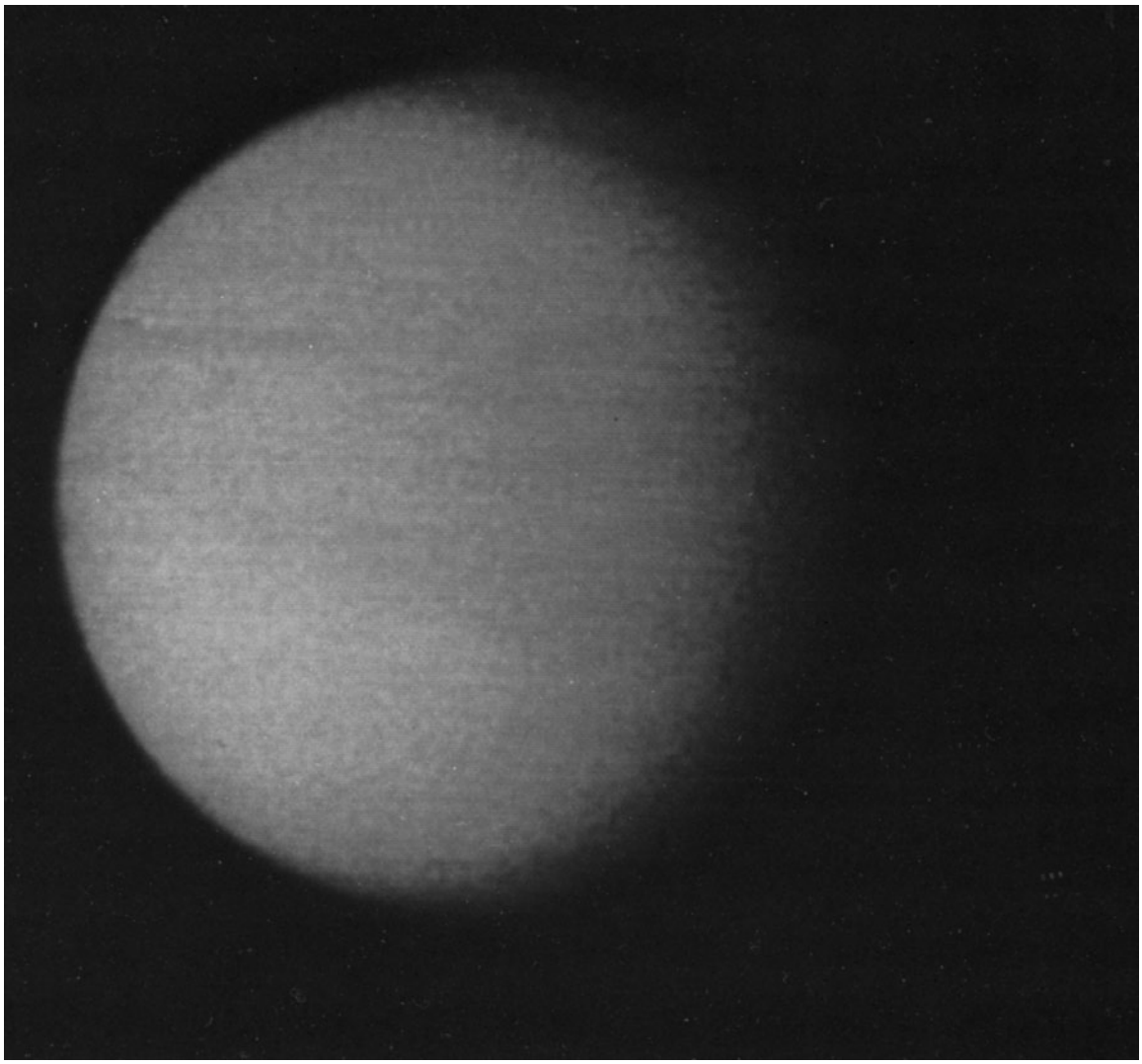


FIGURE 12 Titan, satellite of Saturn photographed by Voyager 1. Titan is the only satellite in the solar system with an appreciable atmosphere. The brownish-orange atmosphere of Titan contains clouds and haze and aerosol layers.

Prior to the Uranus encounter, five satellites were known to be orbiting Uranus: Miranda (distance = 129,000 km from the center of Uranus; diameter = 484 ± 10 km), Ariel (distance = 190,900 km; diameter = 1160 ± 10 km), Umbriel (distance = 266,000 km; diameter = 1190 ± 20 km), Titania (distance = 436,300 km; diameter = 1610 ± 10 km), and Oberon (distance = 583,400 km; diameter = 1550 ± 20 km). Ten new satellites were discovered on Voyager 2 images. All 10 satellites orbit Uranus within the orbit of Miranda (at distances that range from 49,700 to 86,000 km from the center of Uranus) and have diameters that range from about 40 to 80 km. Two of the newly discovered satellites are located within the ring system of Uranus and "shepherd" one of the rings.

Voyager 2 flew within 5000 km of Neptune on Aug. 25, 1989 (Fig. 15). Even though Neptune receives only about 3% as much sunlight as Jupiter does, it is a dynamic planet and, surprisingly, showed several large, dark spots reminiscent of Jupiter's hurricane-like storms. The largest spot, the "Great Dark Spot," is about the size of the earth and is similar to the Great Red Spot on Jupiter. At low northern latitudes, Voyager 2 captured images of cloud streaks casting their shadows on cloud decks below. The strongest winds on any planet were measured on Neptune. Most of the winds blow westward, opposite to the rotation of the planet. Near the Great Dark Spot, winds blow up to 20,000 km/hr. A small irregularly shaped, eastward-moving cloud was observed "scooting" around Neptune every 16 hr or so. This could be a cloud

FIGURE 13 Haze layers of Titan photographed by Voyager 1. The upper level of the thick aerosol layer above the satellite's limb appears orange. The divisions in the haze occur at altitudes of 200, 375, and 500 km above the limb.

plume rising above a deeper cloud deck. Auroras were detected on Neptune, but are much weaker than those on earth.

Six new moons were discovered by Voyager 2, bringing the total of moons around Neptune to eight. The names of the newly discovered satellites in order of increasing distance from Neptune are Nereid, Thalassa, Despoina, Galatea, Larissa, and Proteus. Neptune is also surrounded by four thin rings of varying thickness. Triton, the largest satellite of Neptune, has proved to be one of the most intriguing moons in the entire solar system. It shows evidence of a remarkable geological history, with active geyser-like eruptions spewing nitrogen gas and dark dust particles several kilometers into its tenuous atmosphere. Triton's relatively high density and retrograde orbit offer strong evidence that Triton is not an original member of Neptune's family, but is a captured object.

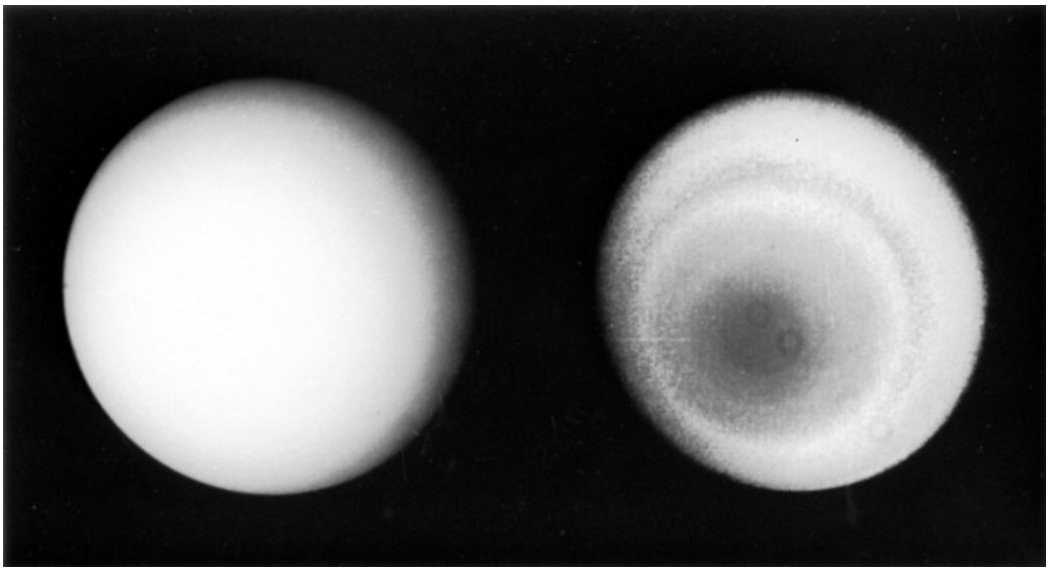


FIGURE 14 Two Voyager 2 views of Uranus—one in true color (left) and the other in false color. The picture on the left shows how Uranus would appear to human eyes from the vantage point of the spacecraft. The picture on the right uses contrast enhancement to bring out subtle details in the polar region of Uranus.

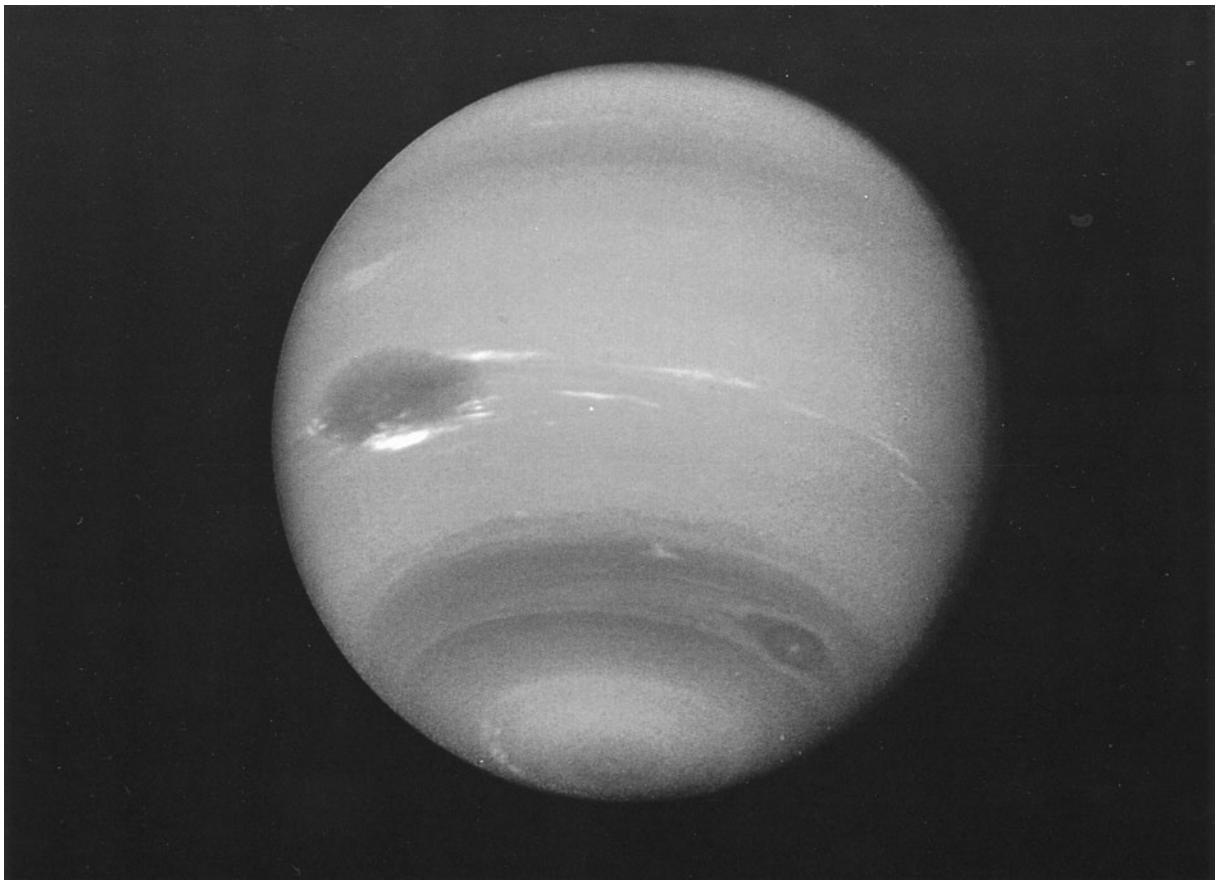


FIGURE 15 Voyager 2 photograph of Neptune obtained during its close encounter within 5000 km of planet on August 25, 1989. This photograph shows two of the four cloud features tracked including the Great Dark Spot (rotation period: 18.3 hr) and a smaller dark spot at the lower right of the planet (rotation period: 16.1 hr). The encounter with Neptune completed the tour of the planets with the exception of Pluto.

SEE ALSO THE FOLLOWING ARTICLES

ASTROCHEMISTRY • EARTH'S MANTLE • GREENHOUSE EFFECT AND CLIMATE DATA • OCEAN-ATMOSPHERIC EXCHANGE • PLANETARY GEOLOGY • PLANETARY SATELLITES, NATURAL • POLLUTION, AIR • PRIMITIVE SOLAR SYSTEM OBJECTS: ASTEROIDS AND COMETS • SOLAR SYSTEM, MAGNETIC AND ELECTRIC FIELDS • TERRESTRIAL ATMOSPHERIC ELECTRICITY • TROPOSPHERIC CHEMISTRY

BIBLIOGRAPHY

Barth, C. A. (1985). *In* "The Photochemistry of Atmospheres: Earth, the Other Planets, and Comets" (J. S. Levine, ed.), pp. 337–392, Academic, Orlando, FL.

Cameron, A. G. W. (1968). *In* "Origin and Distribution of the Elements" (L. H. Ahrens, ed.), pp. 125–143, Pergamon, New York.

Levine, J. S., ed. (1985a). "The Photochemistry of Atmospheres: Earth, the Other Planets, and Comets," Academic, Orlando, FL.

Levine, J. S. (1985b). *In* "The Photochemistry of Atmospheres: Earth, the Other Planets, and Comets" (J. S. Levine, ed.), pp. 3–38, Academic, Orlando, FL.

Lewis, J. S., and Prinn, R. G. (1984). "Planets and Their Atmospheres: Origin and Evolution," Academic, New York.

Prinn, R. G. (1985). *In* "The Photochemistry of Atmospheres: Earth, the Other Planets, and Comets" (J. S. Levine, ed.), pp. 281–336, Academic, Orlando, FL.

Space Studies Board (1994). "An Integrated Strategy for Planetary Sciences 1995–2010," Committee on Planetary and Lunar Exploration, Space Sciences Board, National Research Council, National Academy Press, Washington, DC.

Strobel, D. F. (1985). *In* "The Photochemistry of Atmospheres: Earth, the Other Planets, and Comets" (J. S. Levine, ed.), pp. 393–434, Academic, Orlando, FL.

Walker, J. C. G. (1977). "Evolution of the Atmosphere," Macmillan, New York.

Watters, T. R. (1995). "Planets: A Smithsonian Guide," Macmillan, New York.

Planetary Geology

Raymond E. Arvidson
Washington University, St. Louis

GLOSSARY

Accretion Formation of the planets and satellites by collisional accumulation of smaller bodies called planetesimals.
Insolation Solar energy flux per unit area received by a planet or satellite.
Lithosphere Crust and rigid part of the upper mantle. Lithosphere tends to fracture rather than flow in response to applied stresses.
Obliquity Tilt of planet's spin axis relative to normal to orbital plane. Earth and Mars have similar obliquities.
Planetary stratigraphy Subdiscipline of geology focused on mapping rock units exposed on planetary surfaces and on placing them in relative and absolute chronological sequences.
Plate tectonics Theory that Earth's lithosphere is divided into a number of rigid plates with spreading, convergent, and strike–slip boundaries.
Radar Acronym for radio detection and ranging. Technique using microwave part of spectrum to measure altimetry and scattering properties of planetary surfaces, particularly for Venus.
Siderophile element Element that is readily soluble in molten iron. These elements are deficient on the Moon.
Venera orbiter mission Soviet orbiters *Venera 15* and *16*, that acquired radar images and altimetry of mid- to high-northern latitudes of Venus in 1984 and 1985.
Viking mission Two orbiters and two landers that arrived in Mars orbit in 1976. Landers touched down and successfully operated on the surface for several years.
Volatile element An element that vaporizes at relatively low temperatures. The Moon is depleted of volatile elements.
Voyager mission *Voyager 1* and *2* were launched in 1977 on trajectories for encounters with the outer planets, rings, and satellites.

PLANETARY GEOLOGY is the study of the composition, structure, and distribution of ages of materials found on and within the solid bodies of the solar system and the use of this information to understand the processes involved in the formation and evolution of the objects. Planetary geology has provided insight into the formation of planetary bodies, including the ubiquitous role of

collisions among solar system bodies in accretion, disruption, and reaccretion. Analyses of the nature and history of tectonism and volcanism have provided crucial information on the thermal histories of planetary interiors, including (1) the formation of an early crust on the Moon, (2) the tectonic and volcanic processes on Earth's planetary neighbor, Venus, and (3) the role of Jupiter's tidal heating of the Galilean satellites Europa and Io. Evidence for significant period and secular climatic changes has been found on Mars. The periodic changes are driven by quasi-periodic variations in orbital parameters caused by gravitational interactions with other solar system bodies, while evidence for the secular change suggests that the early climate was warm enough to allow rainfall. Planetary geology is dependent on spacecraft exploration of the solar system. The rate of growth in data volume and complexity associated with missions planned in the 1990s allowed the posing of important, detailed questions about the origin and evolution of Venus, Mars, the Jovian and Saturnian systems, comets, asteroids, and the Moon.

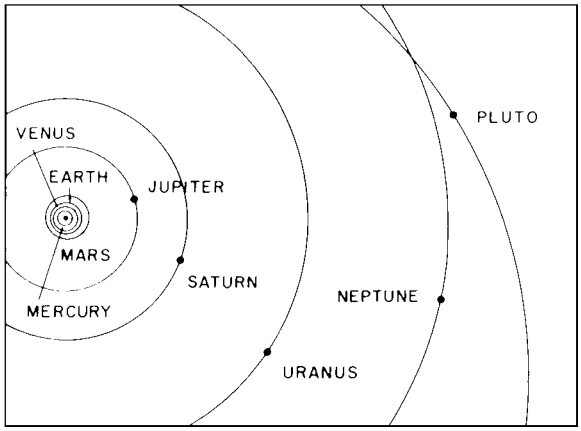


FIGURE 2 Plan view of solar system showing orbits of the planets.

I. INTRODUCTION

Planets and satellites can be separated into two broad classes on the basis of diameter and density (Fig. 1). The first class consists of objects with relatively large radii ($>10^4$ km) and low densities (<2 g/cm^3) and corresponds to the Jovian planets located in the outer parts of the solar system (Fig. 2). The other class contains objects that occupy a range of both size and density, although as a group, the bodies are smaller in size and higher in density than the Jovian planets. These relatively small, dense objects have solid surfaces that have preserved records of a number of processes that have operated over geological time scales. The inner planets, satellites, asteroids, and comets are included in this class. Planetary geology focuses on understanding the nature, magnitude, and frequency of processes that have operated on and within planetary-scale bodies. This information is largely obtained through

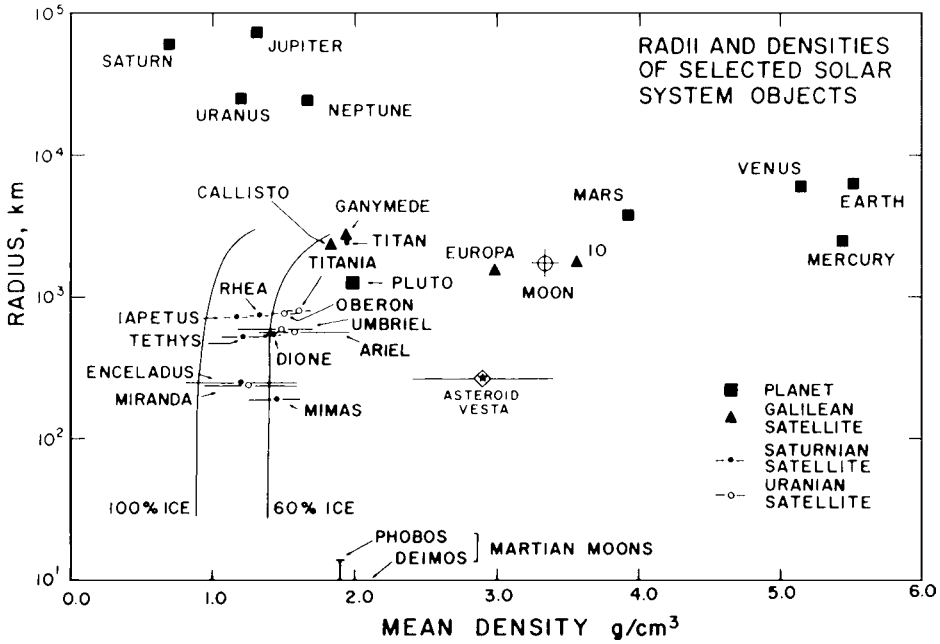


FIGURE 1 Radius is plotted against mean density for a number of solar system objects. The curves for 100 and 60% ice represent the mean densities expected as a function of size for bodies of those compositions.

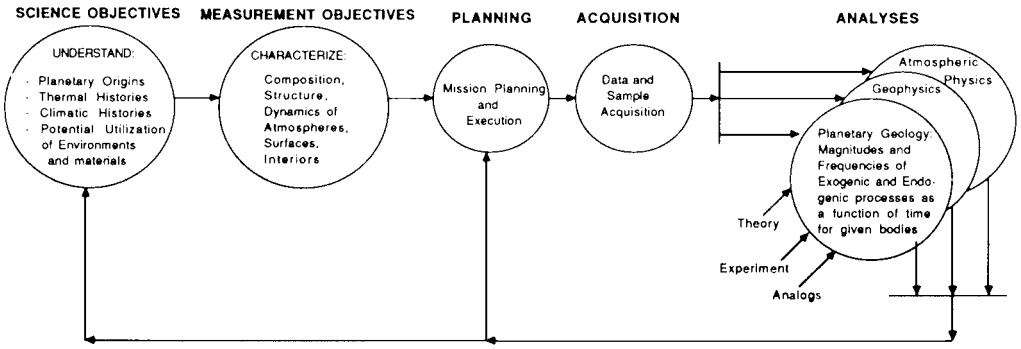


FIGURE 3 Flow diagram showing planetary exploration, including scientific objectives, measurement objectives, and tasks associated with planning, acquiring, and analyzing data. Analyses are shown by disciplines, although cross-discipline research is also common. Exogenic processes are those that are driven by external energy sources, for example, impact cratering or the hydrological cycle. Endogenic processes derive their energy from heat flowing from planetary interiors.

analysis of the composition, structure, and ages of materials exposed on planetary surfaces and is used to infer how these bodies formed and how they have evolved over time.

Planetary geology began as a distinct scientific discipline in the late 1960s, born with the advent of spacecraft missions. Figure 3 shows in schematic form how a typical planetary mission moves from defining science objectives covering a number of relevant disciplines to translating them into measurement objectives, implementing the mission, and analyzing the resultant data. This review is organized roughly along the lines shown in Fig. 3, including examples of how planetary geology has increased our knowledge of planetary origins, interior thermal histories, and climates. We first begin with how the science is conducted, and we then highlight key areas of ongoing research. The last section is a look at future missions, both near and long term.

II. PLANETARY GEOLOGY—THE APPROACH

A. Measurement Objectives

Meeting the broad geological, geophysical, and atmospheric science objectives shown on the left-hand side of Fig. 3 and outlined in Section I requires sets of measurements that provide information on the composition, structure, and dynamics of planetary atmospheres, surfaces, and interiors on a global basis. An additional measurement objective for planetary geology is the determination of the ages of materials that comprise the crusts of planetary surfaces. The information on composition, structure, and ages of rock units exposed on planetary surfaces is used in a planetary geology subdiscipline called planetary stratigraphy to develop global-scale models of the timing and processes involved in the accumulation of rock strata. These stratigraphic models provide a global context for other analyses, such as inferring what types of tectonic processes have operated over time or, for bodies with atmospheres, what climatic regimes have prevailed at various times.

Measurement objectives for planetary geology can be met in part by remote sensing from Earth and spacecraft. For example, both the Viking spacecraft that reached Mars in 1976 and the two Voyager spacecraft launched in 1977 to the outer planets included imaging systems on pointable platforms to acquire multispectral views of atmospheres and surfaces. In addition, both missions included instruments that measured heat emission. The image data have been used to develop a first-order stratigraphy for Mars, its two moons, Phobos and Deimos, and for the satellites of the outer planets, while thermal emission data have been used to infer temperatures and thermophysical properties for various surface units. In addition, other wavelength intervals are utilized. For example, Earth-based and spacecraft radar data have been used to image the surface of Venus, since dense clouds preclude use of shorter wavelength intervals.

In situ observations are also important, including use of landers to directly determine surface composition and seismometers to characterize interior structure. Only the Moon, Mars, and Venus have been explored using lander missions. Generally, samples returned to Earth provide the greatest amount of detailed information because the full range of laboratory techniques, such as radiometric dating, can be brought to bear. Lunar samples were returned to Earth during the Apollo lander missions and during the Russian unmanned Luna missions. Earth proper is also a space mission, sweeping up samples of debris called meteorites as the planet moves through its orbit. Meteorites sample a variety of materials, including asteroids, comets,

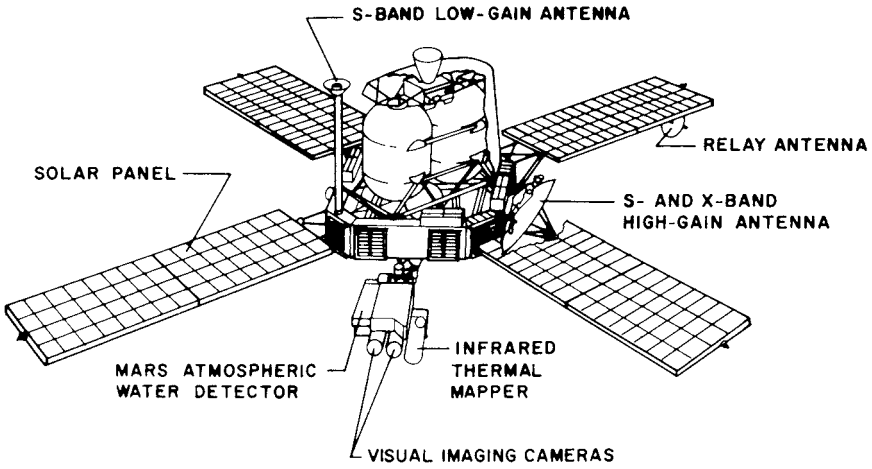


FIGURE 4 Schematic view of Viking orbiter spacecraft.

and perhaps even fragments ejected from the Martian surface during impact events.

B. Mission Planning and Data Acquisition

Meeting the science and measurement objectives associated with planetary geology necessarily requires considerable planning, since the discipline is highly dependent on spacecraft missions associated with national and international programs of solar system exploration. The reason is simple. The large distances between Earth and other solar system objects preclude global observations of high spatial resolution, except from spacecraft that fly by the objects or that are placed in orbits around them (Fig. 4). As noted in Section II.A, *in situ* observations require landers or rovers placed on planet and satellite surfaces. These spaceborne missions require national-scale bases of both financial and programmatic support.

Earth-based observations of planets and satellites do provide important supporting information needed to capitalize on data returned during spacecraft missions. Generally, Earth-based observations can extend over a much longer time period than spaceborne observations. For instance, the Lowell Observatory in Flagstaff, Arizona, maintains an archive containing thousands of photographic plates of Mars acquired over nearly a century. There are also fewer constraints in terms of instrumentation volume, weight, and data collection rates as compared to spaceborne observations. For example, the Arecibo radiotelescope in Puerto Rico and the Gold-stone system in California have obtained numerous images depicting the magnitude of radar backscattered from the Venusian surface. These images have spatial resolutions as small as 1 to 2 km. However, because of an orbital commensurability between the two planets, Venus points the same hemisphere toward Earth during closest approach of the two planets (inferior conjunction), when high-resolution observations can be acquired. Consequently, only one hemisphere can be imaged from Earth.

The rate at which spacecraft acquire data is decidedly nonuniform with respect to time because there are a finite number of missions, spaced over fairly wide time intervals (Fig. 5). Consider the *Mariner 10* mission to Venus and Mercury. *Mariner 10* was launched in November of 1973 using an Atlas/Centaur launch vehicle to reach a parking orbit about Earth. The Centaur upper stage rocket was then used to place the spacecraft into a solar orbit designed to pass close to Venus on its way to close encounters with Mercury. *Mariner 10* approached to within 5794 km of Venus in February 1974 and acquired a variety of data. Close encounters with Mercury occured in March and September 1974 and March 1975, with the spacecraft returning data during each close pass. During the third encounter, the spacecraft passed Mercury at an altitude of only 327 km. In March 1975, the spacecraft began an unprogrammed pitch maneuver, indicating that its attitude control gas was depleted. Commands were then sent to the spacecraft, shutting off its transmitter, and completing the *Mariner 10* mission. During each of the four planetary encounters, a great deal of information was telemetered back to Earth during a time span of a few days, opening new avenues of research each time.

Science-based strategies for the design and implementation of missions to the inner planets and satellites have followed a general trend. A basic premise is the need for an initial, global-scale reconnaissance of a given object (for example, *Mariner 10* to Mercury), followed by systematic and more detailed global observations (for example, Viking orbiters around Mars), and then by *in situ* measurements on the surface (for example, Viking landers). Return

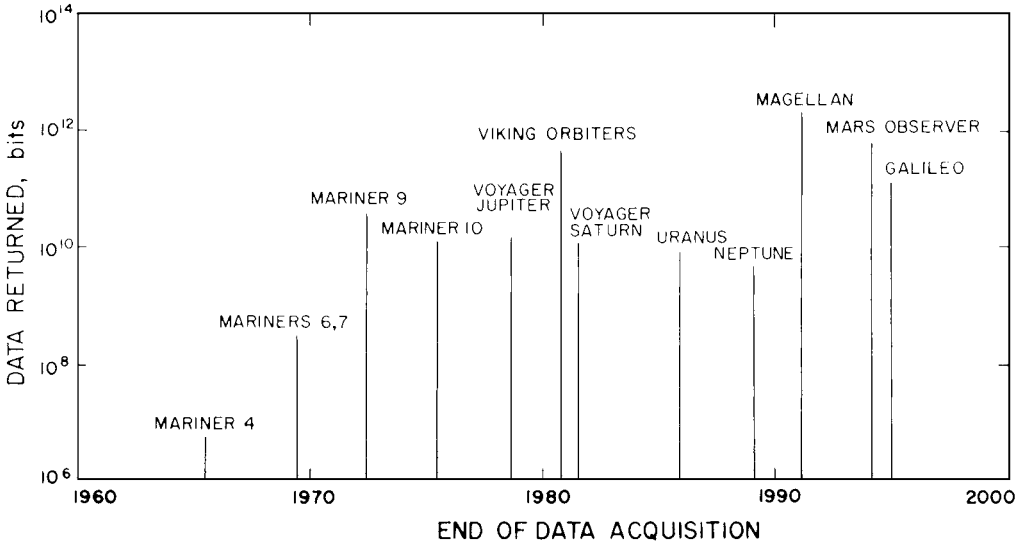


FIGURE 5 Plot showing growth of digital planetary image data. The comet rendezvous asteroid flyby (CRAF), Cassini, and lunar geoscience orbiter (LGO) missions are not shown because they have not yet been funded. The CRAF, Cassini, and LGO missions would each produce approximately 10^{12} bytes of data in the 1990s.

of samples to Earth is usually considered as a longer term objective. However, costs increase rapidly from flyby to orbiter to lander missions. Costs associated with returning samples are high indeed because of the need to get to the target, acquire data and samples, and get back to Earth. With programs of limited resources, balance would soon be lost as a greater number of orbiter and lander missions are required to meet scientific needs. Within the United States, an alternate, less expensive strategy has been developed and centers on the use of a standard spacecraft called the *Mariner-Mark II* and the development of advanced remote-sensing instrumentation that can be used on a number of flyby and orbital missions using the standard *Mariner-Mark II* bus. In fact, as is discussed in Section VI.A, planetary exploration plans of the United States through the end of the twentieth century focused on broad spacecraft reconnaissance of solar system objects, including Venus, Mars, the Jovian system, the Saturnian system, a comet, and an asteroid.

C. Analyses

As noted, Fig. 5 shows the current volume of digital planetary image data and volumes expected from missions to be flown in the 1990s. For example, the Magellan mission to Venus will double the existing amount of digital planetary image data within about 243 days of mapping operations. The overall volume is expected to increase exponentially with a doubling interval of only several years when averaged over the 1990s. Data complexity is expected to increase at a comparable rate. For example, the Galileo orbiter, expected to begin mapping operations in the Jovian orbit in 1995, will carry a near-infrared mapping spectrometer (NIMS) capable of generating discrete images at numerous narrow wavelength intervals. The NIMS will be complemented by a multispectral imaging system using charge-coupled device detectors. Data expected from the two instruments will enable new studies of Jupiter's atmosphere, its ring, and the large group of satellites associated with the planet.

The increase in volume and complexity of data sets expected in the 1990s is a result of two activities, proceeding in parallel. First, science and measurement objectives are becoming more complex as knowledge of planetary bodies and systems increases and more complex questions are posed. Second, technology is keeping pace, offering new observational capabilities and programmable instruments with numerous operating modes.

The expected increase in data volume and complexity requires efficient mechanisms for planning sequences of observations for data access, analysis, and archiving. The increased computation and data management capabilities expected in the 1990s will allow mission operations and data analyses to be geographically distributed, with researchers working at their home institutions and interacting with mission personnel over networks. New analysis techniques are also being developed to cope with growing data volumes and complexity. The NASA Planetary Data System was developed in response to the challenges associated with managing, distributing, processing, and archiving the expected data sets. This system, managed by the Jet Propulsion Laboratory, is based on a

geographically distributed network of sites or nodes that have experts involved for various data sets. Information about data, actual data, and selected processing procedures will be available to NASA's research community through the planetary data system. For example, compact disk read-only memory (CD-ROMs) platters are being utilized as a standard distribution medium for digital data, since each disk holds 550 million bytes of data. With CD-ROMs, hundreds of copies can be generated, significantly lowering reproduction costs over, for example, copying magnetic tapes. Even higher volumes are obtained for data sets that can be compressed. For example, Voyager image data are being distributed on CD-ROMs in a lossless, compressed form, where each disk holds the equivalent of 2 billion bytes of image data.

III. PLANETARY ORIGINS

A. Introduction

One of the most profound aspects of the planetary sciences is the information provided that allows contemplation of the origins of solar system bodies. An understanding of planetary origins requires input from a variety of disciplines, including aspects of astrophysics (star formation), orbital dynamics, and information gathered from geological observations of planetary surfaces and interiors. A recurring theme in the origin and evolution of planetary-scale bodies is the role of collisions, particularly in early geological time. However, collisions among solar system bodies have played major roles in all geological epochs. For example, it has been hypothesized that the impact of an asteroid or comet into Earth's ocean placed a great deal of vapor and dust into the stratosphere approximately 65 Ma, a time period now defined as the stratigraphic boundary between rocks of the Cretaceous (older) and Tertiary (younger) periods of geological time. The lack of sunlight associated with the resultant atmosphere vapor and dust would have catastrophically modified the Cretaceous ecosystem for several years and would explain the catastrophic disappearance of species (including dinosaurs) at the end of the Cretaceous period. This hypothesis is supported by studies of the probability of such impacts and by the worldwide occurrence of a clay layer at the Cretaceous–Tertiary boundary enriched in noble metals, such as iridium and platinum. These metals are rare on Earth's surface because they have been sequestered in Earth's metallic core. They are common in undifferentiated objects such as comets. In this section, we highlight the importance of collisions in earliest geological time, specifically the collisional accretion, disruption, and reaccretion of planetary-scale objects.

FIGURE 6 View of the heavily cratered terrae of the Moon as seen by *Mariner 10* on its way to encounters with Venus and Mercury. Dark circular maria can be seen on the left. Maria are impact basins filled with volcanic deposits of basaltic composition.

B. Evidence from the Cratering Record

The Moon is heavily cratered, as are large parts of Mercury, Mars, and most satellites of the outer planets (Fig. 6). Decades of research has shown that the vast majority of these craters formed by hypervelocity impact of planetesimals, asteroids, and comets with the planetary surface. The Moon provides a particularly good control for understanding how the rate of cratering has changed over time, since samples have been returned from sites of the frontside during the Apollo and Russian Luna missions and radiometrically dated in the laboratory. Further, Earth-based and orbital images for each of the sites allows detailed characterization of the areal abundance of impact craters.

A plot of the cumulative areal abundance of craters per unit area surrounding each of the Apollo landing sites on the Moon is plotted against age in Fig. 7. This plot shows the abundance of craters accumulated over the age of the surface. Considerable uncertainty exists as to whether the heavily cratered terrain on the Moon (called lunar terrae) records craters over a 3.8- or 4.5-Ga period. The terrae crust formed 4.5 Ga, but was intensely bombarded until approximately 3.8 Ga, based on the distribution of radiometric ages of lunar samples that have been impact metamorphosed. In either case, Fig. 7 shows that the rate of impact cratering declined exponentially from torrential values that pertained in early geological time.

Heavily cratered terrains, such as the lunar terrae, are common on many solar system objects and record a period of torrentially high impact bombardment rates. In fact, it seems probable that these surfaces are a record of the last stages of growth of planetary-scale objects in the

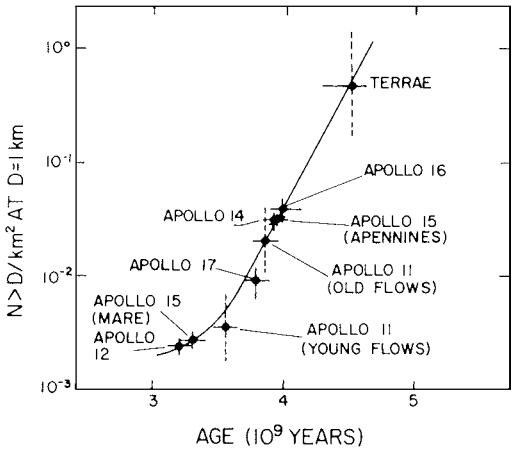


FIGURE 7 Cumulative abundance of impact craters larger than 1 km in diameter per unit area, plotted against ages of Apollo sites on the Moon. The terrae are shown at 4.5 Ga, but the crater abundance could have accumulated over an integral time period as low as 4 Ga. [From Neukum, G. *et al.* (1975). Cratering in the Earth–Moon system: Consequences for age determination by crater counting. *Proc. Lunar Sci. Conf., 6th*, pp. 2597–2620.]

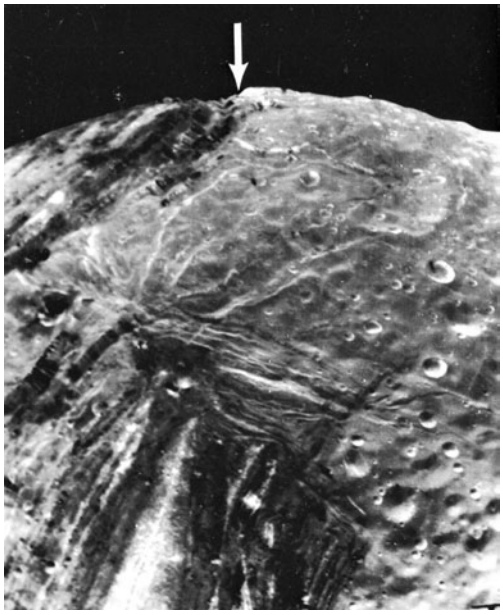


FIGURE 8 *Voyager 2* view of Uranian satellite Miranda, showing bright, heavily crater terrain, and parts of darker, grooved and ridged ovoid (upper left) and trapezoid (lower part of image) terrains. Arrow points to rift valleys separating ovoid and cratered terrain. The ovoid and trapezoid terrains may be a consequence of reaccretion of darker silicate material after disruption of a proto-Miranda.

solar system. These ancient, battered crusts convincingly demonstrate that the earliest history of the solar system was dominated by collisional interactions among solar system objects. This early history has been lost on Earth because of the high rates of surface erosion associated with a vigorous hydrological cycle and because of ongoing tectonism and volcanism.

The solar system formed approximately 4.5 Ga by the gravitational collapse of a gas–dust cloud, producing a protosun with most of the cloud mass and a disk-shaped planetary nebula. Adiabatic heating during collapse at first kept the nebula largely in gaseous form. Dynamical considerations show that as the cloud cooled, kilometer-sized (and larger) bodies called planetesimals first formed. Monte Carlo simulations show that these objects in turn collisionally interacted and accreted to form planetary-scale objects. Heavily cratered terrains record the last sweep-up of planetesimals.

C. Miranda—Reaccreted Satellite of Uranus

Miranda is the innermost satellite of the planet Uranus. Based on *Voyager 2* observations, this satellite is approximately 500 km in diameter and has a surface temperature of about 86 K and a mean density of 1.26 ± 0.39 g/cm^3. The low density implies that this object is primarily composed of water ice, and the low surface temperature ensures that the ice is highly viscous and capable of retaining topography associated with impact craters, tectonism, and volcanism over geological time scales. Miranda is densely cratered, with an abundance of craters between 10 to 50 km in diameter, which is a factor of three higher than for the lunar terrae (Fig. 8). The craters most likely represent the sweeping-up of planetesimals in heliocentric orbit that crossed the path of the early Uranian system. Extrapolating the abundance of impact craters on Miranda and other Uranian satellites back to 4.5 Ga implies that Miranda would have been hit several times with planetesimals large enough to disrupt the satellite, that is, it is probable that Miranda was accreted from icy planetesimals in orbit about Uranus, disrupted by impact with larger planetesimals in heliocentric orbits, and reaccreted a number of times during the last stages of formation of planetary-scale objects.

Miranda consists of two rather diverse terrains, a bright, rolling, highly cratered terrain of relatively uniform brightness and younger, complex ovoid- to trapezoid-shaped terrains characterized by subparallel sets of bright and dark bands, scarps, and ridges (Fig. 8). The abrupt juxtaposition of the trapezoid and ovoid terrains against the more uniform cratered terrain is unusual. It is thought that the darker terrains are the surface expressions of silicate planetesimals that accreted onto the current version of Miranda. The silicate fragments would have come from the core of an earlier version of the satellite, a version that was catastrophically disrupted by collision. Reaccreted

silicate planetesimals, because of their relatively high density, would sink, leaving behind dark silicate contaminates on the surface and ovoid to trapezoid terrains with the types of structures seen in the Voyager images.

D. Origin of the Moon

Analyses of lunar samples collected during the Apollo and Russian Luna missions show that the Moon is depleted in volatile materials. For example, the potassium-to-uranium ratio of the lunar mantle is about $\frac{1}{5}$ of the value postulated for Earth's mantle. Potassium is a volatile element that would have been retained as vapor in the planetary nebula at lower temperatures than uranium. Likewise, lunar samples are depleted in water, carbon dioxide, sulfur, and other elements that evaporate at relatively low temperatures. A second major finding is that the Moon has only about $\frac{3}{4}$ of the iron and ironlike metals (called siderophile elements) as predicted for the solar nebula. Finally, lunar samples show oxygen isotopic ratios similar to those found on Earth. Since the oxygen isotopic signature is thought to have varied with heliocentric distance in the solar nebula, this constraint means that the Moon formed at approximately the same distance from the Sun as did Earth. These results, coupled with the high ratio of lunar to Earth masses ($\sim 10^{-1}$) and the high value of angular momentum for the Earth–Moon system, have, in the past, presented problems for models of how the Moon formed.

An intriguing hypothesis for formation of the Moon that explains the relatively high Moon to Earth mass ratio and angular momentum involves an oblique collision of a Mars-sized planetesimal with an accreting Earth (Fig. 9). Monte Carlo models of planetary growth from formation and accretion of planetesimals show that the average planetesimal sizes increase with time as zones of collisional accretion develop. Collisions among relatively large objects are thus probable. A Mars-sized planetesimal impacting proto-Earth at a grazing angle would disperse vapor and hot debris into orbit around Earth (Fig. 9). Proto-Earth would have completed most of its accretion, and heat associated with accretion would already have triggered differentiation into crust, mantle, and core. Simulations suggest that most of the vapor and debris would come from the impactor, with additions from proto-Earth's crust and mantle. Expanding vapor would accelerate the material and keep some of it from reaccreting onto Earth. Within time scales of centuries, the materials would cool, form into a disk, and begin reaccreting to form the Moon. Volatiles would have been lost from the debris that formed the moon by heating of the ejected material and by vaporization. Siderophiles would have been depleted because the impact would involve proto-Earth's crust and mantle, which are already assumed to be depleted in siderophiles by the formation of a metallic core. Further, if the impactor was also differentiated, impact simulations show that fragments of the metallic core would have been preferentially reaccreted onto Earth because of their mass and structural integrity. Reasonable computer simulations have an impact velocity of about 2 km/s, with an impactor in heliocentric orbit with an eccentricity of about 0.20.

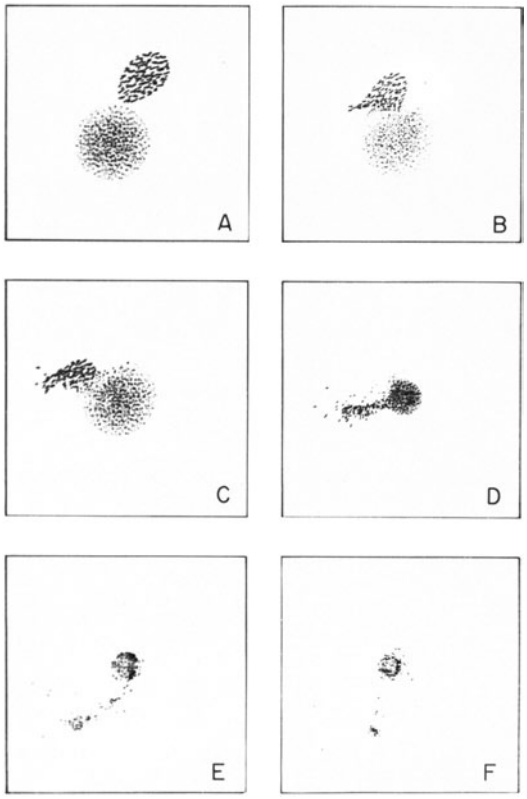


FIGURE 9 Sketch showing Mars-sized planetesimal obliquely impacting proto-Earth, along with trajectories taken by ejecta from the event. Time increases from sketch A to F. Total time shown is 30 min. [From Benz *et al.* (1986). Snapshots from a 3-D modeling of a giant impact. *In* "Origin of the Moon" (W. K. Hartmann, R. J. Phillips, and G. J. Taylor, eds.), pp. 617–620. Lunar and Planetary Institute, Houston, Texas.]

IV. THERMAL EVOLUTION OF PLANETARY INTERIORS

A. Introduction

Heat released through the surface from planetary interiors comes from three primary sources: heat left from initial accretion of the planet, heat formed as a consequence of internal differentiation (for example, forming Earth's metallic core), and heat produced during radioactive decay of uranium, thorium, and the isotope ^{40}K. Earth's surface heat flow from the interior is approximately 60 ergs/cm^2/s.

Models that track the thermal evolution of Earth, including heat of accretion, differentiation, and radioactive decay and both conductive and convective modes of heat transport, show that Earth would have had a higher heat flow early in geological time. It is the heat being lost to the surface that drives both tectonism and volcanism on Earth and on other solid bodies. Thus, an understanding of the tectonic and volcanic processes, deciphered from the geological record, provides first-order constraints on thermal evolution histories of planetary interiors, including temporal variations in rates of dissipation of internal heat.

In effect, tectonism and volcanism are boundary layer phenomena occurring at the interface between warmer planetary interiors and cooler vacuums or atmospheres above. On Earth, the boundary layer processes are dominated by plate tectonics, which account for about 65% of the heat flow. The upper mantle and overlying crust are relatively cool and rigid—this zone is called the lithosphere. Earth's lithosphere is broken into nine major plates. Boundaries between plates are of three types. Divergent boundaries are above upwelling convection currents in the underlying mantle. Upwelling generates elongate topographic highs and rifts as the mantle material decompresses during upwelling and increases in volume. Basaltic volcanism occurs as partial melts of the mantle and ascends through the rifts. The majority of divergent boundaries are in the oceanic crust in regions called spreading centers because the volcanic rocks become new oceanic crust that is then moved out to either side in a conveyor-belt fashion to be replaced by new volcanic rocks along the rift crest (Fig. 10). The new ocean crust cools as it moves away from the rift, eventually becoming denser than the underlying mantle, causing the lithosphere to sink back into the interior. This sinking, called subduction, occurs beneath the deep, curved trenches on the sea floor. The subduction zones represent a second type of plate boundary. Along with collisions of two continental masses, this boundary type is called convergent. The third type is a transform boundary and represents zones of strike–slip motion, such as the San Andreas fault system.

Plate tectonics has probably dominated the evolution of Earth's continental crust, zones of thicker, less dense rock of granitic composition that are too buoyant to be subducted. When continents collide or when subduction occurs at a continental margin, mountain building occurs, adding mass to the continental margins. However, because tectonism and volcanism are vigorous processes on Earth, much of the early continental crust has been destroyed or remobilized. Thus, reconstruction of early tectonic and volcanic processes, including whether or not plate tectonics operated on Earth, is difficult.

A major objective of planetary geology is to utilize the tectonic and volcanic rock records found on planetary bodies to probe how other objects have released their heat over a range of geological time scales. Do other objects have plate tectonics? If not, is heat primarily released by conduction, or does extensive volcanism transport most of the heat to the surface? What processes operated in earliest geological time? In this section, three examples, using tectonic and volcanic records to explore these issues, are considered: (1) evidence preserved in the lunar terrae for early separation of crust from mantle driven by the heat of accretion; (2) tectonic and volcanic processes on Venus; and (3) volcanic resurfacing of Io, powered by tidal heating from Jupiter.

B. Formation of the Lunar Terrae Crust

Figure 11 is a map of the Moon showing the locations of the heavily cratered terrae and the sparsely cratered maria. The terrae represent the surface exposure of an ~60-km-thick crust that formed 4.5 Ga. The maria consists of basalts that fill larger impact basins on the nearside, probably because the terrae crust is thinner on this side and

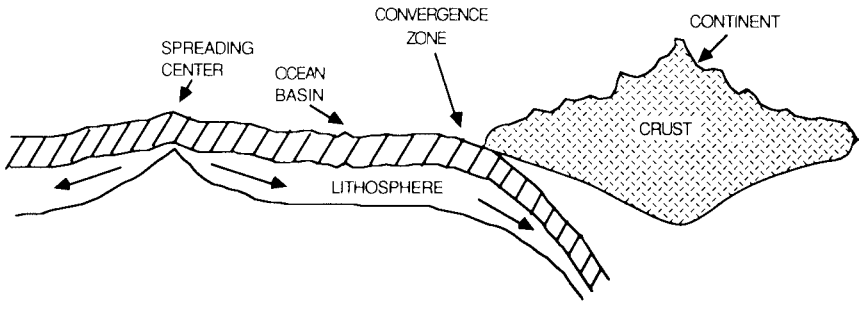


FIGURE 10 Schematic cross section of sea floor spreading center. New oceanic crust is created by volcanism associated with hot upwelling mantle at midocean ridges. The crust then moves out to either side at rates that vary from about 1 to 10 cm/year. The lithosphere is defined to be the crust and the colder, rigid part of the upper mantle. The lithosphere thickens away from ridge as crust and upper mantle cool, and eventually subduction occurs because of negative buoyancy. In the sketch, subduction is shown beneath the edge of a continent.

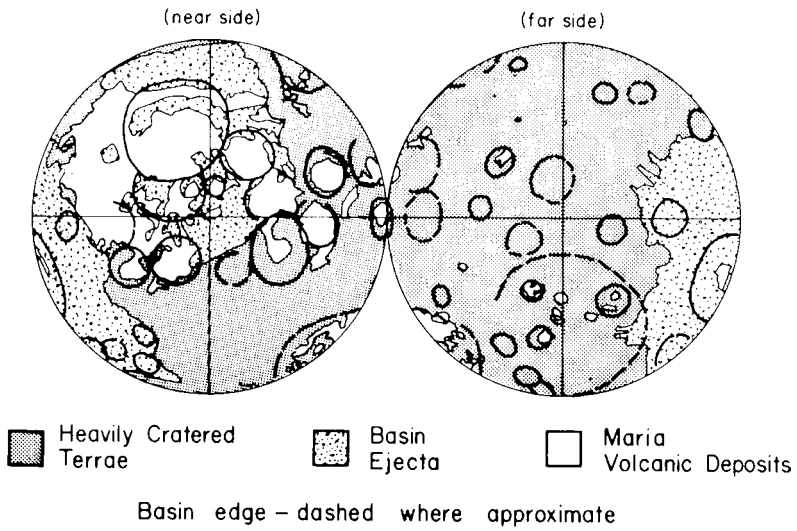


FIGURE 11 Simplified geologic map of the Moon, showing locations of the heavily cratered terrae, impact basins and ejecta, and maria volcanic deposits that fill larger basins on the Moon's Earth-facing or nearside. The basin on the lower right of the farside map with extensive ejecta is Orientale, the youngest of the lunar impact basins.

basaltic magmas generated in the lunar mantle were able to rise buoyantly only in regions with thin terrae crust. The maria basalts formed much later than the terrae crust, approximately 3.8 to 2.5 Ga, when radioactive heating of the lunar mantle led to partial melting and buoyant ascent of basaltic magmas. The maria deposits are several kilometers thick and occupy only a few percent of the crustal volume.

A number of isotopic observations point to separation of the lunar terrae crust and mantle during accretion of the body, approximately 4.5 to 4.6 Ga. For instance, a dunite fragment collected from a boulder at the *Apollo 17* site yielded a rubidium–strontium age of 4.6 Ga. Several other samples also have escaped the severe impact metamorphism associated with early torrential bombardment and date from this very early period. In addition, model isotopic ages for some lunar soils point to a major differentiation event sometime between 4.4 to 4.6 Ga. Niodimium–samarium data provide evidence that, in contrast to Earth, the lunar interior differentiated into a series of discrete source regions very early (about 4.6 Ga) in geologic time. The major differentiation event was probably associated with the formation of the terrae crust.

For a variety of reasons, accretional heating of the Moon is favored as the means of triggering the early separation of terrae crust from the mantle. Various calculations indicate that rapid accretion of the outer parts of the Moon or accretion by large bodies that bury heat deeply would have been sufficient to initiate an early global melting event, perhaps as an ocean of silicate magma. The formation of a global magma ocean is also a convenient means of allowing global differentiation into a crust and upper mantle. In particular, crystallization of rocks called ferroan anorthosites would lead to the formation of a floating outer shell or a series of rockbergs, since this rock type is slightly less dense than the residual magma it would leave behind. Cumulates of olivine and ilmenite would have settled to the bottom, perhaps forming the source regions that would partially melt between approximately 2.5 to 3.8 Ga to form the maria basalts. The residual liquid in the magma ocean would have had an olivine norite composition and formed a series of complex intrusives and other rock types in the terrae crust.

A global magma ocean has many redeeming attributes, although attempts to model even the major compositional trends of terrae rocks illustrate the many complexities associated with a presumably simple Moon-wide crust. Also, it is not clear if it is possible to maintain a silicate magma covering the Moon. For instance, vigorous convection within the heated zone may have served to keep the region from becoming completely fluid in the first place. Whatever the details of such an early heating epoch, it is clear that a major differentiation of crust and upper mantle took place. As noted in Section III.B, accretional modeling of the growth of Earth shows a similar pattern of extensive heat deposition, enough to separate Earth into a primitive crust, mantle, and core during latter stages of accretion.

C. Tectonism and Volcanism on Venus

Of the planets in the solar system, Venus is the closest to Earth in size and mass (Fig. 1). The area to mass ratio scales as the inverse of the body radius for spheroidal objects. Thus, relatively large and massive objects, such as

Earth and Venus, have relatively small ratios, in contrast to smaller objects such as the Moon or Mars. Since interior heat is produced within the planetary mass, whereas it is released through the surface area, it necessarily follows that larger objects will produce more heat and retain it over a longer time period compared to smaller objects. In fact, a plot of the fractional surface area covered by heavily cratered terrain (that is, primitive crust) versus the area to mass ratio of the planet follows a positive trend for solar system objects, with 100% coverage of cratered terrain on very small bodies, such as the Martian moons Phobos and Deimos, and none on Earth. By this argument, Venus should have a rich history of volcanism and tectonism extending over most, if not all, of geological time. This conclusion is predicted by numerical first-order considerations, such as the previous discussion, and by thermal evolution models, although a detailed understanding of the thermal history and consequent tectonism and volcanism requires thorough geological analyses of the composition, structure, and ages of surface materials.

The Venusian surface is obscured from observation by visible and infrared sensors because of sulfuric acid clouds and the dense CO_2 atmosphere. Longer wavelength microwaves penetrate the atmosphere without significant attenuation, and as a consequence, microwaves have been the primary wavelength region used to obtain information about the Venusian surface. The technique that has been used to understand the surface has been radar, from both Earth and orbital spacecraft. The Pioneer-Venus orbiter in 1979 to 1980 included a radar altimeter that provided a broadscale, global view of the distribution of elevations. The Russians used *Venera 15* and *16* in 1984 and 1985 as part of their extensive program of Venus missions that has included landers, balloons, and the two Venus orbiters. The Venera images cover regions from the north pole to approximate 35° N latitude (Figs. 12 and 13). The Venera radar system looked down with an angle tilted from the normal to the surface by approximately 10°. At this small incidence angle, radar returns are typically governed by specular reflections from slope facets, where the radii of curvature of the facets are much larger than the 8-cm Venera radar wavelength. Thus, the image brightness values shown in Figs. 12 and 13 are largely controlled by topography.

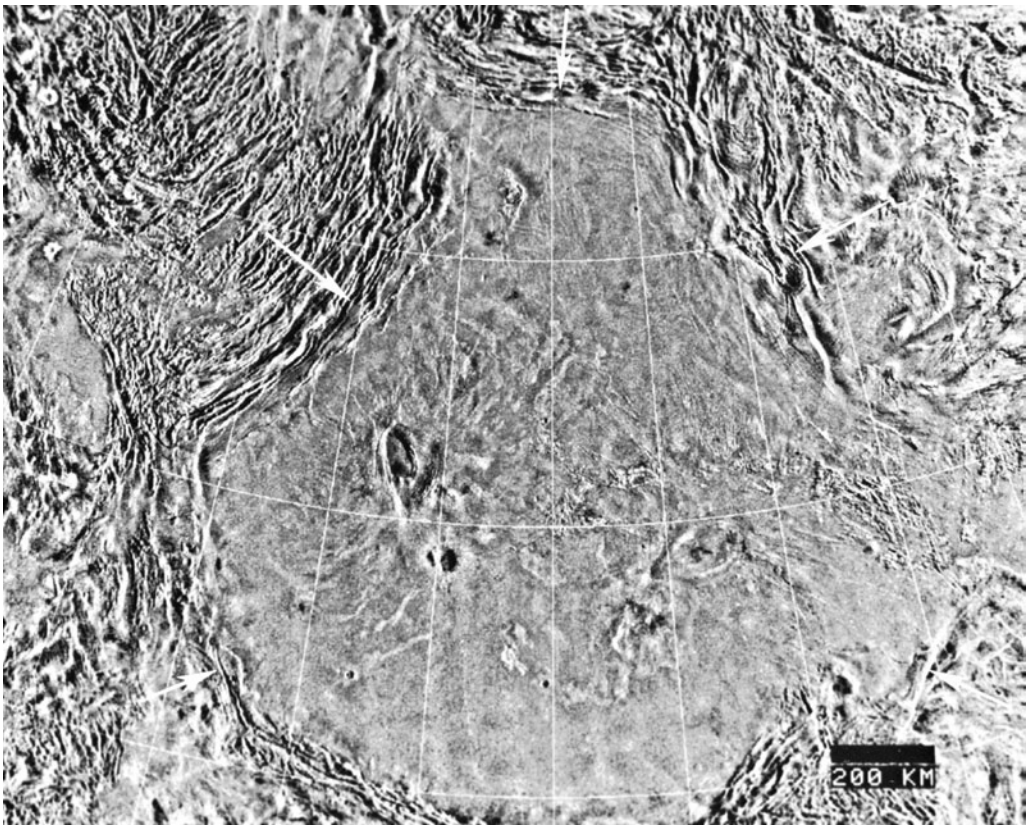


FIGURE 12 Mosaic of Venera radar images covering the plateau Lakshmi Planum and immediate surroundings. Arrows delineate edge of Lakshmi Planum. Structures expressed as curvilinear grooved and ridged terrain surrounding the plateau may be thrust faults and folds. Lines represent 5° in longitude and latitude. Radar illumination from left.

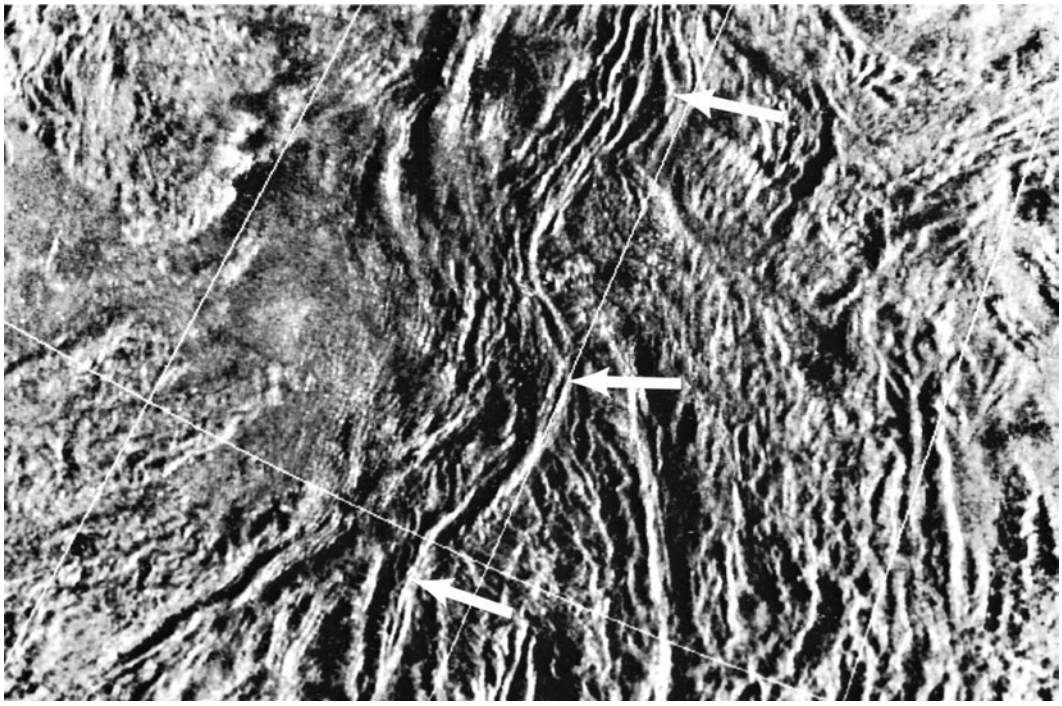


FIGURE 13 Mosaic of Venera radar images for region to northeast of Ishtar Terrae. Mosaic covers about 600 km in width. Arrows show probable surface expression of young faults and faults that have disrupted older structures. Lines are in 5° latitude and longitude intervals. Radar illumination from left.

Evidence can be found in the ensemble of Earth-based and spacecraft radar images of Venus for extension, compression, and strike–slip motions of the crust and lithosphere. For example, Beta Regio (latitude 15° to 40° N; longitude 60° to 85° E) is a 2500-km-long uplift that stands a couple of kilometeres above the surrounding rolling plains. Venera images of the northern segments of Beta Regio show a number of north–south oriented rifts arranged in an *en echelon* fashion. This pattern continues to the south, based on the examination of Earth-based Arecibo and Goldstone images. Further, Earth-based images show a pair of young volcanoes and numerous lava flows, based on analysis of the morphology and scattering properties inferred from the radar data. Thus, Beta Regio is an uplifted region with associated rifting and volcanism. The combined topography and gravity signature derived from Pioneer-Venus orbiter data for Beta Regio is consistent with dynamic support of the uplifted region caused by upwelling convection cells within the mantle, with fracturing and volcanism associated with the extension of the Venusian lithosphere at the apex of the uplift.

Figure 12 is a view from the Venera orbiters of the region known as Lakshmi Planum, a plateau standing several kilometers above surrounding plains. The structural patterns surrounding Lakshmi Planum suggest compressional forces have thrust segments of Venusian lithosphere over one another. Figure 13 is a close-up of a complexly faulted location to the northeast of Ishtar Terrae, again suggesting that compressional forces have been at work in the lithosphere. Other regions seen in the Venera data show evidence for strike–slip motion. Many areas are highly deformed, perhaps because of multiple deformation events. The abundance of impact craters is relatively low in all regions, with crater retention ages measured in only hundreds of millions of years. This age is comparable to the retention age for impact craters on Earth's continents and suggests that tectonism and volcanism is an ongoing process on Venus.

No other planetary body, other than Earth, seems to have as complex a tectonic and volcanic pattern as Venus. Although the limited coverage and resolution of current radar data have precluded understanding whether Venus supports a global plate tectonics regime, a number of hypotheses can be posed. For example, because of the high surface temperatures, the lithosphere may be no more than tens or so of kilometers in thickness and may be too buoyant to subduct. Thus, stresses associated with uplift and drag on the lithosphere may lead to extensive thin-skinned tectonics, but without the extensive sea floor spreading found on Earth. However, such a model is currently being debated, and it has been argued that Aphrodite Terrae in the southern hemisphere is a spreading center.

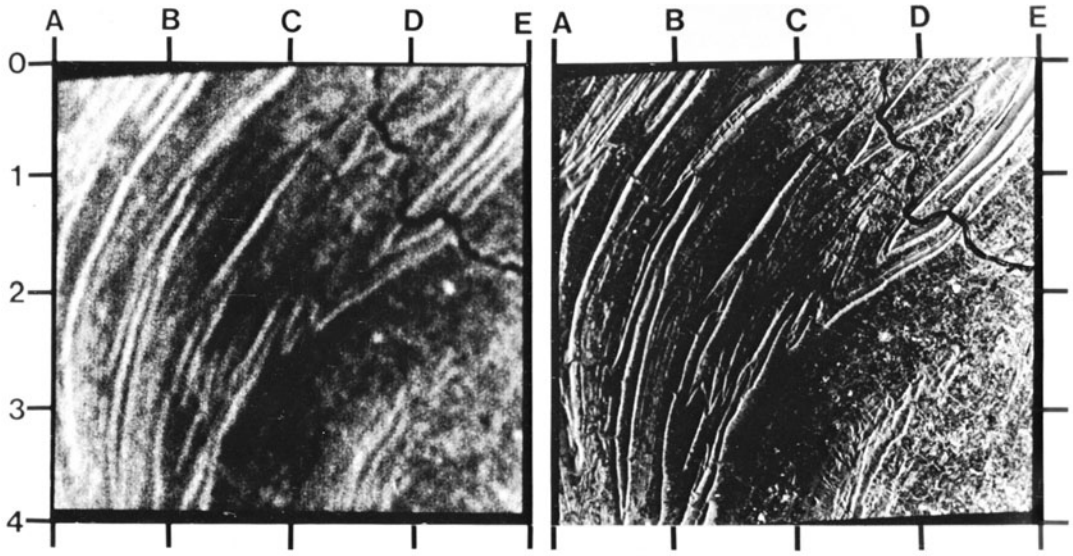


FIGURE 14 The Valley and Ridge Province, Appalachian Mountains, seen by the Seasat radar system, degraded to an image with Venera (left) and Magellan (right) system characteristics. [From Arvidson, R. E., *et al.* (1988). Construction and analysis of simulated Venera and Magellan images of Venus. *Icarus* **75,** 163–181.]

A more thorough understanding of the planet must await receipt of radar data from the Magellan mission. The spacecraft was placed into orbit about Venus and began mapping the planet in the summer of 1990. The Magellan mission acquired data over both the northern and southern hemisphere, with resolution an order of magnitude (Fig. 14). By the end of the mission in 1994, topography and gravity observations produced during the course of the Magellan mission was significantly improved. The detailed, global data bases from the Magellan mission allowed an in-depth understanding of how the interior of Earth's nearest planetary neighbor in location, mass, and size has evolved over time.

D. Resurfacing of the Galilean Satellites Europa and Io

Jupiter has a number of satellites. The four largest are named the Galilean satellites. These satellites were observed in some detail by both *Voyager 1* and *2*. The inner two, Io (closest to Jupiter) and Europa, show evidence for young tectonism and volcanism driven by tidal heating by Jupiter. Europa has the highest surface reflectance of any object in the solar system, reflecting 80% of the incident sunlight from its water-ice dominated crust. Since Europa is a small body with a large surface area to mass ratio, it should have lost most of its internal heat associated with accretion, differentiation, and radiative decay long ago. However, Voyager observations show that Europa is very smooth, largely devoid of impact craters, and transected by sets of linear (great and small circle) reddish zones (Fig. 15). The scarcity of craters implies a very young surface as compared to the surfaces of the outer two Galilean satellites Ganymede and Callisto. Earth-based spectral reflectance data indicate the surface of Europa, like the grooved terrain of Ganymede, is covered with relatively clean water ice, probably erupted as lava flows from water melted beneath an icy crust and upper mantle. The linear reddish zones are probably the surface expressions of rifts or strike–slip faults that have lost most of their topographic expression because of viscous relaxation. Three families of fractures can be discerned. The first set is aligned along roughly concentric circles centered about the middle of the hemisphere facing away from Jupiter. The second set

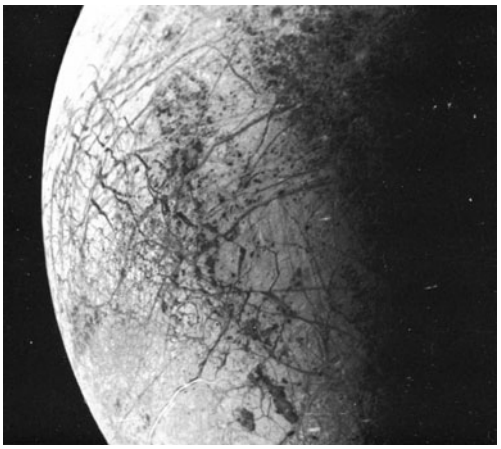


FIGURE 15 Voyager view of Galilean satellite Europa, showing a craterless surface of water ice cut by a number of dark fractures.

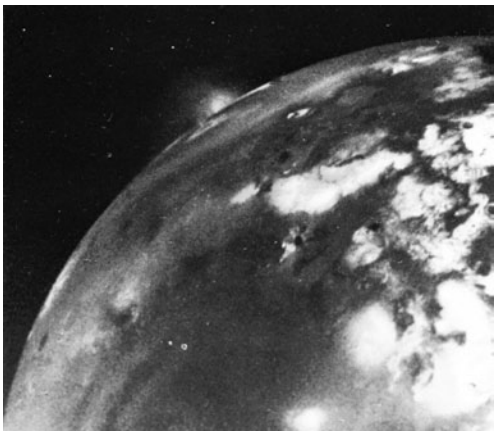


FIGURE 16 Voyager view of erupting volcano on Io.

forms conjugate pairs inclined at angles of about 30° to 60° from north. The third set forms a series of concentric cracks centered about an apparent pole. It appears that a principlal axis of deformation for the fractures is approximately along a line radial to Jupiter, suggesting that gravitational interactions with Jupiter may have dominated the satellite's tectonic and volcanic history.

Io has perhaps the most spectacularly active surface of any of the terrestrial bodies, with 8 active volcanic eruptions over the time span of the Voyager encounters (Fig. 16). Variations in the yellowish color of the surface, combined with Earth-based reflectance data, demonstrate that allotropes of sulfur, combined with SO_2, dominate the optical properties of Io. Approximately 200 calderas with diameters larger than 20 km across dot the surface, an abundance that is three times that found on Earth, normalized to surface area. Numerous long, low-relief lava flows emanate from a number of the calderas. Voyager thermal infrared observations of a dark area south of the Loki eruption center indicate that the dark material exhibited a 300-K brightness temperature. This temperature is 180 K higher than the temperature of the surrounding areas. A temperature of 500 K was measured over another of the eruptive sites, Pele. Other evidence for hot regions on Io can be found in Earth-based observations at the 5-μm wavelength, which indicate deviation from simple blackbody thermal behavior. These data are consistent with about 1% of the surface being covered with hot spots comparable in temperature to that measured for the volcanoes. In addition, thermal brightness outbursts seen in Earth-based thermal infrared observations of Io over the past few years are now interpreted as being caused by the eruptive activity.

The eruptions on Io hurl ash into space to altitudes of 70 to 280 km, values consistent with ballistic exit velocities of 0.6 to 1 km/s, approximately an order of magnitude higher than typical exit velocities for basaltic eruptive events on Earth. Based on the amount of material erupted during the Voyager encounters, a volcanic resurfacing rate of 10^{-2} to 10^{-4} cm/year has been computed. If such a rate has prevailed over a billion-year time scale, the crust and upper mantle of Io would be literally turned inside out.

Both Europa and Io are maintained in forced ecocentric orbits about Jupiter because of orbital interactions with the other Galilean satellites, Ganymede and Callisto. For example, the free eccentricity of Io would be about 10^{-5}, while orbital resonance calculations give a value of 4.3×10^{-3}. Tidal flexing of Europa and Io occurs as the satellites move through their eccentric orbital paths. The flexing heats the interiors of the two satellites and has been shown to dominate their thermal histories, maintaining active tectonism and volcanism long after the majority of heats of accretion, differentiation, and radioactive decay have dissipated. In fact, the model of tidal heating was published just before the *Voyager 1* Jupiter encounter. The spectacular eruptions on Io and the young icy surface of Europa graphically showed how both theory and observation complement one another in planetary geology.

V. CLIMATIC EVOLUTION OF MARS

A. Introduction

Mars has a carbon dioxide dominated atmosphere that ranges between 2 to 10 mbars in pressure at the surface. The average pressure is 6.1 mbars and the mean equatorial temperature is 220 K. There is abundant evidence for reservoirs of both carbon dioxide and water at and beneath the surface. Further, the climate seems to have changed over a variety of time scales. For example, channels that look like they were carved by water are ubiquitous in parts of the heavily cratered terrains (Fig. 17). In this section, evidence for the subsurface carbon dioxide and water reservoirs is first discussed and then evidence and models for past climates are high-lighted.

B. Evidence for Carbon Dioxide and Water Reservoirs

1. Evidence from Elemental and Isotopic Atmospheric Abundances

The abundance and isotopic composition of various elements in the Martian atmosphere were measured by mass spectrometers during the entry and landed phases of the Viking mission. These data provide a fairly detailed characterization of the present atmosphere and an indication of the total extent of volatile degassing during Martian

FIGURE 17 Viking orbiter image showing interconnected channels system in heavily cratered terrain on Mars. Image is located at about 48° south latitude, 98° west longitude and covers approximately 250 km in width.

history. The present atmosphere has carbon dioxide as the overwhelming constituent, with lesser quantities of N_2, Ar, Co, Ne, H_2O, O_2, Kr, and Xe. One of the most spectacular isotopic effects seen in the Mars data is in the value of the $^{15}N/^{14}N$ ratio, which is 1.7 times that found on Earth. The lighter ^{14}N has been preferentially lost from the atmosphere by mass-dependent processes, such as photoinduced fragmentation and thermal escape into the exosphere. The data indicate that perhaps as much as 10–150 times more than the current 0.13-mbar partial pressure of N_2 existed in the past. The mass-dependent stripping process would act on any polyatomic species, and it is therefore of interest that the carbon and oxygen isotopic ratios deviate by not more than 5% from the terrestrial values. This result implies that most of the carbon and oxygen has been kept in a reservoir that is immune to such processes. The reservoir must also exchange with the atmosphere on a fairly regular basis to maintain the observed isotopic normalcy. The most likely reservoir is CO_2 and H_2O ice buried beneath the surface and exposed at the permanent polar caps.

Viking measurements showed that the abundance of noble gases in the Martian atmosphere follows the typical planetary trend, but is depleted by nearly two orders of magnitude (per planetary mass) as compared to Earth. Martian xenon, as on Earth, is depleted relative to the planetary pattern found in ordinary chondrites. Conceivably, xenon may be preferentially trapped on dust grains as adsorbed species, as has been proposed to account for Earth's xenon depletion. There are a number of estimates of the column abundance (g/cm^2) of H_2O and CO_2 that have been degassed from the Martian interior, based on utilizing noble gas abundances as tracers of the extent of degassing. The estimates range from 290 to 500 g/cm^2 of CO_2 and 600 to 940 g/cm^2 of H_2O.

The N/C and H_2O/N_2 ratios for Mars, Earth, and Venus, together with a two-stage degassing history, have been used to model the extent of outgassing. A best fit to the Martian data predicts (1) an abundance of CO_2 between 2700 and 8000 g/cm^2, (2) an abundance of H_2O of 800 to 1600 g/cm^2, and (3) an outgassing history that was $\frac{1}{5}$ to $\frac{1}{20}$ as much as Earth's for the assumed early period and about $\frac{1}{5}$ as much as Earth's for the assumed later period.

Considering that the amount of CO_2 in the atmosphere is currently only about 23 g/cm^2 and the H_2O content is only about 0.0009 g/cm^2, the isotopic, noble gas abundance data and the degassing model just presented strongly indicate that most of the degassed CO_2 and H_2O are trapped at or beneath the surface as solids, as adsorbed or absorbed species, or chemically bound with surface materials. Loss by photodissociation processes, together with the amount of material found in the seasonal Martian polar caps, accounts for only a small fraction of the predicted amount of degassed material. Various lines of evidence that indicate the ways in which H_2O and CO_2 are stored on and beneath the surface, together with the volatile surface interactions that may have affected the history of Mars because of these reservoirs, are now considered.

2. Polar Caps and Layered Deposits

Beneath the thin (meters) seasonal polar CO_2 ice deposits is a thick deposit of dust and ice at both poles. Viking thermal infrared measurements showed that the ground temperature over the northern permanent cap during the

FIGURE 18 Viking orbiter image of the permanent north polar cap of Mars, which is approximately 1000-km wide and 4- to 6-km deep. It is composed of water ice and dust.

summer season, when the seasonal cap had evaporated, was about 210 K (Fig. 18). This temperature is far above the value that would exist if the volatile in the permanent deposits was CO_2 (148 K) or clathrates ($5.75H_2O \cdot CO_2$) (151 K) in equilibrium with the atmosphere. The only plausible volatile is water ice. The north polar cap covers about 8×10^5 km^2 and is thick enough to obscure the underlying topography. An average thickness of 100 m of H_2O would be the equivalent planetwide column abundance of 56 g/cm^2, while 3 km of ice thickness would be the equivalent of 1700 g/cm^3. Most likely, the amount of H_2O trapped in these deposits lies somewhere between these extreme limits. Better delineation of the amount is hampered by a lack of knowledge concerning the dust to ice ratio and the detailed topography of the deposits.

The composition of the south polar permanent cap deposits is uncertain since global dust storms obscured the southern hemisphere during the summer when the Viking measurements were conducted. The temperatures over the cap apparently never exceeded 160 to 180 K. The high albedo of the cap, even at those temperatures, would make it difficult for much CO_2 ice to evaporate. Thus, the southern polar deposits may consist of a thin (meters?) CO_2 ice cover over a thicker clathrate or water ice and dust deposit.

3. The Regolith

Soils imaged and sampled at the Viking landing sites are hydrated, oxidized weathering products. It is not known whether these soils formed by the interaction of magma and buried ice reservoirs to produce palagonites, by UV-induced photooxidation, or by some other process. It is significant that thick deposits of such materials could house, as chemically bound constituents, much of the degassed H_2O predicted in Section V.B.1. In fact, numerous regions on Mars can be found that appear to be underlain by ash or eolian deposits that are hundreds of meters thick. It is thus conceivable that the entire column abundance of CO_2 and H_2O predicted from noble gas data could be stored as chemically bound, adsorbed and absorbed species within the Martian regolith. For instance, 100 to 1000 g/cm^2 of CO_2 could be stored in adsorbed and absorbed form in a regolith of average thickness equal to 100 m, if the mineralogy were dominated by the iron-rich smectite clay, nontronite. In addition, appreciable amounts of CO_2 could be stored as carbonates.

4. Other Geological Evidence

Permanent deposits of water ice can exist in equilibrium with the atmosphere at depths below a few centimeters for latitudes poleward of 40°. Permanent reservoirs can also exist at the equator if the deposits are buried deep enough to inhibit diffusion and equilibration with the atmosphere. The depth of the 273-K isotherm within the crust is dependent on the assumed heat flow, but is probably about 1 km at the equator and somewhat more than 2 km at the poles. Thus, a substantial volume exists for storage of water ice.

Four major types of geologic features provide evidence that at least some of the surface is underlain by water-ice reservoirs. The features are (1) quasi-dendritic channel systems located in the ancient cratered terrain that indicate a warmer, wetter climate in early geologic time (see Section V.C); (2) large outflow channels that seem to have formed by high fluid discharge rates; (3) polygonally fractured ground in the higher northern latitudes, ranging in size from meters to kilometers in width; and (4) peculiar lobe-shaped ejecta deposits surrounding some craters.

The large channel systems on Mars are unique morphologic features that indicate formation by very high fluid discharge rates. A variety of forms can be discerned, but high water-discharge rates seem to be a needed ingredient in all of the large channels. Some of the larger channels are located in regional topographic lows. A considerable hydrostatic head (several kilometers) of water pressure would have existed in low areas, such as in the Chryse basin, if a regional groundwater system extended to the surrounding highlands. Pressures induced by such a system would have been sufficient to generate the 10^6 to 10^8 m^3/s discharge rates needed to carve the large channels. While other mechanisms have been proposed to carve the channels, such as the liquifaction of a water-soil mass or autosuspension of gravel within a water flow, the point remains that a considerable reservoir of water is needed.

Polygonally fractured terrain covers a portion of the higher latitude northern plains, with size ranges of meters per polygon at the *Viking 2* landing site to kilometers per polygon as seen from the Viking orbiters. The origin of these features has been ascribed to three processes: (1) cooling and extension fracturing of thick lava flows, (2) desiccation-induced fracturing of thick (∼1 km) deposits of clay minerals, and (3) freeze–thaw cycling of deep, water-saturated soil deposits. Two of the three explanations require the presence of a water reservoir.

Finally, the morphologies of ejecta deposits surrounding a number of Martian craters are unique, with lobe-shaped deposits extending radially outward from the crater rims. Simulation of these features by impact into layered, water-saturated soils produces strikingly similar deposits. The uniqueness of the deposits is thought to be due to impact into materials with a high content of ice. Steam or liquid released during the impact events would act as a carrier medium that lubricates or carries the ejecta as a ground-hugging debris flow. While such a hypothesis is still in the speculative state, the explanation is consistent with the other evidence for a considerable reservoir

of underground ice. In fact, considering the polar caps, volatiles stored in and on soil particles, and the evidence for underground reservoirs, it seems probable that the column abundances of degassed H_2O and CO_2 based on noble gas arguments may be lower bounds.

C. Current and Past Climates of Mars

The obliquity or inclination of the Martian spin axis is 24°, which compares closely to the 23.5° value for Earth's spin axis obliquity. Thus, Mars experiences seasons as does Earth. However, because the eccentricity of the Martian orbit is an order of magnitude higher than Earth's orbital eccentricity, seasons are much more asymmetric in terms of heating and length as compared to seasons on Earth. Specifically, the subsolar latitudes during the short Martian southern summer season receive 45% more incident radiation per unit time than do the subsolar latitudes during the northern summer. The reason is that the southern summer solstice occurs only 15° from the planet's perihelion position in its orbit. Earth-based telescopic observations over the past several decades show that major dust clouds tend to obscure the surface during the southern summer and early fall. Most of the storms begin in the south, roughly along subsolar latitudes, and the dust then migrates to become a planetwide shroud. On the other hand, northern summers are relatively dust free. It was within this seasonal asymmetry that the Viking landers conducted observations in the northern latitudes (*Viking 1* at 24°N; *Viking 2* at 48°N) for two Martian years for *Viking 2* and three years for *Viking 1*.

Observations conducted during the course of the Viking missions that are pertinent to monitoring atmosphere–surface interactions included (1) measuring atmospheric temperature, pressure, wind velocity, and wind direction; (2) six-channel (0.4 to 1.0 μm) imaging of the surface and sky; and (3) estimating atmospheric optical depths by imaging the Sun. Imaging of the surfaces from the Viking cameras within the first few weeks after landing demonstrated the presence of eolian deposits that ranged from centimeters to a meter or so in thickness and that indicated formation by winds blowing from north to south. Orbiter observations during the beginning of the southern summer (approximately 200 days after landing) showed the beginning of a number of dust storms at midsouthern latitudes. Tracking of the optical depths of the atmosphere from *Viking 1* showed that the dust passed over the landing sites within a few days of the start of the storm. After reaching a peak load, the dust content of the atmosphere began to decrease, decaying to prestorm opacities within a few tens of solar days. A second global major storm soon followed. Peak winds blew from north to south at the landing sites during these periods, although no wind velocities of sufficient magnitude to locally erode material were measured at the sites. Storms also occurred during the second and third years, although without the intensity of the first-year storms. In addition, local storms also occurred, including one that was imaged from the Viking orbiter while data were being acquired on the surface (Figs. 19 and 20).

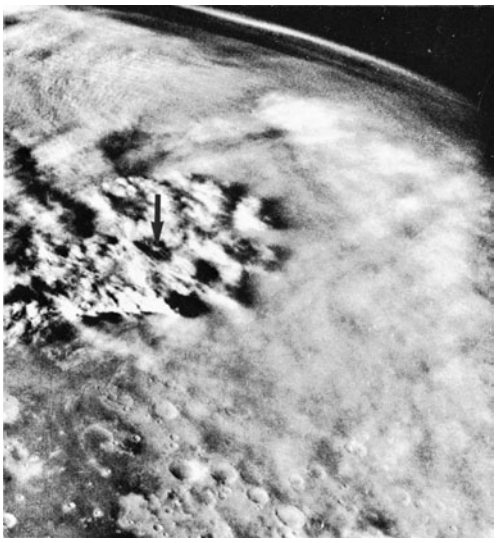


FIGURE 19 Viking orbiter image of a dust storm over *Viking 1* approximately 420 days after landing. The storm is about 500-km wide and moving toward the east. Season is late northern winter. Arrow points to approximate landing site.

The progress of the two first-year dust storms was obscured at the *Viking 2* site because of the effects of ice condensates on the optical properties of the atmosphere. During the height of the second storm, which occurred close in time to the northern winter solstice, a grayish surface condensate began to form at the *Viking 2* site, eventually reaching a thickness of a few tens of micrometers before evaporating back into the atmosphere within about 120 days. High temperatures observed during midday imply that the condensate was dominated by water ice. The color and albedo of the soils at the *Viking 2* site were markedly different after the condensate evaporated. The surface was brighter, redder, and closer in optical properties to the red deposits of dust that earlier only covered parts of the site. The *Viking 1* site also appeared to brighten and redden, but by no means as dramatically as the *Viking 2* site. The condensates at the *Viking 2* site may have formed as a fall of ice particles, cored by dust particles. Since the settling velocity of particles scales as the square of the particle diameter, this process may be an effective way of cleansing the atmosphere of dust, especially at higher latitudes. The bright dust accumulations were removed in part by high winds during the northern summer seasons.

FIGURE 20 *Viking 1* image taken during dust storm shown in Fig. 19. Sediments have been deposited on the lander's deck by the surface sampler. Note that the horizon can be seen, so the total amount of dust suspended in the atmosphere must be relatively small, only a fraction of a micrometer in equivalent thickness.

The Viking observations suggest an effective way of accumulating the layered deposits that are part of the permanent polar ice caps at both Martian poles. That is, they may have formed by a fall of ice and dust during a storm. However, the fact that the polar deposits are layered indicates that such a process must vary in intensity with time. Also, the fact that the layers can be seen means that they have been partially eroded and, consequently, that any depositional process must be interspersed with periods of erosion or that some other mechanism of landform evolution has also been active.

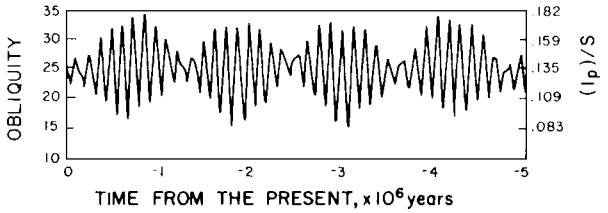


FIGURE 21 Sketch showing variations in obliquity of the Martian spin axis (left axis) and resultant variation in average annual polar insolation [normalized to the solar constant (right axis)]. [From Ward *et al.* (1974). Climatic variation on Mars: 1. Astronomical theory of insolation. *J. Geophys. Res.* **79,** 3375–3386.]

The probable mechanism for modulating the extent of dust in the atmosphere and the accumulation of dust and ice in the layered deposits is the variation in the orbital characteristics of Mars due to gravitational interaction with the Sun and other planets. The orbital eccentricity, the obliquity, and the orientation of the spin axis relative to a star-fixed coordinate system all change. These variations in turn lead to major changes in the latitudinal and seasonal distribution of solar insolation (Fig. 21). The patterns of obliquity and eccentricity oscillations are much larger than those predicted for Earth, with the obliquity oscillating between 15° to 35° and the eccentricity moving from 0.001 to 0.14. Changes in the latitudinal distribution of insolation are primarily dependent on the obliquity variations. During periods of high obliquity, increased insolation at the poles would tend to raise surface temperatures and lead to evaporation of carbon dioxide and water ice. On the other hand, during periods of low obliquity, polar surface temperatures should decrease and permanent caps composed of carbon dioxide and water ice should form.

It appears that the overall Martian atmospheric pressure is buffered by an atmospheric–cap–regolith system because of the possibility of adsorption and absorption of gaseous species onto regolith particles. Considering the three component system, variations are predicted in atmospheric pressure between about $\frac{1}{10}$ mbar periods of low obliquity to 20 to 30 mbar during periods of high obliquity. Since the wind shear stress needed to erode particles scales as the inverse of the square root of atmospheric density, it seems probable that eolian activity would be high during periods of high obliquity and essentially nonexistent during periods of low obliquity. It also seems probable that any hemispheric asymmetry in polar cap composition must be related to changes in eccentricity. As an example, during periods of low eccentricity, there would not be a preferential season for dust storms. Thus, the two polar caps would tend toward a uniform composition.

Although no model has been developed that explains in detail the erosional and depositional patterns in the polar layered deposits and permanent ice caps, it seems probable

that these first-order orbital changes are the forcing functions underlying the origins of these deposits. The variatoins would have been even more enhanced during the period of time before the Tharsis Plateau, when obliquity variations are predicted to have reached values as high as 45°.

As noted in Section V.B, the smaller, interconnected channel systems on Mars are found preferentially in the more heavily cratered parts of the cratered terrain (Fig. 17). Because these features look like drainage networks and because they are only found in the oldest geologic units, it is tempting to ascribe their origin to rainfall and runoff during an earlier period in Martian history. Conditions then may have been warmer and thus the partial pressure of water may have been sufficiently high to allow the liquid phase to be stable. Hypotheses have been invoked that suggest an early atmosphere, containing CH_4 and NH_3, would have been opaque enough in the thermal infrared to provide the requisite environment through greenhouse effects. This topic is an area of ongoing debate. For example, the channel system junction angles (angles between tributary and main channel) deviate significantly from the values predicted based on runoff patterns for channel networks on Earth. This observation has led some workers to invoke origins of Martian channel networks based on groundwater sapping and scarp retreat. Testing of these hypothesis will require further exploration of Mars, together with collection and analyses of a variety of data sets.

VI. THE FUTURE OF PLANETARY GEOLOGY

In the 1990s, the United States committed to a wide-ranging reconnaissance of the solar system involving orbital spacecraft. As discussed in Section IV.C, the Magellan spacecraft was placed into orbit around Venus in the summer of 1990 and acquired detailed radar images and altimetry and gravity data over a number of 243-day mapping cycles. *Mars Observer*, launched in 1992, was an orbiter designed to track the meteorology of Mars in detail, to better understand the complicated climatic regime that exists at present. In addition, the orbiter provided the first detailed maps of the composition of materials exposed at the surface and measured the magnetic and gravitational field characteristics of the planet, along with the topography of the surface. A camera acquired synoptic and very high resolution images of the surfaces, including sites for future missions (Fig. 22).

The Galileo mission went to the Jovian system in 1995 and consisted of both a probe into Jupiter's atmosphere and an orbiter to observe Jupiter and its system of satellites. It is expected that the outer solar system reconnaissance

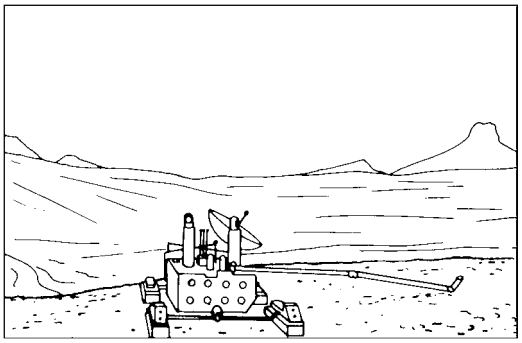


FIGURE 22 Sketch showing Mars rover surveying surrounding deposits. The rover would travel a number of kilometers, select and acquire samples, and deliver them to an ascent vehicle for return to Earth.

will be completed with the approval and implementation of the CRAF/Cassini mission. CRAF stands for comet rendezvous asteroid flyby and would be a *Mariner-Mark II* sent to acquire data during a close flyby of an asteroid, on its way for an encounter with a comet. The spacecraft will fly with the comet and observe it as the comet approaches the Sun and forms a dust and plasma tail. Cassini is a Saturn mission to complement the Galileo mission to Jupiter, including an orbiter to observe the Saturnian system and an atmospheric probe for the Saturnian satellite Titan.

A lunar geoscience orbiter is being planned that would be used to map the kinds of minerals found on the lunar surface, including searching for ices that may have been trapped in the eternal darkness of craters at the lunar poles. Measurements of the interior will also be conducted.

In contrast to the broad reconnaissance of the solar system planned by the United States and its allies, Russia will focus on Mars. They have launched *Phobos I* and *II* to Mars, although communications have been lost with *Phobos I*. *Phobos II* was placed in orbit around Mars in January 1989, but contact was lost with the spacecraft in late March 1989. Russia launched a pair of spacecraft, called the Mars 94 mission, which included an orbital component, instrumented balloons, and perhaps surface rovers. Discussions are underway to utilize *Mars Observer* to transmit data acquired from the Soviet 94 balloons back to Earth. The Russians also planned to return Martian samples to Earth, probably by the end of the 1990s. The United States also planned a rover and sample return mission and there were discussions for a joint United States–Russian mission.

The future of planetary geology will be a rapid increase in data volumes and complexity (Fig. 5) for a number of solar system objects. These data will provide the grist needed for the discipline to continue to mature and address global-scale geological problems associated with the formation and evolution of planetary-scale bodies. Finally, existing

and future data will further demonstrate that Earth is but one member of a suite of planetary-scale objects. Understanding the origin, thermal evolution, and climates of these bodies will increase our understanding of the entire collection, including Earth.

SEE ALSO THE FOLLOWING ARTICLES

ASTEROID IMPACTS AND EXTINCTIONS • CHEMICAL COMPOSITION AND ELEMENT DISTRIBUTION IN THE EARTH'S CRUST • EARTH SCIENCES, HISTORY OF • GEOMORPHOLOGY • IMPACT CRATERING • LUNAR ROCKS • MOON (ASTRONOMY) • PLANETARY ATMOSPHERES • PLANETARY SATELLITES, NATURAL • PLATE TECTONICS • SOLAR SYSTEM, GENERAL

BIBLIOGRAPHY

Carr, M. H. (ed.) (1984). "The Geology of the Terrestrial Planets," NASA Special Publ. 469. U.S. Govt. Printing Office, Washington, D.C.

Clifford, S. M., Greeley, R., and Haberle, R. M. (1988). NASA Mars project: Evolution of climate and atmosphere. *EOS, Trans. Am. Geophys. Union* **69,** 1585–1596.

Hartmann, W. K., Phillips, R. J., and Taylor, G. J. (eds.) (1986). "Origin of the Moon," Lunar and Planetary Institute, Houston, Texas.

NASA Solar System Exploration Committee (1986). "Planetary Exploration Through the Year 2000, An Augmented Program," Report of NASA's Solar System Exploration Committee. U.S. Govt. Printing Office, Washington, D.C.

Spohn, T. (1991). *ICARUS* **90**(2), 222–236.

Stone, E. C., and Miner, E. D. (1979). Voyager 1 encounter with the Jovian system. *Science* **204,** 945–948.

Stone, E. C., and Miner, E. D. (1986). Voyager 2 encounter with the Uranian system. *Science* **233,** 39–43.

Planetary Satellites, Natural

Bonnie J. Buratti
California Institute of Technology

GLOSSARY

Bond albedo Fraction of the total incident radiation reflected by a planet or satellite.
Carbonaceous material Carbon–silicate material rich in simple organic compounds. It exists on the surfaces of several satellites.
Differentiation Melting and chemical fractionation of a planet or satellite into a core and mantle.
Geometric albedo Ratio of the brightness at a phase angle of zero degrees (full illumination) compared with a diffuse, perfectly reflecting disk of the same size.
Greenhouse effect Heating of the lower atmosphere of a planet or satellite by the transmission of visible radiation and subsequent trapping of reradiated infrared radiation.
Lagrange points Five equilibrium points in the orbit of a satellite around its primary. Two of them (L4 and L5) are points of stability for a third body.
Magnetosphere Region around a planet dominated by its magnetic field and associated charged particles.
Opposition effect Surge in brightness as a satellite becomes fully illuminated to the observer.
Phase angle Angle between the observer, the satellite, and the Sun.
Phase integral Integrated value of the function which describes the directional scattering properties of a surface.
Primary body Celestial body (usually a planet) around which a satellite, or secondary, orbits.
Regolith Surface layer of rocky debris created by meteorite impacts.
Roche's limit Distance (equal to 2.44 times the radius of the primary) at which the tidal forces exerted by the primary on the satellite equal the internal gravitational forces of the satellite.
Synchronous rotation Dynamical state caused by tidal interactions in which the satellite presents the same face towards the primary.

A NATURAL planetary satellite is a celestial body in orbit around one of the nine principal planets of the solar system. The central body is known as the primary and the orbiting satellite its moon or secondary. Among the nine planets only Mercury and Venus have no known companions. There are 54 known natural planetary satellites in the solar system; there may exist may more undiscovered satellites, particularly small objects encircling the giant outer planets. The satellites range in size from planet-sized objects such as Ganymede and Titan to tiny, irregular bodies tens of kilometers in diameter (see Table I and Fig. 1).

TABLE I Summary of the Properties of the Natural Planetary Satellites

Satellite	Distance from primary (10^3 km)	Revolution period (days) R = retrograde	Orbital eccentricity	Orbital inclination (degrees)	Radius (km)	Density (gm/cm^3)	Visual geometric albedo	Discoverer	Year of discovery
Earth									
Moon	384.4	27.3	0.055	18 to 29	1738	3.34	0.11		
Mars									
M1 Phobos	9.38	0.32	0.018	1.0	14×10	1.9	0.05	Hall	1877
M2 Diemos	23.50	1.26	0.002	2.8	8×6	2.1	0.05	Hall	1877
Jupiter									
J15 Adrastea	128	0.30	0.0	0.0	10		<0.1	Jewitt *et al.*	1979
J16 Metis	129	0.30	0.0	0.0	20		<0.1	Synnott	1979
J5 Amalthea	181	0.49	0.003	0.4	$131 \times 73 \times 67$		0.05	Barnard	1892
J14 Thebe	222	0.67	0.015	0.8	50		<0.1	Synnott	1979
J1 lo	422	1.77	0.004	.04	1,818	3.53	0.6	Galileo	1610
J2 Europa	671	3.55	0.010	0.5	1,560	2.99	0.6	Galileo	1610
J3 Ganymede	1,070	7.15	0.002	0.2	2,634	1.94	0.4	Galileo	1610
J4 Callisto	1,883	16.69	0.007	0.5	2,409	1.85	0.2	Galileo	1610
J18 1975 J1 (also 2000 J1)	7,507	130	0.242	43	~4			Kowal and Roemer	1975
J13 Leda	11,094	239	0.148	26.7	5			Kowal	1974
J6 Himalia	11,480	251	0.163	27.6	85		0.03	Perrine	1904
J10 Lysithea	11,720	259	0.107	29.0	12			Nicholson	1938
J7 Elara	11,737	260	0.207	24.8	40		0.03	Perine	1904
J28 2000 J11	12,557	287	0.248	28	~2			Sheppard *et al.*	2000
J27 2000 J10	22,988	716R	0.159	166	~2			Sheppard *et al.*	2000
J20 2000 J3	20,210	585R	0.218	150	~3			Sheppard *et al.*	2000
J22 2000 J5	21,132	625R	0.227	149	~2			Sheppard *et al.*	2000
J24 2000 J7	20,929	629R	0.219	149	~3			Sheppard *et al.*	2000
J12 Ananke	21,200	631R	0.17	147	10			Nicholson	1951
J26 2000 J9	23,140	722R	0.252	165	~2			Sheppard *et al.*	2000
J21 2000 J4	23,169	723R	0.270	165	~2			Sheppard *et al.*	2000
J11 Carme	22,600	692R	0.21	163	15			Nicholson	1938
J23 2000 J6	23,074	719R	0.261	165	~2			Sheppard *et al.*	2000
J8 Pasiphae	23,500	735R	0.38	145	18			Melotte	1908
J25 2000 J8	23,913	758R	0.426	153	~3			Sheppard *et al.*	2000
J9 Sinope	23,700	758R	0.28	153	14			Nicholson	1914
J17 1999 J1	24,103	759R	0.282	147	~2			Scotti *et al.*	1999
J19 2000 J2	23,746	751R	0.243	165	~3			Sheppard *et al.*	2000
Saturn									
S18 Pan	134	0.57	0.0	0.0	10	—	—	Showalter	1990
S15 Atlas	138	0.60	0.000	0.0	$19 \times 17 \times 14$		0.4	*Voyager*	1980
S16 Prometheus	139	0.61	0.002	0.0	$74 \times 50 \times 34$		0.6	*Voyager*	1980
S17 Pandora	142	0.63	0.004	0.05	$55 \times 44 \times 31$		0.6	*Voyager*	1980
S10 Janus	151	0.69	0.007	0.14	$97 \times 95 \times 77$	0.65	0.6	Dollfus	1966
S11 Epimetheus	151	0.69	0.009	0.34	$69 \times 55 \times 55$	0.65	0.5	Fountain and Larson	1978
S1 Mimas	186	0.94	0.020	1.5	199	1.4	0.8	Herschel	1789
S2 Enceladus	238	1.37	0.004	0.0	249	1.2	1.0	Herschel	1789

continues

TABLE I (*Continued*)

Satellite	Distance from primary (10^3 km)	Revolution period (days) R = retrograde	Orbital eccentricity	Orbital inclination (degrees)	Radius (km)	Density (gm/cm^3)	Visual geometric albedo	Discoverer	Year of discovery
S3 Tethys	295	1.89	0.000	1.1	530	1.2	0.8	Cassini	1684
S14 Calypso	295	1.89	0.0	1.1	15 × 8 × 8		0.6	Pascu *et al.*	1980
S13 Telesto	295	1.89	0.0	1.0	15 × 12 × 8		0.9	Smith *et al.*	1980
S4 Dione	377	2.74	0.002	0.02	560	1.4	0.55	Cassini	1684
S12 Helene	377	2.74	0.005	0.15	16		0.5	Laques and Lecacheux	1980
S5 Rhea	527	4.52	0.001	0.35	764	1.3	0.65	Cassini	1672
S6 Titan	1,222	15.94	0.029	0.33	2,575	1.88	0.2	Huygens	1655
S7 Hyperion	1,481	21.28	0.104	0.4	180 × 140 × 112		0.3	Bond and Lassell	1848
S8 Iapetus	3,561	79.33	0.028	14.7	718	1.2	0.4–0.08	Cassini	1671
S9 Phoebe	12,952	550.4R	0.163	150	110		0.06	Pickering	1898
Uranus									
U6 Cordelia	49.7	0.33	0.0005	0.14	13			*Voyager 2*	1986
U7 Ophelia	53.8	0.38	0.010	0.09	15			*Voyager 2*	1986
U8 Bianca	59.2	0.43	0.001	0.16	21			*Voyager 2*	1986
U9 Cressida	61.8	0.46	0.0002	0.04	31		∼0.04	*Voyager 2*	1986
U10 Desdemona	62.7	0.47	0.0002	0.16	27		∼0.04	*Voyager 2*	1986
U11 Juliet	64.4	0.49	0.0006	0.06	42		∼0.06	*Voyager 2*	1986
U12 Portia	66.1	0.51	0.0002	0.09	54		∼0.09	*Voyager 2*	1986
U13 Rosalind	69.9	0.56	0.00009	0.28	27		∼0.04	*Voyager 2*	1986
U14 Belinda	75.3	0.62	0.0001	0.03	33			*Voyager 2*	1986
U15 Puck	86.0	0.76	0.00005	0.31	77		0.07	*Voyager 2*	1985
U5 Miranda	130	1.41	0.003	3.4	236	1.2	0.35	Kuiper	1948
U1 Ariel	191	2.52	0.003	0.0	579	1.6	0.36	Lassell	1851
U2 Umbriel	266	4.14	0.005	0.0	585	1.5	0.20	Lassell	1851
U3 Titania	436	8.71	0.002	0.0	789	1.7	0.30	Herschel	1787
U4 Oberon	583	13.46	0.001	0.0	761	1.6	0.22	Herschel	1787
U16 Caliban	7,169	580R	0.08	137.6	∼20		0.07	Gladman *et al.*	1997
U20 Stephano	7,948	678R			∼8			Gladman *et al.*	1999
U17 Sycorex	12,213	1290R	0.5	97.8	∼40		0.07	Nickolson *et al.*	1999
U18 Prospero	16,568	2039R			∼10			Holman *et al.*	1993
U19 Setebos	17,681	2248R			∼10			Kavelaars *et al.*	1999
Neptune									
N8 Naiad	48.2	0.29	0.000	0.0	29			*Voyager 2*	1989
N7 Thalassa	50.1	0.31	0.0002	4.5	40			*Voyager 2*	1989
N5 Despina	52.5	0.33	0.0001	0.0	74		0.05	*Voyager 2*	1989
N6 Galatea	62.0	0.43	0.0001	0.0	79			*Voyager 2*	1989
N4 Larissa	73.6	0.55	0.000	0.0	104 × 89		0.06	*Voyager 2*	1989
N3 Proteus	117.6	1.12	0.0004	0.0	208		0.06	*Voyager 2*	1989
N1 Triton	354.8	5.87R	0.000015	157	1,353	2.08	0.73	Lassell	1846
N2 Nereid	5,513	360.1	0.751	29	170		0.14	Kuiper	1949
Pluto									
P1 Charon	19.4	6.39R	0	0	586	2.24	0.40	Christy	1978

Note. Twelve additional Satellites of Saturn, S19–S30 (2000 S1 to 2000 S12) were discovered by Gladman *et al.* in late 2000.

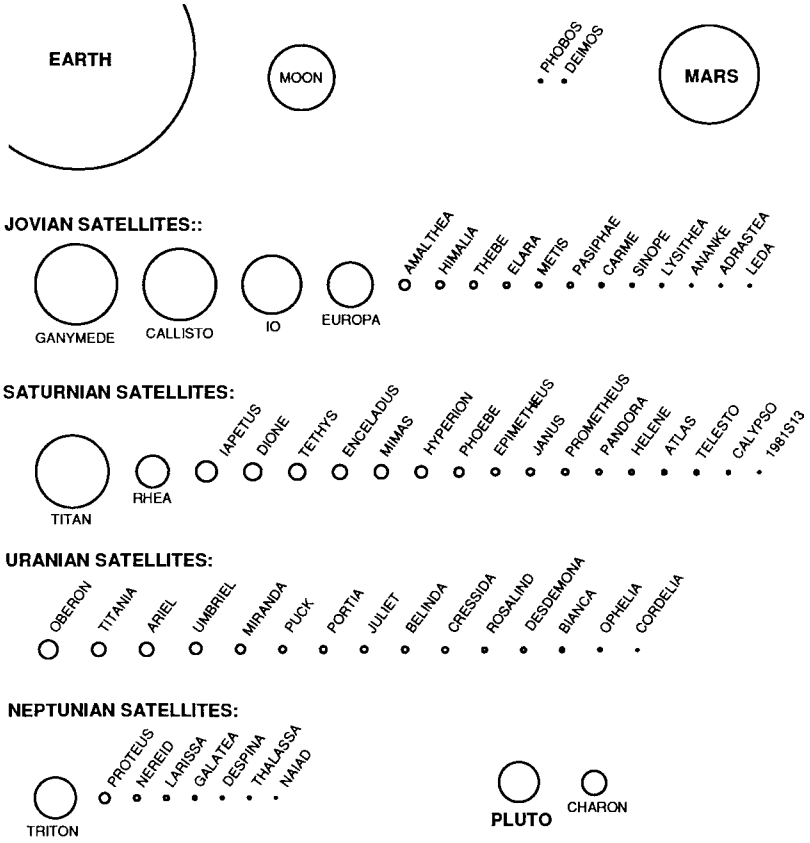


FIGURE 1 Relative shapes and sizes of the largest natural planetary satellites. For comparison, the relative sizes of smaller planets are shown. Jupiter and Saturn would span about twice the height of the figure and Uranus and Neptune would be about two-thirds as wide as the figure.

I. SUMMARY OF CHARACTERISTICS

A. Discovery

The only natural planetary satellite known before the advent of the telescope was the Earth's moon. Phenomena such as the lunar phases and the ocean tides have been studied for centuries. When Galileo turned his telescope to Jupiter in 1610, he discovered the four large satellites in the Jovian system. His observations of their orbital motion around Jupiter in a manner analogous to the motion of the planets around the sun provided critical evidence for the acceptance of the heliocentric (sun-centered) model of the solar system. These four moons—Io, Europa, Ganymede, and Callisto—are sometimes called the Galilean satellites.

In 1655 Christian Huygens discovered Titan, the giant satellite of Saturn. Later in the 17th century, Giovanni Cassini discovered the four next largest satellites of Saturn. It was not until over 100 years later that the next satellite discoveries were made: the Uranian satellites. Titania and Oberon and two smaller moons of Saturn. As telescopes acquired more resolving power in the 19th century, the family of satellites grew (see Table I). Many small satellites of Jupiter and Saturn were discovered during flybys of the *Pioneer* and *Voyager* spacecraft. Recent discoveries were made by highly sensitive electronic detectors known as charged coupled devices (CCDS) (see Table II).

The natural planetary satellites are generally named after figures in classical Greek and Roman mythology who were associated with the namesakes of their primaries. They are also designated by the first letter of their primary and an Arabic numeral assigned in order of discovery: Io is J1, Europa, J2, and so on. When satellites are first discovered but not yet confirmed or officially named, they are known by the year in which they were discovered, the initial of the primary, and a number assigned consecutively, e.g., 1980J27.

When planetary scientists were able to observe and map geologic formations of the satellites from spacecraft images, they named many of the features after characters or locations from Western and Eastern mythologies.

TABLE II Summary of Major Missions to the Planetary Satellites

Mission name	Object	Encounter dates	Type of mission
Luna 3 (Russia)	Moon	1959	Flyby (far side)
Ranger 7, 8, 9	Moon	1964–1965	Crash landing; image return
Luna 9, 13 (Russia)	Moon	1966	Soft landing
Luna 10, 12, 14 (Russia)	Moon	1966–1968	Orbiter
Surveyor 1, 3, 5, 6, 7	Moon	1966–1968	Soft landing
Lunar Orbiter 1–5	Moon	1966–1968	Orbiter
Apollo 7–10	Moon	1968–1969	Manned orbiter
Apollo 11, 12, 14–17	Moon	1969–1975	Manned landing
Luna 16, 20 (Russia)	Moon	1970–1972	Sample return
Luna 17, 21, 24 (Russia)	Moon	1970–1976	Rover
Mariner 9	Deimos	1971	Orbiter
	Phobos		
Pioneer 10	Jovian satellites	1979	Flyby
Pioneer 11	Jovian satellites	1979	Flyby
	Saturnian satellites	1979	Flyby
Viking 1,2	Phobos	1976	Orbiter
	Deimos		
Voyager 1	Jovian satellites	1979	Flyby
	Saturnian satellites	1980	Flyby
Voyager 2	Jovian satellites	1979	Flyby
	Saturnian satellites	1981	Flyby
	Uranian satellites	1986	Flyby
	Neptunian satellites	1989	Flyby
Phobos 2 (Russia)	Phobos	1989	Flyby
Clementine	Moon	1994	Orbiter
Galileo	Jovian satellites	1995–2002	Orbiter
Lunar prospector	Moon	1998	Orbiter
Mars global surveyor	Phobos	1998	Orbiter
Cassini	Saturnian satellites	2004–2008	Orbiter

B. Physical and Dynamic Properties

The motion of a satellite around the center of mass of itself and its primary defines an ellipse with the primary at one of the foci. The orbit is defined by three primary orbital elements: (1) the semimajor axis, (2) the eccentricity, and (3) the angle made by the intersection of the plane of the orbit and the plane of the primary's spin equator (the angle of inclination). The orbits are said to be regular if they are in the same sense of direction (the prograde sense) as that determined by the rotation of the primary. The orbit of a satellite is irregular if its motion is in the opposite (or retrograde) sense of direction or if it has a high angle of inclination. The majority of satellites move in regular, prograde orbits. Those satellites that do not move in regular, prograde orbits are believed to be captured objects (see Table I).

Except for those that are small and irregular, planetary satellites present the same hemisphere toward their primaries, which is the result of tidal evolution. When two celestial bodies oroit each other, the gravitational force exerted on the near side is greater than that exerted on the far side. The result is an elongation of each body to form tidal bulges, which can consist of solid, liquid, or gaseous (atmospheric) material. The primary will tug on the satellite's tidal bulge to lock its longest axis onto the primary-satellite line. The satellite, which is said to be in a state of synchronous rotation, keeps the same face toward the primary. Since this despun state occurs rapidly (usually within a few million years), most natural satellites are in synchronous rotation.

The natural satellites are unique worlds, each representing a vast panorama of physical processes. The small satellites of Jupiter and Saturn are irregular chunks of ice and rock, perhaps captured asteroids, which have been subjected to intense meteoritic bombardment. Several of the satellites, including Phoebe (which is in orbit around Saturn) and the Martian moon Phobos, are covered with dark carbonaceous material believed to be representative

of the primordial, unprocessed material from which the Solar System formed. The medium-sized satellites of Saturn and Uranus are large enough to have undergone internal melting and subsequent differentiation and resurfacing. The Saturnian satellite, Iapetus, presents a particular enigma: one hemisphere is 10 times more reflective than the other. Three of the Galilean satellites show evidence of geologically active periods in their history; Io is presently undergoing intense volcanic activity. The Earth's Moon experienced a period of intense meteoritic bombardment and melting soon after its formation.

Saturn's largest satellite, Titan, has a predominantly nitrogen atmosphere thicker than that of the Earth. Triton, the large satellite of Neptune, also has an appreciable atmosphere. Io has a thin, possibly transient sulfur dioxide atmosphere that is thought to be related to outgassing from active volcanoes. None of the other satellites has a significant atmosphere, although tenuous molecular atmospheres have been detected on Europa, Ganymede, Callisto, Rhea, and Dione.

II. FORMATION OF SATELLITES

A. Theoretical Models

Because the planets and their associated moons condensed from the same cloud of gas and dust at about the same time, the formation of the natural planetary satellites must be addressed within the context of the formation of the planets. The solar system formed 4.65 billion years ago. This age is derived primarily from radiometric dating of meteorites, which are believed to consist of primordial, unaltered matter. In the radiometric dating technique, the fraction of a radioactive isotope (usually rubidium, argon, or uranium), which has decayed into its daughter isotope, is measured. Since the rate at which these isotopes decay has been measured in the laboratory, it is possible to infer the time elapsed since formation of the meteorites and thus of the solar system.

The Sun and planets formed from a disk-shaped rotating cloud of gas and dust known as the protosolar nebula. When the temperature in the nebula cooled sufficiently, small grains began to condense. The difference in solidification temperatures of the constituents of the protosolar nebula accounts for the major compositional differences of the satellites. Since there was a temperature gradient as a function of distance from the center of the nebula, only those materials with high melting temperatures (e.g., silicates, iron, aluminum, titanium, and calcium) solidified in the central (hotter) portion of the nebula. The Earth's Moon consists primarily of these materials. Beyond the orbit of Mars, carbon, in combination with silicates and organic molecules, condensed to form a class of asteroids known as carbonaceous chondrites. Similar carbonaceous material is found on the surfaces of Phobos, several of the Jovian and Saturnian satellites, and perhaps the Uranian satellites. Beyond the outer region of the asteriod belt, formation temperatures were sufficiently cold to allow water ice to condense and remain stable. Thus, the Jovian satellites are primarily ice–silicate admixtures (except for Io, which has apparently outgassed all its water). On Saturn and Uranus, these materials are joined by methane and ammonia. For the satellites of Neptune and Pluto, formation temperatures were probably low enough for other volatiles, such as nitrogen and carbon monoxide, to exist in liquid and solid form. In general, the satellites, which formed in the inner regions of the solar system, are denser than the outer planets' satellites because they retained a lower fraction of volatile materials.

After small grains of material condensed from the protosolar nebula, electrostatic forces caused them to stick together. Collisions between these larger aggregates caused meter-sized particles, or planetesimals, to be accreted. Finally, gravitational attraction between the largest particles led the the formation of even larger, kilometer-sized planetesimals. The largest of these bodies swept up much of the remaining material to create the protoplanets and their companion satellite systems. One important concept of planetary satellite formation is that a satellite cannot accrete within Roche's limit, the distance at which the tidal forces of the primary become greater than the internal cohesive forces of the satellite.

The formation of the regular satellite systems of Jupiter, Saturn, and Uranus is sometimes thought to be a smaller scaled version of the formation of the solar system. A density gradient as a function of distance from Jupiter does exist for the Galilean satellites (see Table I). This implies that more volatiles (primarily ice) are included in the bulk composition as the distance increases. However, this simple scenario cannot be applied to Saturn or Uranus because their regular satellites do not follow this pattern.

The retrograde satellites are probably captured asteroids or large planetesimals left over from the major episode of planet formation. Except for Titan and perhaps Triton, the satellites are too small to possess gravitational fields sufficiently strong to retain an appreciable atmosphere against thermal escape.

B. Evolution

Soon after the satellites accreted, they began to heat up from the release of gravitational potential energy. An additional heat source was provided by the release of mechanical energy during the heavy bombardment of their surfaces by remaining debris. The satellites Phobos, Mimas, and Tethys all have impact craters caused by

bodies that were nearly large enough to break them apart; probably such catastrophes did occur. The decay of radioactive elements found in silicate materials provided another major source of heat. The heat produced in the larger satellites was sufficient to cause melting and chemical fractionation; the dense material, such as silicates and iron, went to the center of the satellite to form a core, while ice and other volatiles remained in the crust.

Some satellites, such as the Earth's Moon, Ganymede, and several of the Saturnian satellites, underwent periods of melting and active geology within a billion years of their formation and then became quiescent. Others, such as Io, Triton, and possibly Enceladus and Europa, are currently geologically active. For nearly a billion years after their formation, the satellites all underwent intense bombardment and cratering. The bombardment tapered off to a slower rate and presently continues. By counting the number of craters on a satellite's surface and making certain assumptions about the flux of impacting material, geologists are able to estimate when a specific portion of a satellite's surface was formed. Continual bombardment of satellites causes the pulverization of the surface to form a covering of fine material known as a regolith.

III. OBSERVATIONS OF SATELLITES

A. Telescopic Observations

1. Spectroscopy

Before the development of interplanetary spacecraft, all observations from Earth of objects in the solar system were obtained by telescopes. One particularly useful tool of planetary astronomy is spectroscopy, or the acquisition of spectra from a celestial body.

Each component of the surface or atmosphere of a satellite has a characteristic pattern of absorption and emission bands. Comparison of the astronomical spectrum with laboratory spectra of materials which are possible components of the surface yields information on the composition of the satellite. For example, water ice has a series of absorption features between 1 and 4 μ. The detection of these bands on three of the Galilean satellites and several satellites of Saturn, Uranus, Neptune, and Charon demonstrated that water ice is a major constituent of their surfaces. Other examples are the detections of SO_2 frost on the surface of Io, methane in the atmosphere of Titan, and nitrogen on Triton.

2. Photometry

Photometry of planetary satellites is the accurate measurement of radiation reflected to an observer from their surfaces or atmospheres. These measurements can be compared to light-scattering models that are dependent on physical parameters, such as the porosity of the optically active upper surface layer, the albedo of the material, and the degree of topographic roughness. These models predict brightness variations as a function of solar phase angle (the angle between the observer, the Sun, and the satellite). Like the Earth's Moon, the planetary satellites present changing phases to an observer on Earth. As the face of the satellite becomes fully illuminated to the observer, the integrated brightness exhibits a nonlinear surge in brightness that is believed to result from the disappearance of mutual shadowing among surface particles. The magnitude of this surge, known as the "opposition effect," is greater for a more porous surface. Optical effects, such as the coherent backscattering of radiation, can also increase the brightness of a fully illuminated satellite's disk.

One measure of how much radiation a satellite reflects is the geometric albedo, p, which is the disk-integrated brightness at "full moon" (or a phase angle of zero degrees) compared to a perfectly reflecting, diffuse disk of the same size. The phase integral, q, defines the angular distribution of radiation over the sky as follows:

$$q = 2\int_0^{\pi} \Phi(\alpha)\sin\alpha\, d\alpha,$$

where $\Phi(\alpha)$ is the disk integrated brightness and α is the phase angle.

The Bond albedo, which is given by $A = p \times q$, is the ratio of the integrated flux reflected by the satellite to the integrated flux received. The geometric albedo and phase integral are wavelength dependent, whereas a true (or bolometric) Bond albedo is integrated over all wavelengths.

Another ground-based photometric measurement, which has yielded important information on the satellites surfaces, is the integrated brightness of a satellite as a function of orbital angle. For a satellite in synchronous rotation with its primary, the subobserver geographical longitude of the satellite is equal to the longitude of the satellite in its orbit. Observations showing significant albedo and color variegations for Io, Europa, Rhea, Dione, and especially Iapetus suggest that diverse geologic terrains coexist on these satellites. This view was confirmed by images obtained by the Voyager spacecraft.

Valuable contributions here also have been made by the Hubble Space Telescope, which can resolve the disks of the largest satellites.

3. Radiometry

Satellite radiometry is the measurement of radiation which is absorbed and reemitted at thermal wavelengths. The

distance of each satellite from the Sun determines the mean temperature for the equilibrium condition that the absorbed radiation is equal to the emitted radiation as follows:

$$\pi R^2(F/r^2)(1-A) = 4\pi R^2 \varepsilon \sigma T^4$$

or

$$T = \left(\frac{(1-A)F}{4\sigma \varepsilon r^2}\right)^{1/4},$$

where R is the radius of the satellite, r is the Sun–satellite distance, ε is the emissivity, τ is Stefan–Boltzmann's constant, A is the Bond albedo, and F is the incident solar flux (a slowly rotating body would radiate over $2\pi R^2$). Typical mean temperatures in degrees Kelvin for the satellites are: the Earth's Moon, 280; Io, 106; Titan, 97; the Uranian satellites, 60; and the Neptunian satellites, 45. For thermal equilibrium, measurements as a function of wavelength yield a blackbody curve characteristic of T: With the exception of Titan, the temperatures of the satellites closely follow the blackbody emission values. The discrepancy for Titan may be due to a weak greenhouse effect in the satellite's atmosphere.

Another possible use of radiometric techniques, when combined with photometric measurements of the reflected portion of the radiation, is the estimate of the diameter of a satellite. A more accurate method of measuring the diameter of a satellite from Earth involves measuring the light from a star as it occulted by the satellite. The time the starlight is dimmed is proportional to the satellite's diameter.

A third radiometric technique is the measurement of the thermal response of a satellite's surface as it is being eclipsed by its primary. The rapid loss of heat from a satellite's surface indicates a thermal conductivity consistent with a porous surface. Eclipse radiometry of Phobos, Callisto, and Ganymede suggests these objects all lose heat rapidly.

4. Polarimetry

Polarimetry is the measurement of the degree of polarization of radiation reflected from a satellite's surface. The polarization characteristics depend on the shape, size, and optical properties of the surface particles. Generally, the radiation is linearly polarized and is said to be negatively polarized if it lies in the scattering plane, and positively polarized if it is perpendicular to the scattering plane. Polarization measurements as a function of solar phase angle for atmosphereless bodies are negative at low phase angles; comparisons with laboratory measurements indicate this is characteristic of complex, porous surfaces consisting of multisized particles. In 1970, ground-based polarimetry of Titan that showed it lacked a region of negative polarization led to the correct conclusion that it has a thick atmosphere.

5. Radar

Radar is the active investigation of the surface of a planet or satellite gained through the study of returned radio signals. Radar techniques are used to measure the distances to satellites and their spin states, their topography, and the physical properties of their surfaces. The most distant satellite detected by radar is Titan.

B. Spacecraft Exploration

1. Imaging Observations

Interplanetary missions to the planets and their moons have enabled scientists to increase their understanding of the solar system more since 1970 than over the previous total years of scientific history. Analysis of data returned from spacecraft has led to the development of whole new fields of scientific endeavor, such as planetary geology. From the earliest successes of planetary imaging, which included the flight of a Soviet *Luna* spacecraft to the far side of the Earth's Moon to reveal a surface unlike that of the visible side, devoid of smooth lunar plains, and the crash landing of a United States *Ranger* spacecraft, which sent back pictures showing that the Earth's Moon was cratered down to meter scales, it was evident that interplanetary imaging experiments had immense capabilities. Table II summarizes the successful spacecraft missions to the planetary satellites.

Most images from spacecraft are obtained by charged-coupled devices (CCDs), which are large arrays of electronic solid-state devices. A computer onboard the spacecraft records these numbers and sends them by means of a radio transmitter to the Earth, where another computer reconstructs the image.

The first spacecraft to send pictures of a moon other then the Earth's was *Mariner 9*, which began orbiting Mars in 1971 and sent back images of Phobos and Deimos showing that these satellites are heavily cratered, irregular objects. Even more highly resolved images were returned by the *Viking* orbiters in 1976 (see Fig. 2). The *Pioneer* spacecraft, which were launched in 1972 and 1973 toward an encounter with Jupiter and Saturn, returned the first disk-resolved images of the Galilean satellites. By far the greatest scientific advancements were made by the *Voyager* spacecraft, which returned thousands of images of the Satellites of Jupiter, Saturn, Uranus, and Neptune, some of which are shown in Section IV. Color information for the objects was obtained by means of six broadband filters attached to the camera. The return of large numbers of images with resolution down to a

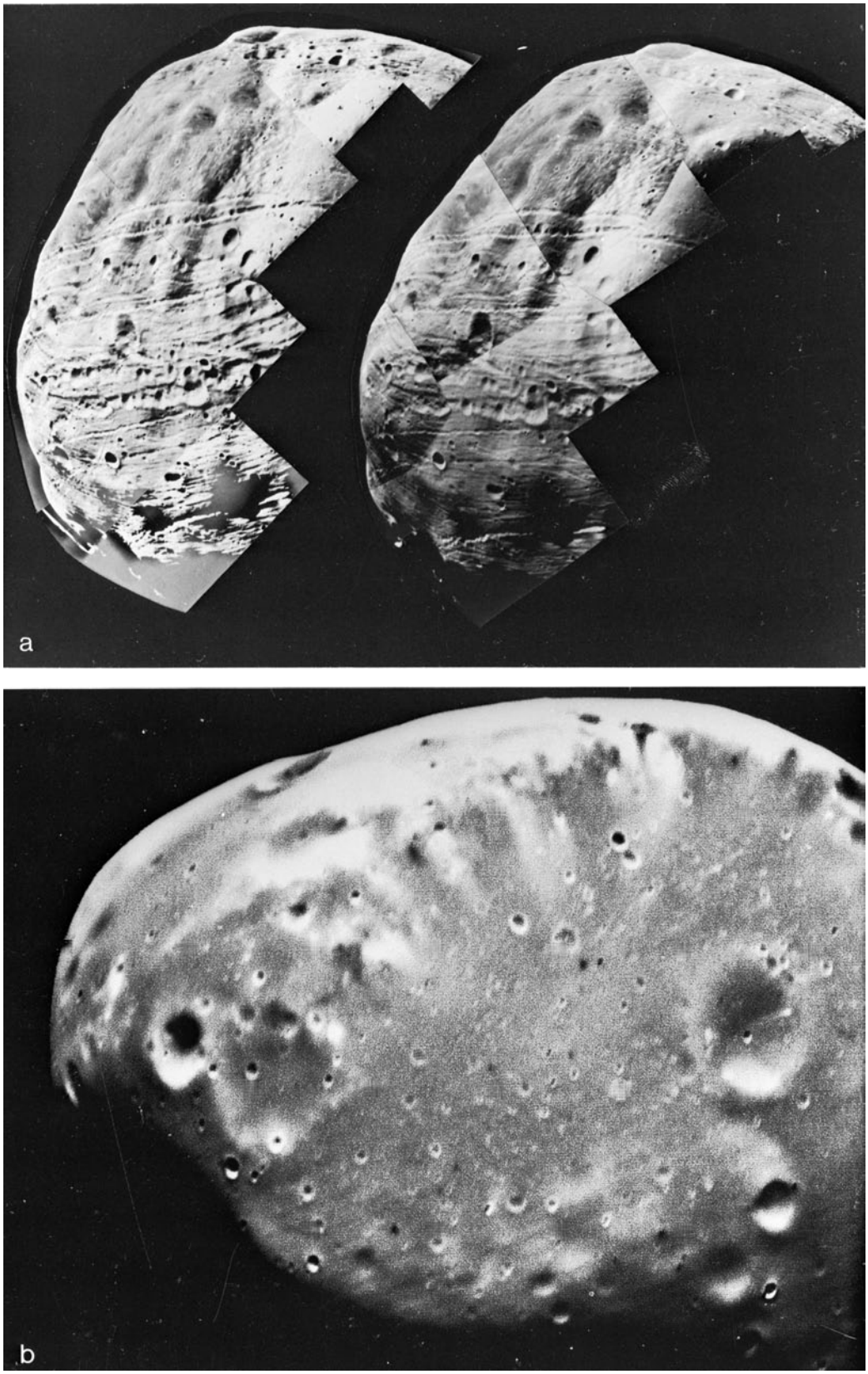


FIGURE 2 The two moons of Mars: (a) Phobos and (b) Deimos. Both pictures were obtained by the *Viking* spacecraft.

kilometer has enabled geologists to construct geologic maps, to make detailed crater counts, and to develop realistic scenarios for the structure and evolution of the satellites. The *Galileo* spacecraft, which has been in orbit around Jupiter since 1995, has returned far more detailed images (10s-of-meters resolution) and spectra of the planet's satellite system (see Section IV). In the case of Io, scientists were able to detect significant changes in the extent of its lava flows which occurred during the 16-year period between *Voyager* and *Galileo*.

The upcoming *Cassini* mission, launched in 1997, will conduct a detailed investigation of the Saturnian system between 2004 and 2008. The payload includes 12 instruments (imaging systems, spectrometers, and field and particle detectors) as well as a probe that will measure the properties of the atmosphere of Titan and land on its surface (the probe is not designed to operate beyond a few minutes after landing).

2. Other Experiments

Although images are the most spectacular data returned by spacecraft, a whole array of equally valuable experiments are included in each scientific mission. For example, a gamma ray spectrometer aboard the lunar orbiters was able to map the abundance of iron and titanium across the Moon's surface. Seismometers placed on the Moon recorded waves from small moonquakes. Measurements of heat flow from the interior of the Moon enabled scientists to understand something about its composition and present evolutionary state. Onboard the *Voyager* spacecraft were several experiments which are valuable for satellite investigations, including an infrared spectrometer capable of mapping temperatures; an ultraviolet spectrometer; a photopolarimeter, which simultaneously measured the color, intensity, and polarization of light; and a radio science experiment that was able to measure the pressure of Titan's atmosphere by observing how radio waves passing through it were attenuated. The *Galileo* magnetometer obtained crucial evidence for subsurface liquid water on Europa and perhaps Ganymede and Callisto.

IV. INDIVIDUAL SATELLITES

A. The Earth's Moon

1. Introduction

The Earth's Moon has played a key role in the lore and superstition of the world's peoples. When Galileo first turned his telescope on the lunar disk in the early 1600s he clearly perceived the most obvious fact of lunar morphology: the demarcation of the Moon's surface into a dark smooth terrain and a brighter more mountainous terrain. This delineation is responsible for the appearance of the "Man in the Moon." The dark areas were called *maria* (Latin for "seas"; the singular is *mare*) because of their visual resemblance to oceans and the bright areas were called *terrae* (Latin for "land"; the singular is *terra*), or the lunar uplands. The lunar mountains were named after terrestrial mountain chains and the craters after famous astronomers. The appearance of the Moon through a large telescope is shown in Fig. 3.

The physical characteristics of the Moon are listed in Table I. Since the Moon has been tidally despun, it is locked in synchronous rotation, and it was not until 1959 that the back side of the Moon was observed by a Soviet spacecraft. Because gravitational perturbations on the Moon cause it to wobble, or librate, in its orbit, about 60% of the surface of the Moon is visible from the Earth. The far hemisphere has fewer maria than the visible side.

According to theories of tidal evolution, the Moon is receding from the Earth and the Earth's spin rate is in turn slowing down. Thus in the past the lunar month and terrestrial day were shorter. Evidence for this is found in the fossilized shells of certain sea corals, which deposit calcareous material in a cycle following the lunar tides.

2. Origin

The four standard theories for the origin of the Moon are (1) capture, in which the Moon forms elsewhere in the solar system and is gravitationally captured by the Earth; (2) fission, which asserts that the Moon broke off from the Earth early in its formation; (3) coaccretion, which asserts that the Earth and Moon formed independently but nearly simultaneously near their present locations; and (4) impact, in which a Mars-sized body collides with the proto-Earth to create a debris disk that rapidly reaccretes to form the Moon. The latter theory is currently favored.

3. Early History of the Moon

Soon after the Moon accreted it heated up due to the reasons outlined in Section II.B. The result was the melting and eruption of basaltic lava onto the lunar surface between 3.8 to about 2.8 billion years ago to form the lunar maria. This lava was highly fluid under the weaker gravitational field of the Moon and spread over vast distances.

Before Soviet and American spacecraft explored the Moon, there was considerable debate over whether the craters on the Moon were of impact or of volcanic origin. Morphological features, such as bright rays and ejecta blankets, around large craters show that they were formed by impacts.

The ringlike structures that delineate the maria are the outlines of impact basins, which filled in with lava. The

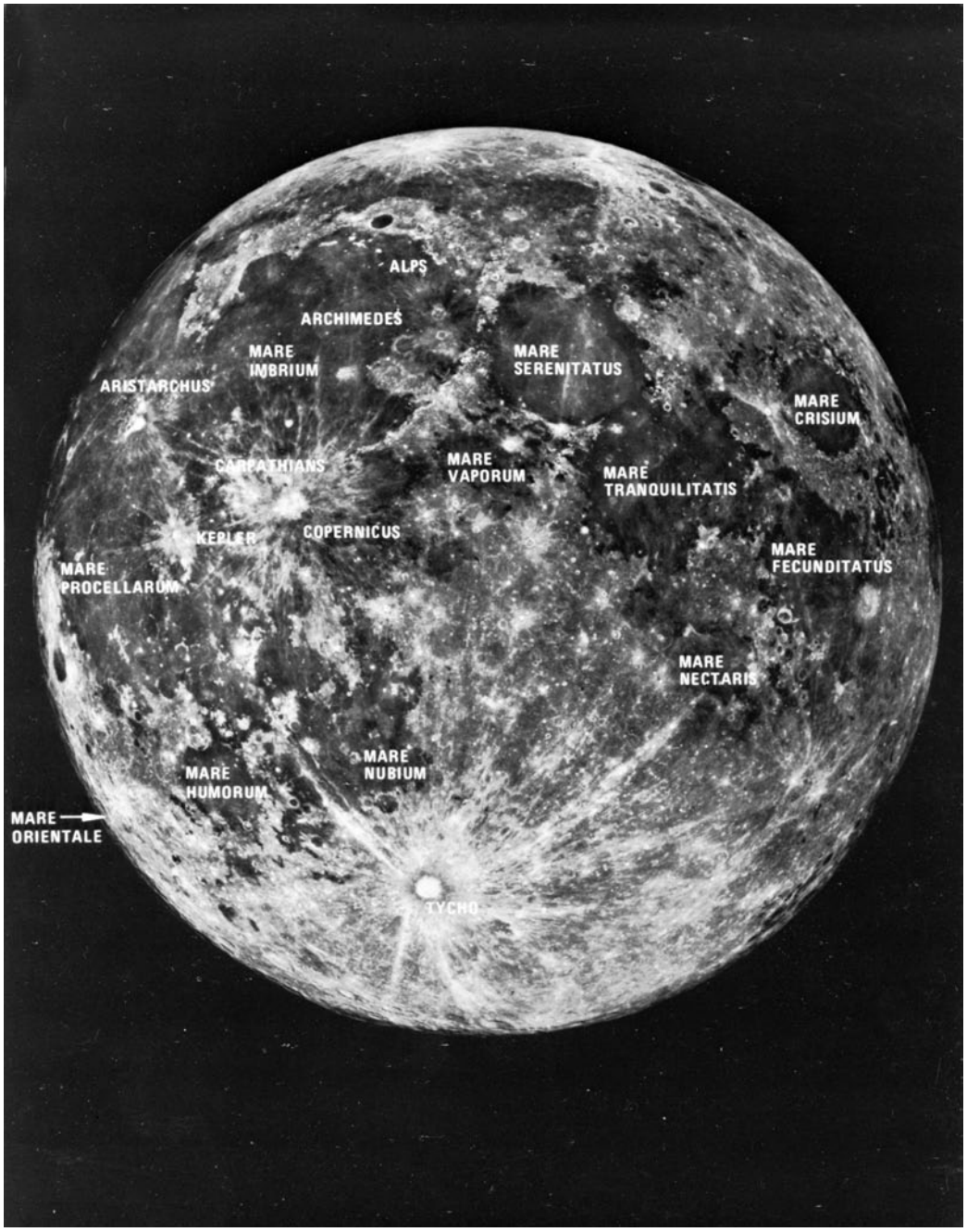


FIGURE 3 A telescopic view of the Moon with the major features marked. (Photograph courtesy of Lick Observatory.)

maria are not as heavily cratered as the uplands because the lava flows which created them obliterated preexisting craters.

4. Nature of the Lunar Surface

The physical properties of the Moon are known more accurately than those of any other satellite because of the extensive reconnaissance and study by spaceraft (see Table II). Rock samples have been returned by both American manned and Soviet unmanned missions.

The astronauts who walked on the Moon found that the upper few centimeters of the lunar surface was covered by fine dust and pulverized rock. This covering, or regolith, is the result of fragmentation of particles from constant bombardment of the Moon's surface during its history

and may be structurally similar to the regoliths of other satellites.

Analysis of lunar samples revealed important chemical differences between the Earth and Moon. The lunar uplands consist of a low-density, calcium-rich igneous rock known as anorthosite. The younger maria are composed of a dark basalt rich in the minerals olivine and pyroxene. Some of the lunar basalts, known as KREEP, were found to be anonymously rich in potassium (K), rare earth elements (REE), and phosphorous (P). Formations such as lava tubes and vents, which are similar to terrestrial volcanic features, are found in the maria.

The Moon has no water or significant atmosphere. Surface temperatures range from 100 to 380 K.

B. Phobos and Deimos

Mars has two small satellites, Phobos and Deimos, which were discovered by the American astronomer Asaph Hall in 1877. In Jonathan Swift's moral satire *Gulliver's Travels* (published in 1726), a fanciful but coincidentally accurate prediction of the existence and orbital characteristics of two small Martian satellites was made. These two objects are barely visible in the scattered light from Mars in Earth-based telescopes. Most of what is known about Phobos and Deimos was obtained from the *Mariner 9* and the *Viking 1* and *2* missions to Mars (see Table II). Their physical and orbital properties are listed in Table I. Both satellites are shaped approximately like ellipsoids and are in synchronous rotation. Phobos, and possibly Deimos, has a regolith of dark material similar to that found on carbonaceous asteroids common in the outer asteroid belt. Thus the satellites may have been asteroids or asteroidal fragments, which were perturbed into a Mars-crossing orbit and captured.

Both satellites are heavily cratered, which indicates that their surfaces are at least 3 billion years old (Fig. 2). However, Deimos appears to be covered with a fine, light-colored dust, which gives its surface a smoother appearance. The dust may exist because the surface is more easily pulverized by impacts or simply because it is easier for similar material to escape from the gravitational field of Phobos, which is closer to Mars. The surface of Phobos is extensively scored by linear grooves that appear to radiate from the huge impact crater Stickney (named after the surname of Asaph Hall's wife, who collaborated with him). The grooves are probably fractures caused by the collision that produced Stickney. There is some evidence that tidal action is bringing Phobos, which is already inside Roche's limit, closer to Mars. The satellite will either disintegrate (perhaps to form a ring) or crash into Mars in about 100 million years. The suggestion that Phobos' orbit is decaying because it is a hollow extraterrestrial space station has no basis in fact.

C. The Galilean Satellites of Jupiter

1. Introduction and Historical Survey

When Galileo trained his telescope on Jupiter he was amazed to find four points of light which orbited the giant planet. These were the satellites Io, Europa, Ganymede, and Callisto, planet-sized worlds known collectively as the Galilean satellites. Analysis of telescopic observations since the early 1600s revealed certain basic features of their surfaces. There was spectroscopic evidence for water ice on the outer three objects. The unusually orange color of Io was hypothesized to be due to elemental sulfur. Orbital phase variations were significant, particularly in the cases of Io and Europa, which indicated the existence of markedly different terrains on their surfaces. Large opposition effects observed on Io and Callisto suggested their surfaces were porous, whereas the lack of an opposition effect on Europa suggested a smooth surface. The density of the satellites decreases as a function of distance from Jupiter (Table 1).

Theoretical calculations suggested the satellites had differentiated to form silicate cores and (in the case of the outer three) ice crusts. There is the good probability that the mantles of the outer three satellites are liquid water. In 2000, the *Galileo* spacecraft's magnetometer detected induced magnetic fields on the three outer satellites that strongly suggested the existence of a conducting, water ice lower crust. The induced field coaligned with the Jovian magnetic field, regardless of the satellite's orbital position. A camera and spectrometers on Galileo, as well as the instruments on the two Voyager missions (Table 2), revealed the Galilean satellites to be unique geological worlds.

2. Io

About the size of the Moon, Io is the only body in the Solar System other than the Earth on which active volcanism has been observed. The *Voyager* spacecraft detected nine currently erupting plumes, scores of calderas (volcanic vents), and extensive lava flows consisting of nearly pure elemental sulfur (Fig. 4). As sulfur cools, it changes from dark brown to red, to orange, and finally to yellow, which accounts for the range of colors on Io's surface. There are nearly black liquid sulfur lava lakes with floating chunks of solid sulfur. Sulfur dioxide is driven out of the volcanoes to condense or absorb onto the surface as white deposits. A thin, transient atmosphere with a pressure less than 1 millionth of the Earth's and consisting primarily of sulfur dioxide has been detected. There appears to be a total absence of water on the surface of

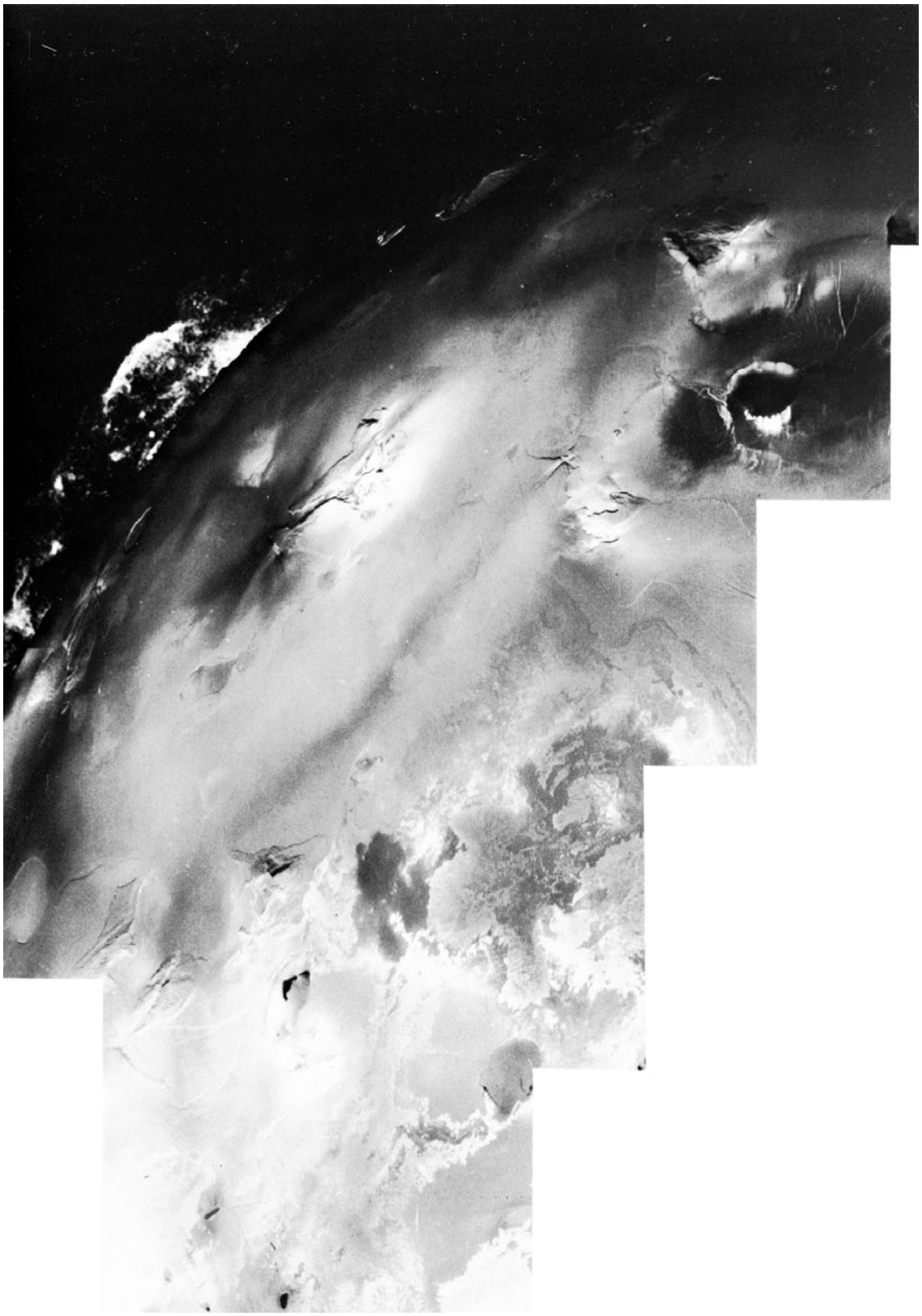


FIGURE 4 A highly processed *Voyager* image of the Galilean satellite Io, showing wispy structures in the volcanic plumes (upper left), volcanic vents, and sulfur lava flows.

Io, probably because it has all been degassed from the interior from extensive volcanism and escaped into space. The total lack of impact craters means the entire surface is young and geologically active. Among the findings of the *Galileo* spacecraft were silicate (as opposed to pure sulfur) vulcanism, a 9-km-high mountain (although most of the satellite's surface consists of flat plains rising no more than 1 km), lava fountains and curtains erupting onto the surface, a high-altitude ionosphere, and an iron core.

The heat source for melting and subsequent volcanism is the dissipation of tidal energy from Jupiter and the other Galilean satellites. As Io moves in its orbit, its distance from Jupiter changes as the other satellites exert different forces depending on their distance from Io. The varying tidal stresses cause Io to flex in and out. This mechanical energy is released as heat, which causes melting in Io's mantle.

A spectacular torus of ionized particles, primarily sulfur, oxygen, and sodium, corotates with Jupiter's strong magnetic field at Io's orbital position. High energy ions in the Jovian magnetosphere knock off and ionize surface particles (about a ton each second), which are swept up and entrained in the field lines. An additional source of material for the torus is sulfur and sulfur dioxide from the plumes. Aurorae seen on Jupiter are caused by particles from the torus being conducted to the planet's polar regions. Because Io is a conducting body moving in the magnetic field of Jupiter it generates a flux tube of electric current between itself and the planet. Radio emissions from the Jovian atmosphere, which correlate with the orbital position of Io, appear to be triggered by the satellite.

3. Europa

When the Voyagers encountered the second Galilean satellite, Europa, they returned images of bright, icy plains crisscrossed by an extensive network of darker fractures (see Fig. 5). The existence of only a handful of impact craters suggested that geological processes were at work on the satellite until a few hundred million years ago or less. Europa is very smooth: the only evidence for topographical relief is the scalloped ridges with a height of a few hundred meters (see bottom of Fig. 5).

Part of Europa is covered by a darker mottled terrain. Dark features also include hundreds of brown spots of unknown origin and larger areas, which appear to be the result of silicate-laden water erupting onto the surface (bottom left of Fig. 5). The reddish hue of Europa is believed to be due to contamination by sulfur from Io.

The mechanism for the formation of cracks on Europa is probably some form of tidal interaction and subsequent heating, melting, and refreezing. *Galileo* images show evidence for both compression and separation of plates of ice on the satellite's surface. The likelihood of a subsurface liquid ocean on Europa has led scientists to discuss the possibility of a primitive lifeform teeming in the mantle. However, no evidence for life there exists, and such ideas are pure speculation. The Hubble Space Telescope (HST) detected a tenuous molecular oxygen atmosphere on Europa.

4. Ganymede

The icy moon Ganymede, which is the largest Galilean satellite, also shows evidence for geologic activity as recently as a billion years ago. A dark, heavily cratered terrain is transected by more recent, brighter grooved terrain (see Fig. 6). Although they show much diversity, the grooves are typically 10 km wide and $\frac{1}{3}$ to $\frac{1}{2}$ km high. They were implaced during several episodes between 3.5 and 4 billion years ago. Their formation may have occurred after a melting and refreezing of the core, which caused a slight crustal expansion and subsequent faulting and flooding by subsurface water.

The grooved terrain of Ganymede is brighter because the ice is not as contaiminated with rocky material that accumulates over the eons from impacting bodies. The satellite is also covered with relatively fresh bright craters, some of which have extensive ray systems. In the cratered terrain there appear outlines of old, degraded craters, which geologists called palimpsests. The polar caps of Ganymede are brighter than the equatorial regions; this is probably due to the migration of water molecules released by evaporation and impact toward the colder high latitudes. HST detected a thin atmosphere of molecular oxygen, similar to that of Europa.

5. Callisto

Callisto is the only Galilean satellite that does not show evidence for extensive resurfacing at any point in its history. It is covered with a relatively uniform, dark terrain saturated with craters (Fig. 7). There is, however, an absence of craters larger than 150 km. Ice slumps and flows over periods of billions of years and is apparently not able to maintain the structure of a large crater as long as does rocky material. One type of feature unique to Callisto is the remnant structures of numerous impacts. The most prominent of these, the Valhalla basin, is a bright spot encircled by as many as 13 fairly regular rings (as shown in Fig. 7). The leading hemisphere of Callisto appears to be fluffier than the trailing side, possibly due to more intense meteoritic bombardment. High-resolution *Galileo* images show a dark mantling deposit (Fig. 8) that may be the remains from the preferential volatization of ice. The

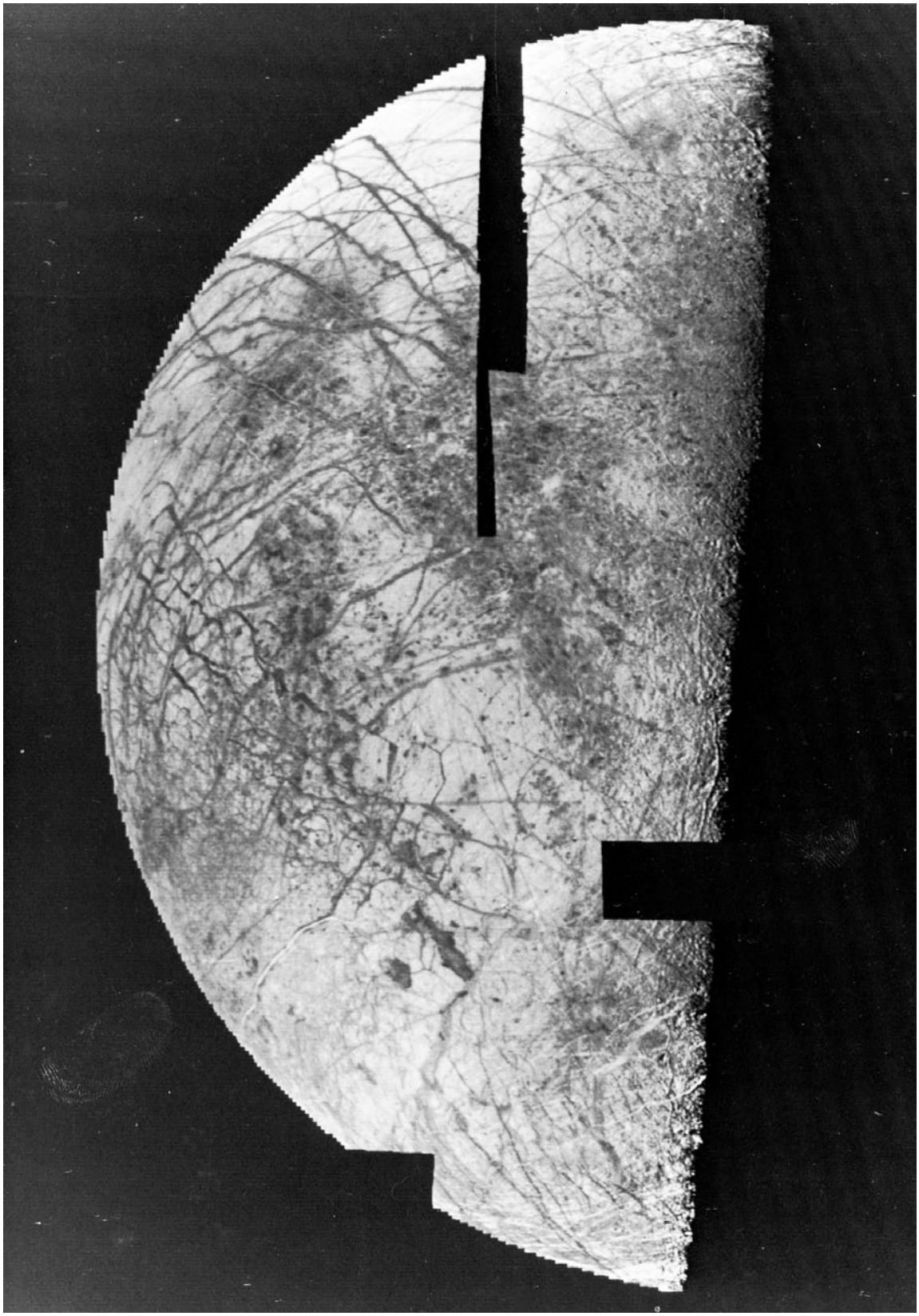


FIGURE 5 A photomosaic of Europa assembled from *Voyager 2* images.

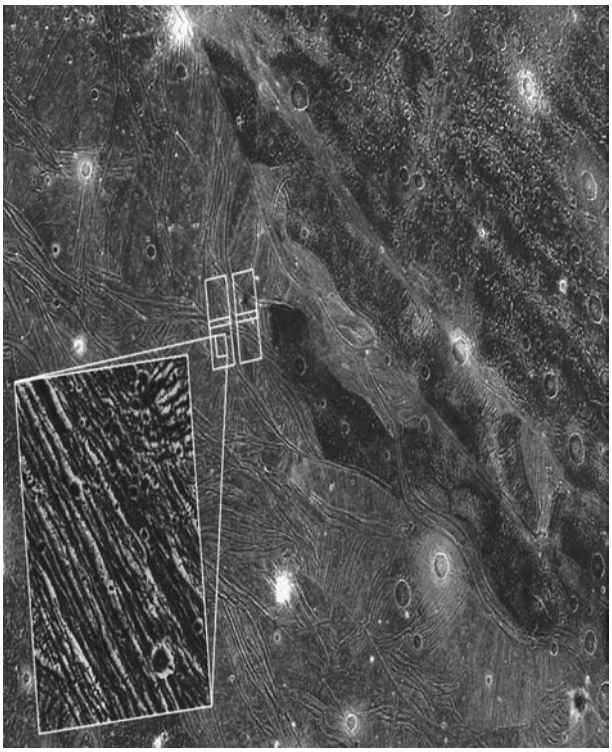


FIGURE 6 A high resolution (~80-M) *Galileo* image of the Uruk Sulcus regim placed on an earlier *Voyager* image. Bright grooved terrain, older darker terrain, and bright impact craters are visible in the *Voyager* image. The four boxes show the entire *Galileo* coverage.

infrared spectrometer on *Galileo* also detected a possible tenuous CO_2 atmosphere.

6. The Small Satellites of Jupiter

Jupiter has 28 known small satellites, including three discovered by the *Voyager* mission. They are all probably irregular in shape (see Table I). Within the orbit of Io are at least three satellites: Amalthea, Adrastea, and Metis. Amalthea is a dark, reddish heavily cratered object reflecting less than 5% of the radiation it receives; the red color is probably due to contamination by sulfur particles from Io. Little else is known about its composition except that the dark material may be carbonaceous.

Adrastea and Metis, both discovered by *Voyager*, are the closest known satellites to Jupiter and move in nearly identical orbits just outside the outer edge of the thin Jovian ring, for which they may be a source of particles. Between Amalthea and Io lies the orbit of Thebe, also discovered by Voyager. Little is known about the composition of these satellites, but they are most likely primarily rock–ice mixtures. The three inner satellites sweep out particles in the Jovian magnetosphere to form voids at their orbital positions.

Moving outward from Jupiter, we find a class of four satellites moving in highly inclined orbits (Lysithea, Elara, Himalia, and Leda). They are dark, spectrally neutral objects, reflecting only 2 or 3% of incident radiation and may be similar to carbonaceous asteroids.

Another family of objects is the outermost four satellites, which also have highly inclined orbits, except they move in the retrograde direction around Jupiter. They are Sinope, Pasiphae, Carme, and Ananke, and they may be captured asteroids.

D. The Saturnian System

1. The Medium Sized Icy Satellites: Rhea, Dione, Tethys Mimas, Enceladus, and Iapetus

The six largest satellites of Saturn are smaller than the Galilean satellites but still sizable—as such they represent a unique class of icy satellite. Earth-based telescopic measurements showed the spectral signature of ice for Tethys, Rhea, and Iapetus; Mimas and Enceladus are close to Saturn and difficult to observe because of scattered light from the planet. The satellites' low densities and high albedos (Table I) imply that their bulk composition is largely water ice, possibly combined with ammonia. They have smaller amounts of rocky silicates than the Galilean satellites. Resurfacing appears to have occurred on several of the satellites. Most of what is presently known of the Saturnian system was obtained from the *Voyager* flybys in 1980 and 1981.

The innermost medium-sized satellite Mimas is covered with craters, including one (named Arthur), which is as large as a third of the satellite's diameter (upper left of Fig. 9). The impacting body was probably nearly large enough to break Mimas apart; such disruptions may have occurred to other objects. There is a suggestion of surficial grooves that may be features caused by the impact. The craters on Mimas tend to be high-rimmed, bowl-shaped pits; apparently surface gravity is not sufficient to have caused slumping.

The next satellite outward from Saturn is Enceladus, an object that was known from telescopic measurements to reflect nearly 100% of the visible radiation incident on it (for comparison, the Moon reflects only about 11%). The only likely composition consistent with this observation is almost pure water ice. *Voyager 2* images of Enceladus show an object that has been subjected, in the recent geologic past, to extensive resurfacing; grooved formations similar to those on Ganymede are evident (see Fig. 10). The lack of impact craters on this terrain is consistent with an age less than a billion years. It is possible that some form of ice volcanism is presently active on Europa. The heating mechanism is believed to be tidal interactions, perhaps with Dione. About half of the surface observed

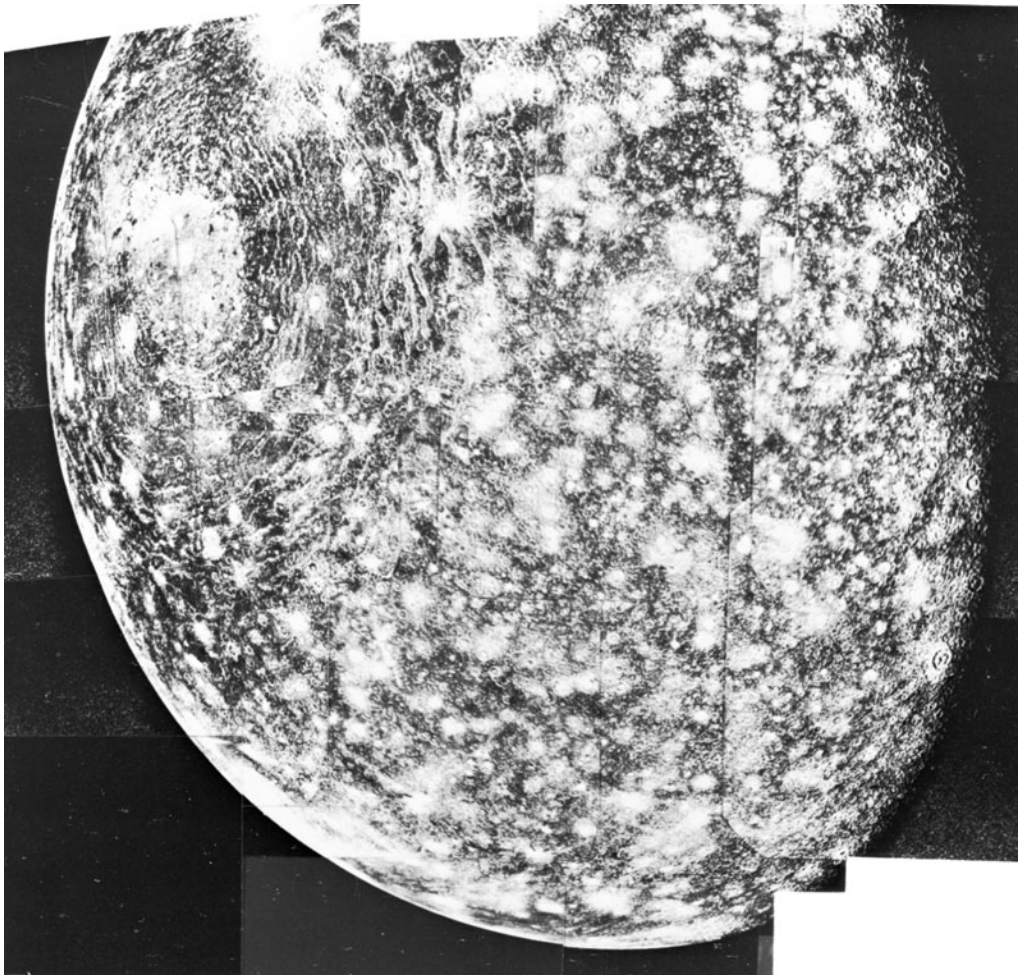


FIGURE 7 A *Voyager 1* photomosaic of Callisto. The Valhalla impact basin, which is 600 km wide, dominates the surface.

by *Voyager* is extensively cratered and dates from nearly 4 billion years ago.

A final element to the enigma of Enceladus is the possibility that it is responsible for the formation of the E-ring of Saturn, a tenuous collection of icy particles that extends from inside the orbit of Enceladus to past the orbit of Dione. The position of maximum thickness of the ring coincides with the orbital position of Enceladus. If some form of volcanism is presently active on the surface, it could provide a source of particles for the ring. An alternative source mechanism is an impact and subsequent escape of particles from the surface.

Tethys is covered with impact craters, including Odysseus, the largest known impact structure in the solar system. The craters tend to be flatter then those on Mimas or the Moon, probably because of relaxation and flow over the eons under Tethys's stronger gravitational field. Evidence for resurfacing episodes is seen in regions that have fewer craters and higher albedos. In addition, there is a huge trench formation, the Ithaca Chasma, which may be a degraded form of the grooves found on Enceladus.

Dione, which is about the same size as Tethys, exhibits a wide diversity of surface morphology. Most of the surface is heavily cratered (Fig. 11), but gradations in crater density indicate that several periods of resurfacing events occurred during the first billion years of its existence. One side of the satellite is about 25% brighter than the other. Wispy streaks (see Figs. 9 and 11), which are about 50% brighter than the surrounding areas, are believed to be the result of internal activity and subsequent implacement of erupting material. Dione modulates the radio emission from Saturn, but the mechanism for this phenomenon is unknown.

Rhea appears to be superficially very similar to Dione (see Fig. 9). Bright wispy streaks cover one hemisphere. However, there is no evidence for any resurfacing events early in its history. There does seem to be a dichotomy between crater sizes—some regions lack large craters

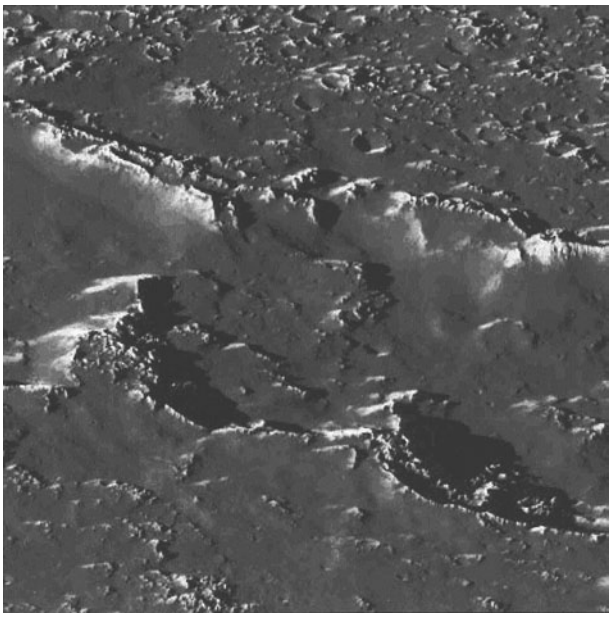


FIGURE 8 (Near Picture) *Galileo* image of Callisto showing the dark mantling deposits found in certain regions.

while other regions have a preponderance of such impacts. The larger craters may be due to a population of larger debris more prevalent during an earlier episode of collisions.

HST discovered a tenuous atmosphere of ozone (O_3) on both Rhea and Dione. When Cassini discovered Iapetus in 1672, he noticed that at one point in its orbit around Saturn it was very bright, whereas on the opposite side of the orbit it nearly altogether disappeared. He correctly deduced that one hemisphere is composed of highly reflective material, while the other side is much darker. Voyager images show that the bright side, which reflects nearly 50% of the incident radiation, is fairly typical of a heavily cratered icy satellite. The other side, which is centered on the direction of motion, is coated with a material with a reflectivity of about 3–4% (see Fig. 9).

Scientists still do not agree on whether the dark material originated from an exogenic source or was endogenically created. One scenario for the exogenic deposit of material entails dark particles being ejected from Phoebe and drifting inward to coat Iapetus, The major criticism of this model is that the dark material on Iapetus is redder than that on Phoebe, although the material could have undergone chemical changes after its expulsion from Phoebe to make it redder. One observation lending credence to an internal origin is the concentration of material on crater floors, which implies an infilling mechanism. In one model, methane erupts from the interior and is subsequently darkened by ultraviolet radiation.

Other aspects of Iapetus are unusual. It is the only large Saturnian satellite in a highly inclined orbit. It is less dense than objects of similar albedo; this implies a higher fraction of ice or possibly methane or ammonia in its interior.

2. Titan

Titan is a fascinating world that one member of the *Voyager* imaging team called "a terrestrial planet in a deep freeze." It has a thick atmosphere that includes a layer of photochemical haze (Fig. 12) and a surface possibly covered with lakes of methane or ethane. Methane was discovered by G. P. Kuiper in 1944: the *Voyager* experiments showed that the major atmospheric constituent is nitrogen, the major component of the Earth's atmosphere. Methane (which is easier to detect from Earth because of prominent spectroscopic lines) may comprise only a few percentages or less. The atmospheric pressure of Titan is 1.5 times that of the Earth's; however, Titan's atmosphere extends much further from the surface (nearly 100 km) on account of the satellite's lower gravity. The atmosphere is thick enough to obscure the surface, although HST has observed bright surface markings through methane atmospheric windows.

Titan's density (Table I) implies a bulk composition of 45% ice and 55% silicates. It probably has a differentiated rocky core. Titan was able to retain an appreciable atmosphere while the similarly sized Ganymede and Callisto were not because more methane and ammonia condensed at Titan's lower formation temperature. The methane has remained, whereas ammonia has been photochemically dissociated into molecular nitrogen and hydrogen, the latter being light enough to escape the gravitational field of Titan. The escaped hydrogen forms a tenuous torus at the orbital position of Titan. Although it has not been directly detected, argon may comprise a few percentages of the atmosphere.

The infrared spectrometer on *Voyager* detected nearly a dozen organic compounds such as acetylene (C_2H_2), ethane (C_2H_6), and hydrogen cyanide (HCN), which play an important role in prebiological chemistry. These molecules, which are constantly being formed by the interaction of ultraviolet radiation with nitrogen and methane, constitute the haze layer of aerosol dust in the upper atmosphere. Much of this material, which gives Titan a reddish color, "rains" onto the surface and possibly forms lakes of ethane or methane. The role of methane on the surface and in the atmosphere of Titan may be similar to that of water on Earth; the triple point of methane (where it can coexist as a solid, liquid, or gas) is close to the surface temperature of Titan (93° K). In this scenario, methane in the atmosphere covers methane lakes and ice on the surface. Radar measurements suggest that most of Titan's surface is solid: Any liquid exists as lakes rather than global oceans.

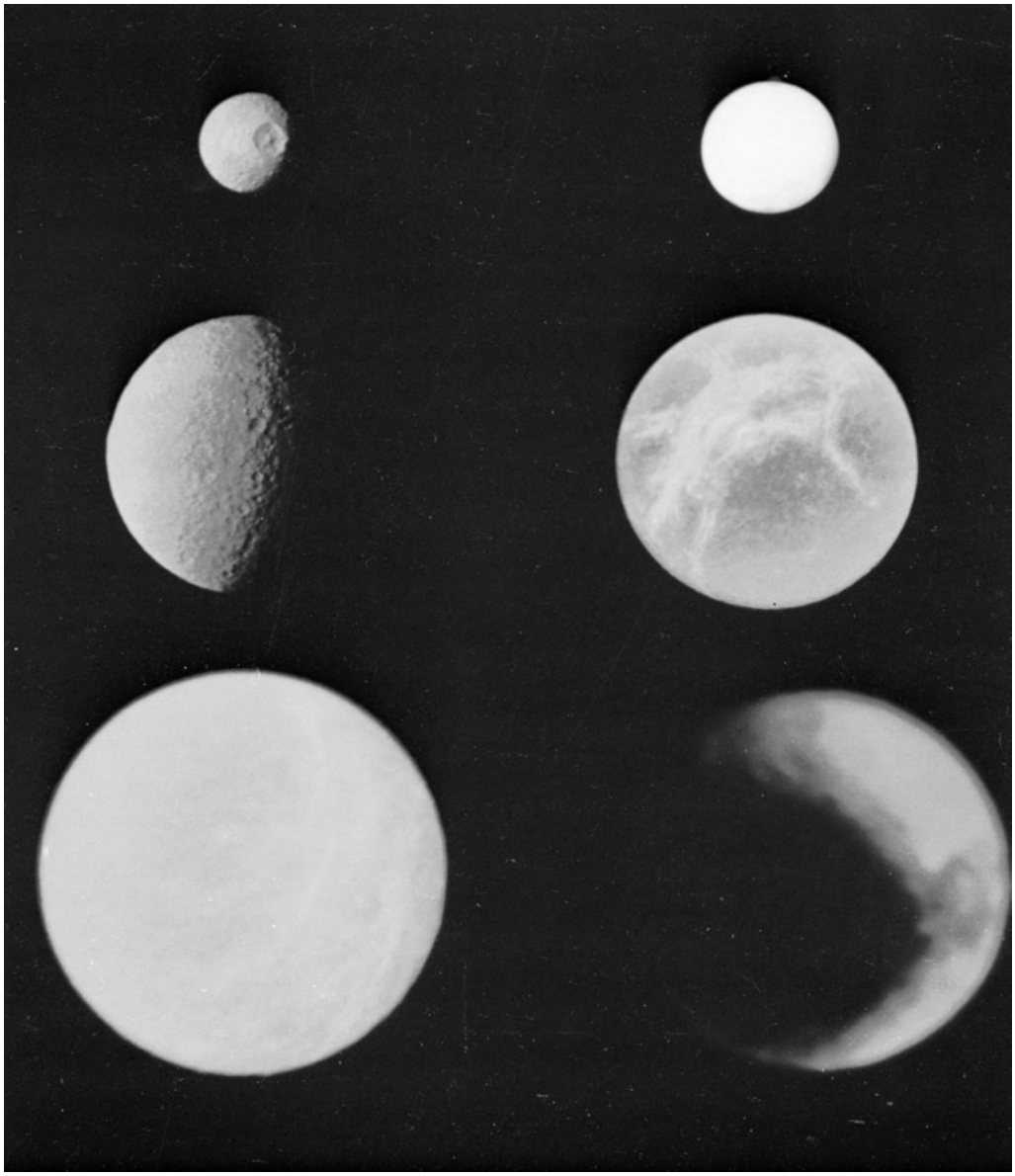


FIGURE 9 The six medium-sized icy Saturnian satellites. From the upper left in order of size: Mimas, Enceladus, Tethys, Dione, Rhea, and Iapetus.

The northern polar area of Titan is darker than the southern cap. Ground-based observations detected long-term variations in the brightness of Titan. Both observations may be related to the existence of 30-year seasonal cycles on Titan. Even though only a small fraction of the incident solar radiation reaches the surface, Titan is probably subjected to a slight greenhouse effect.

3. The Small Satellites

The Saturnian system has a number of unique small satellites. Telescopic observations showed that the surface of Hyperion, which lies between the orbits of Iapetus and Titan, is covered with ice. Because Hyperion has a visual geometric albedo of 0.30, this ice must be mixed with a significant amount of darker, rocky material. It is darker than the medium-sized inner Saturnian satellites, presumably because resurfacing events have never covered it with fresh ice. Although Hyperion is only slightly smaller than Mimas, it has a highly irregular shape (see Table 1). This suggests, along with the satellite's battered appearance, that it has been subjected to intense bombardment and fragmentation. There is also evidence for Hyperion being in nonsynchronous rotation—perhaps

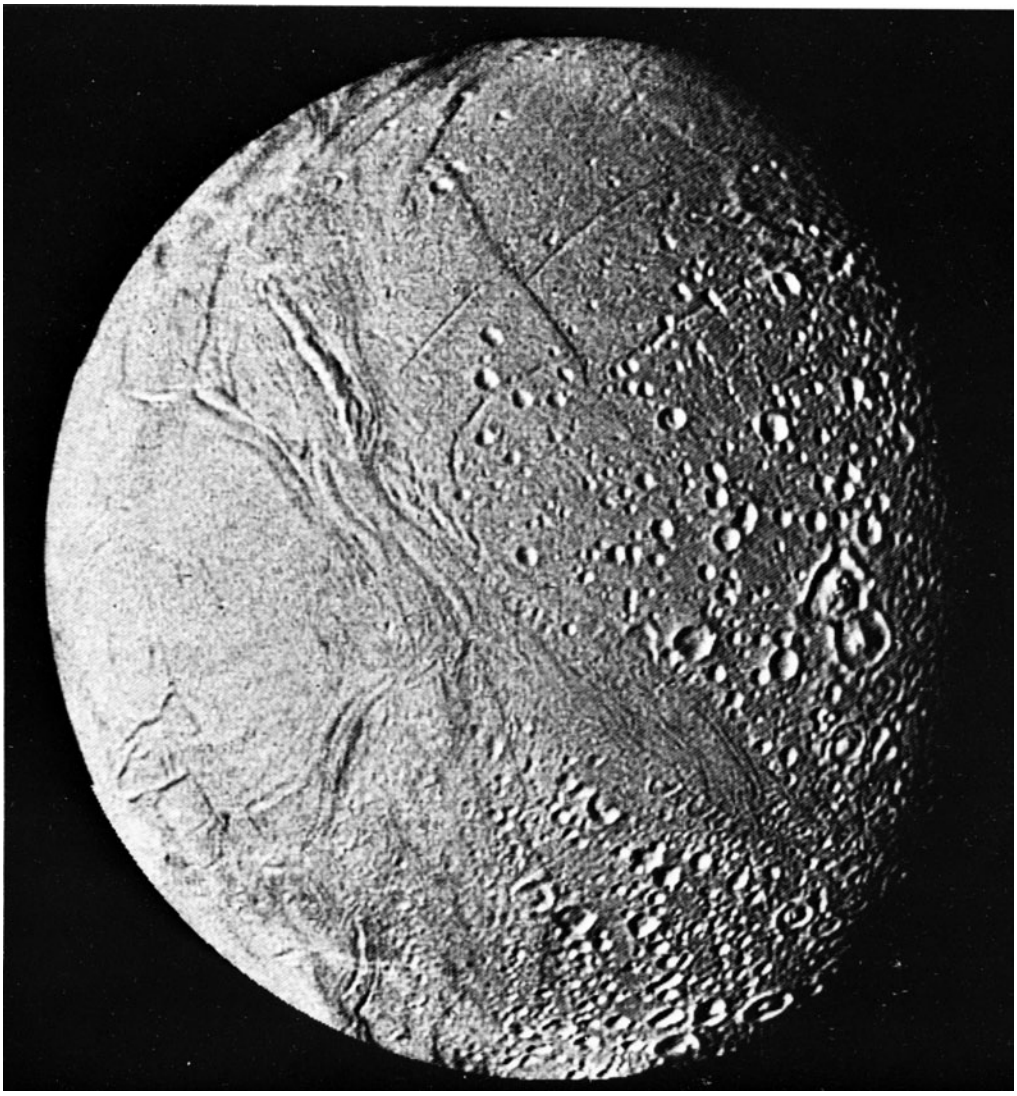


FIGURE 10 A *Voyager 2* photomosaic of Enceladus. Both the heavily cratered terrain and the recently resurfaced areas are visible.

a collision within the last few million years knocked it out of a tidally locked orbit.

Saturn's outermost satellite, Phoebe, a dark object (Table I) with a surface similar to that of carbonaceous asteroids, moves in a highly inclined, retrograde orbit. *Voyager* images show definite variegations consisting of dark and bright (presumably icy) patches on the surface. Although it is smaller than Hyperion, Phoebe has a nearly spherical shape.

Three types of small satellites have been found only in the Saturnian system: the sheparding satellites, the co-orbitals, and the lagrangians. All these objects are irregularly shaped (Fig. 13) and probably consist primarily of ice. The three shepherds, Atlas, Pandora, and Prometheus, are believed to play a key role in defining the edges of Saturn's A and F rings. The orbit of Saturn's innermost satellite Atlas lies several hundred kilometers from the outer edge of the A ring. The other two shepherds, which orbit on either side of the F ring, not only constrain the width of this narrow ring, but may cause its kinky appearance.

The co-orbital satellites Janus and Epimetheus, which were discovered in 1966 and 1978, exist in an unusual dynamical situation. They move in almost identical orbits at about 2.5 Saturn radii. Every 4 years the inner satellite (which orbits slightly faster than the outer one) overtakes its companion. Instead of colliding, the satellites exchange orbits. The 4-year cycle then begins over again. Perhaps these two satellites were once part of a larger body that disintegrated after a major collision.

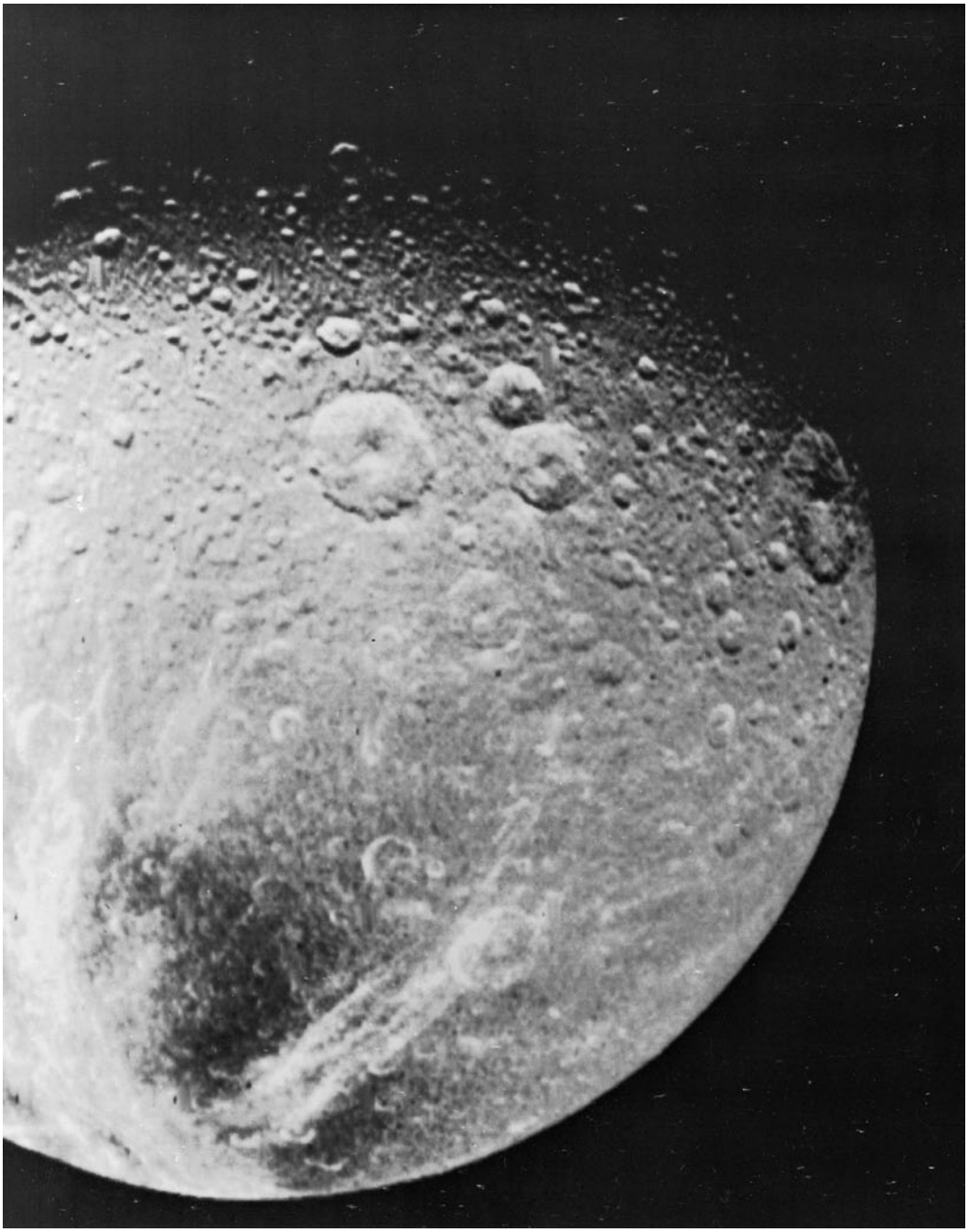


FIGURE 11 The heavily cratered face of Dione is shown in this *Voyager 1* image. Bright wispy streaks are visible on the limb of the satellite.

Three other small satellites of Saturn orbit in the Lagrangian points of larger satellites: one is associated with Dione and two with Tethys. The Lagrangian points are locations within an object's orbit in which a less massive body can move in an identical, stable orbit. They lie about 60°s in front of and in back of the larger body. Although no other known satellites in the solar system are Lagrangians, the Trojan asteroids orbit in two of the Lagrangian points of Jupiter.

In 1990, an 18th satellite of Saturn was discovered on archived *Voyager* images. The small body orbits in the Encke division, a gap between Saturn's C and D rings, and it is believed to cause wavelike structures observed in the rings.

FIGURE 12 A *Voyager 1* image of Titan showing the extended haze layer.

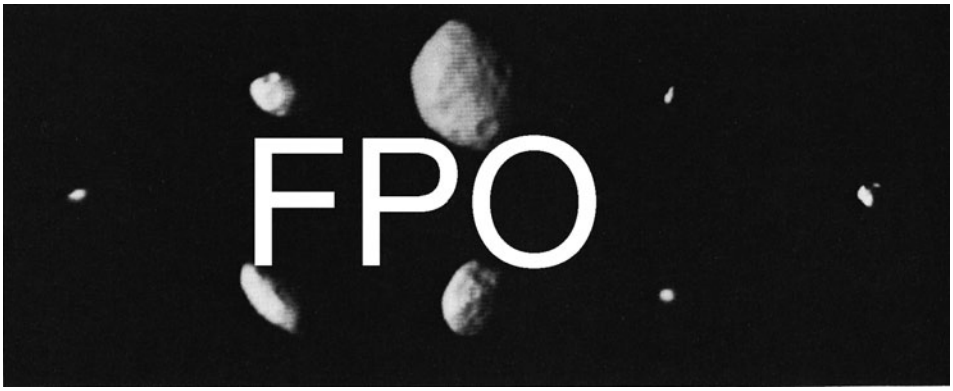


FIGURE 13 Eight small satellites of Saturn. They are, clockwise from far left, Atlas, Pandora, Janus, Calypso, Helene, Telesto, Epimetheus, and Prometheus.

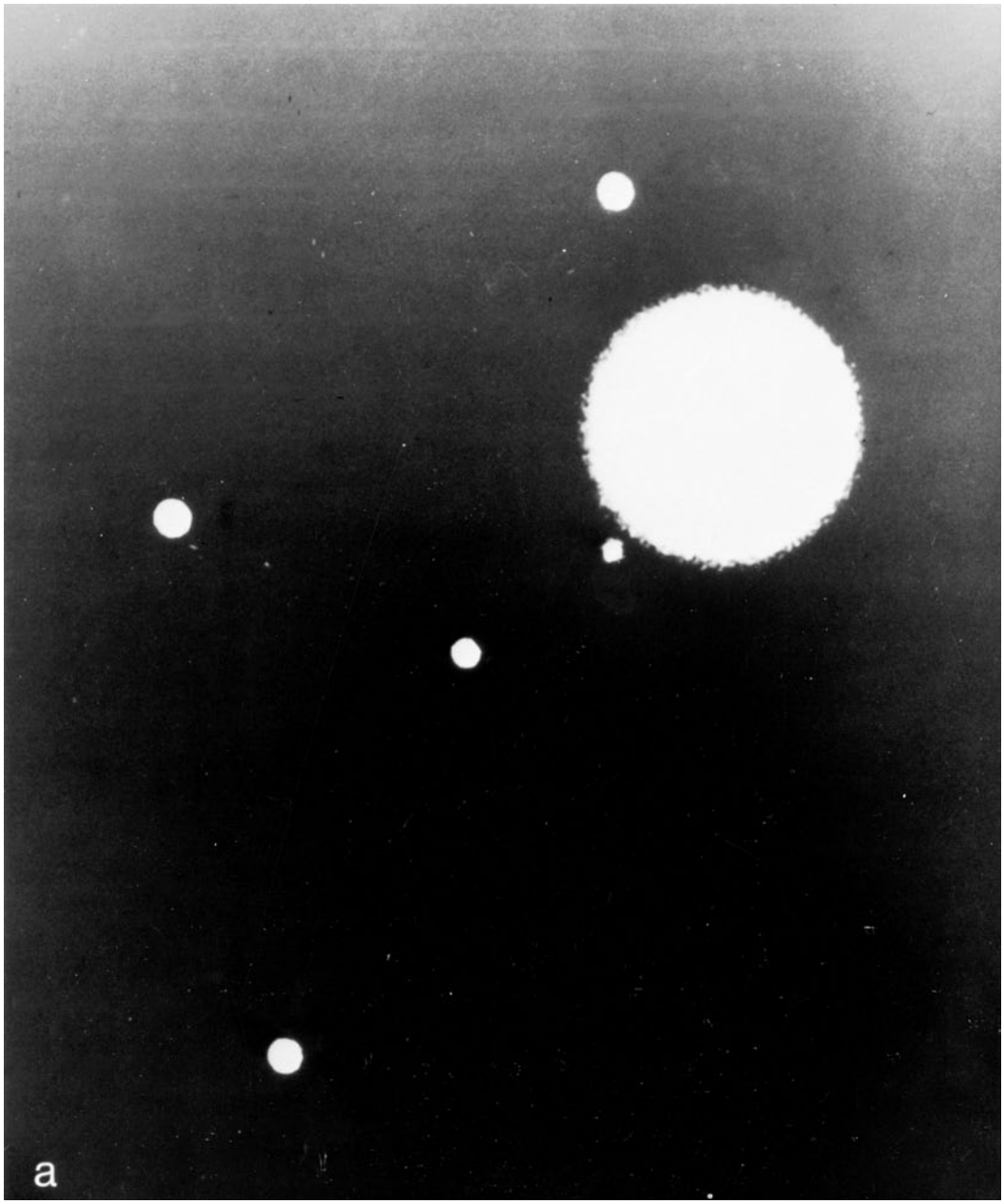


FIGURE 14a Telescopic view of Uranus and its five satellites obtained by C. Veillet on the 154-cm Danish–ESO telescope. Outward from Uranus they are as follows: Miranda, Ariel, Umbriel, Titania, and Oberon.

E. The Satellites of Uranus

The rotational axis of Uranus is inclined 98°s to the plane of the Solar System; observers on Earth thus see the planet and its system of satellites nearly pole-on. The orbits of Ariel, Umbriel, Titania, and Oberon are regular, whereas Miranda's orbit is slightly inclined. Figure 14 is a telescopic image of the satellites. The oretical models suggest the satellites are composed of water ice (possibly bound with carbon monoxide, nitrogen, and methane) and silicate rock. The higher density of Umbriel implies its bulk composition includes a larger fraction of rocky material. Melting and differentiation have occurred on some of the satellites. Theoretical calculations indicate that tidal interactions may provide an additional heat source in the case of Ariel.

Water ice has been detected spectroscopically on all five satellites. Their relatively dark albedos (Table I) are

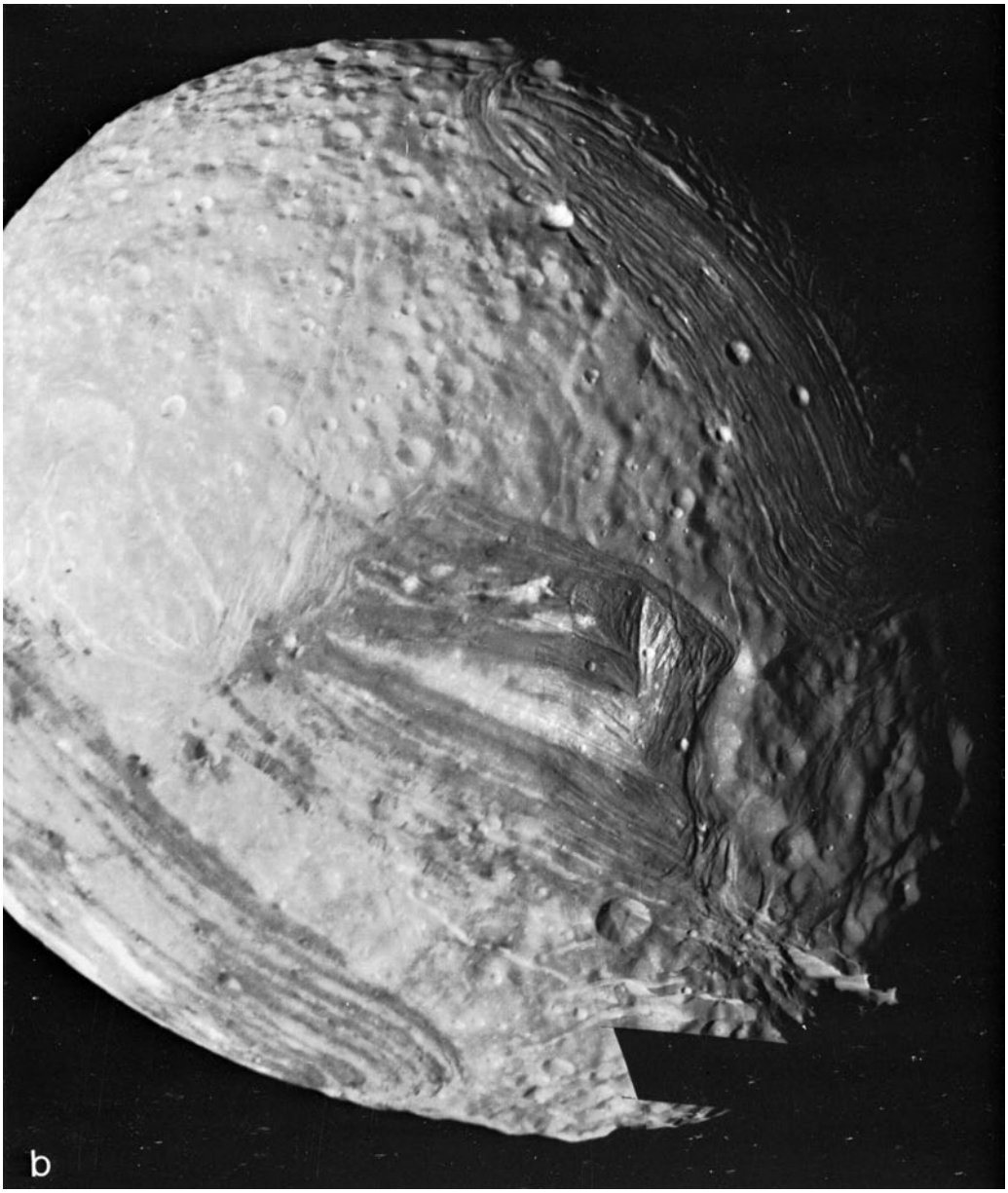


FIGURE 14b A mosaic of Miranda produced from images taken by the *Voyager 2* spacecraft at 30,000–40,000 km from the moon. Resolution is 560 to 740 m. Older, cratered terrain is transected by ridges and valleys, indicating more recent geologic activity.

probably due to surficial contamination by carbonaceous material. Another darkening mechanism that may be important is bombardment of the surface by ultraviolet radiation. The four outer satellites all exhibit large opposition surges, which may indicate that the regoliths of these objects are composed of very porous material.

The *Voyager 2* spacecraft encountered Uranus in January 1986 to provide observations indicating that at least some of the major satellites have undergone melting and resurfacing. One feature on Miranda consists of a series of ridges and valleys ranging from 0.5 to 5 km in height (Fig. 14b). Ariel, which is the geologically youngest of the five satellites, and Titania are covered with cratered terrain transected by grabens, which are fault-bounded valleys. Umbriel is heavily cratered and is the darkest of the major satellites, which indicates that its surface is the oldest. Oberon is similarly covered with craters, some of which have very dark deposits on their floors. The satellites are spectrally flat, with visual geometric albedos ranging from 0.2 to 0.4, which is consistent with a composition of water

ice (or methane-water ice) mixed with a dark component such as graphite or carbonaceous chondritic material.

Voyager 2 also discovered 10 new small moons, including two which act as shepherds for the outer (epsilon) ring of Uranus (Table I). These satellites have visual geometric albedos of only 4–9%. They move in orbits that are fairly regularly spaced in radial distance from Uranus and have low orbital inclinations and eccentricities. Five additional small satellites (10–20-km radii) were subsequently discovered by ground-based observers (see Table I).

F. The Satellites of Neptune

Neptune has an unusual family of satellites (Table I and Fig. 15). Its large satellite, Triton, moves in a highly inclined retrograde orbit. The orbit of Nereid, Neptune's

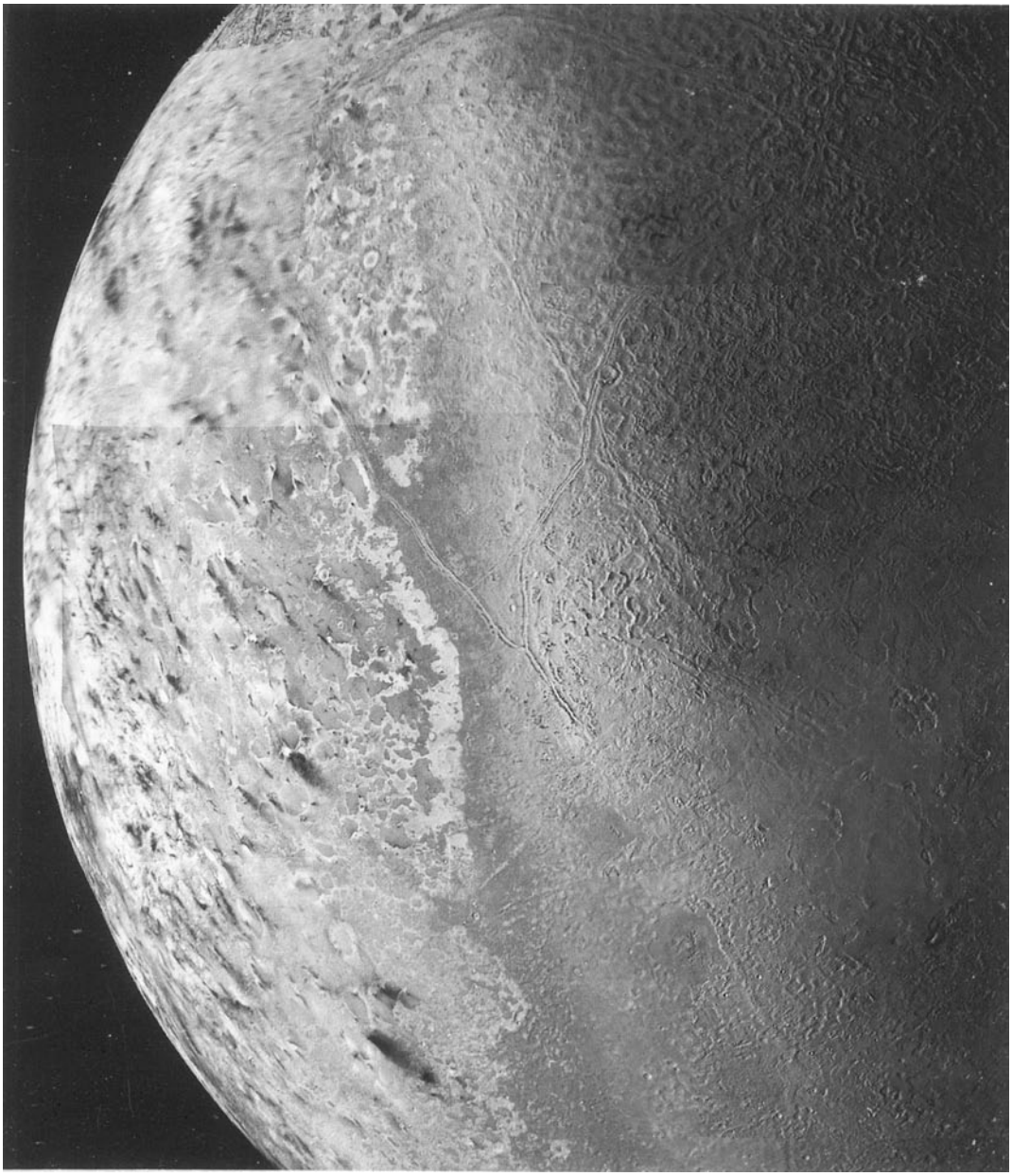


FIGURE 15 This mosaic of Triton was produced from about 12 high-resolution images obtained by the *Voyager* spacecraft on August 25, 1989. The bright polar cap is slightly red and may be composed partly of a currently evaporating deposit of nitrogen frost deposited during the previous winter. The darker terrain also has a reddish hue and consists of both smooth plains and areas with more complex surface texture. The dark streaks on the polar cap may be plume deposits of organic material from geyserlike eruptions.

only other satellite that was observed from Earth, is prograde, eccentric, and somewhat inclined. This situation implies an anomolous origin, perhaps involving capture of two remnant planetesimals by Neptune. The suggestion that Triton and Pluto were once both satellites of Neptune that experienced a near encounter, causing Triton to go into a retrograde orbit and the expulsion of Pluto from the system, is not plausible on dynamical grounds (see Section IV.G). During the *Voyager* encounter in 1989, six small satellites were discovered, including one (Proteus) larger than Nereid.

The bulk composition of Triton is water ice, methane, ammonia, and silicates; CO_2, N_2, CH_4, and CO frost have been detected on its surface. The center of the satellite may consist of a differentiated rock–ice core and a possibly liquid water–ammonia mantle. Because of the high inclination of Triton's orbit, the satellite undergoes a complex seasonal cycle of 165 years (the Neptunian year). *Voyager 2* observations of Triton show that this satellite is highly diverse and geologically active (Fig. 15). High-resolution images show a seasonal polar cap, which is composed primarily of nitrogen and smaller amounts of methane. Triton's surface has a reddish color, believed to result from deposits of organic compounds formed from photochemical reactions and particle bombardment of the veneer of methane and nitrogen ice which seems to cover its entire surface.

The relatively uncratered surface of Triton implies that it has experienced episodes of resurfacing (Fig. 15). Among the most unusual regions is the cantaloupe terrain (so named because of its morphological similarity to the skin of that fruit), which consists of pit-like depressions and crisscrossing ridges. Other areas of the satellite are covered by frozen lakes rimmed by layered terraces,

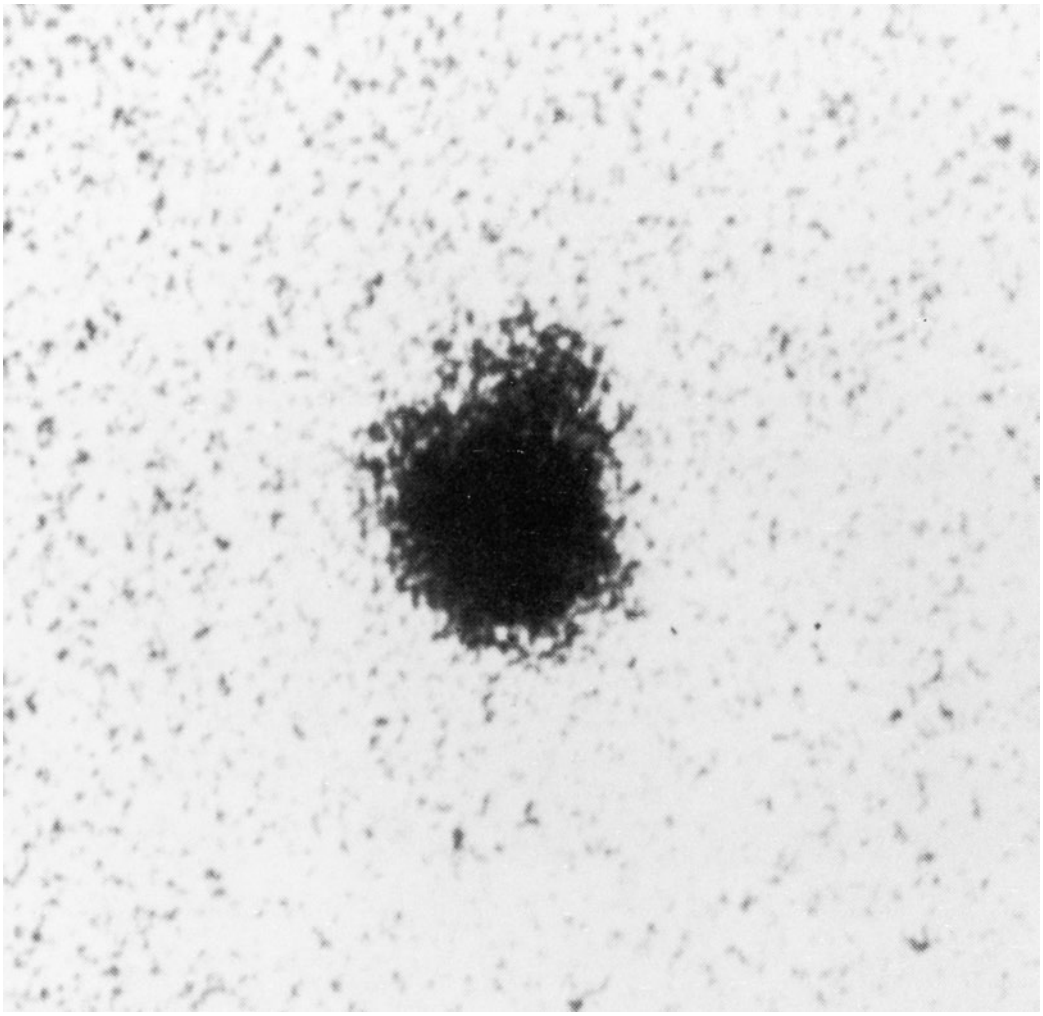


FIGURE 16 A negative image of the Pluto–Charon system obtained by J. W. Christy. Charon is the extended blob to the upper right of Pluto. (Photo courtesy of the U.S. Naval Observatory.)

indicating successive periods of flooding and freezing. In the seasonal cap, dozens of low-albedo, young (<1000 years) streaks were discovered. At least two actively erupting geyserlike plumes were observed. These features rise to a height of 8 km and disperse in a horizontal wind-entrained trail of more than 100 km. The geysers may be explosive eruptions of sun-heated nitrogen with admixtures of dark, organic material.

The surface temperature of Triton is 38 K—lower than that of any other body in the Solar System. Its tenuous atmosphere, including thin clouds which appear on Triton's limb in *Voyager* images, has a surface pressure of only 16 microbars and is composed primarily of nitrogen with about 0.01% methane. There is a haze layer in the upper atmosphere and an ionosphere with a peak concentration at 350 km above Triton's surface.

The small satellites of Neptune have low albedos, implying surface compositions similar to carbonaceous asteroids. Nereid has a geometric albedo of 0.14, more consistent with an ice–silicate surface; ice has been detected on its surface. Its unusual orbit means that the satellite is probably not in synchronous rotation. Because albedo variegations on it appear to be less than 10%, no rotational period has been deduced from its light curve.

G. The Pluto–Charon System

Soon after Pluto was discovered in 1930, scientists conjectured that the planet was an escaped satellite of Neptune. More recent calculations have shown that it is unlikely that Pluto could have acquired its present amount of angular momentum during an ejection event. Pluto is most likely a large planetesimal dating from the formation period of the solar system.

In 1978, Pluto was shown to have a moon of its own, which was named Charon (Fig. 16). Charon appears in specially processed high-quality images as a fuzzy mass orbiting close to Pluto. More recent images obtained by HST show the two bodies that are clearly separated. It is more massive in comparison with its primary than any other satellite: its mass is about 7% that of Pluto, if one assumes their densities are equal. The satellite has rocky core and a water–ammonia ice mantle. Spectral measurements show water ice absorption bands, but there is no evidence for methane on its surface. Apparently methane (which is abundant on the surface of Pluto) escaped from Charon's weaker gravitational field. Charon's surface should be sufficiently cold for nitrogen, which is difficult to detect spectrally, to exist in solid form. The hemisphere of the satellite which faces Pluto is neutral in color, whereas the opposite hemisphere has a reddish hue; this dichotomy implies surficial compositional variegations. Between 1985 and 1990, Pluto and Charon underwent a series of mutual eclipses, which allowed a more accurate determination of their radii, masses, orbital elements, and individual surface properties.

SEE ALSO THE FOLLOWING ARTICLES

CELESTIAL MECHANICS • IMPACT CRATERING • LUNAR ROCKS • MOON (ASTRONOMY) • PLANETARY ATMOSPHERES • PLANETARY GEOLOGY • POLARIZATION AND POLARIMETRY • PRIMITIVE SOLAR SYSTEM OBJECTS: ASTEROIDS AND COMETS • RADIOACTIVITY • RADIOMETRY AND PHOTOMETRY • SOLAR SYSTEM, GENERAL

BIBLIOGRAPHY

Beatty, J. K., Petersen, C. C., and Chaikin, A. (eds.) (1998). "The New Solar System," 4th ed. Cambridge, UK, Cambridge Univ. Press.

Belton, M. J. S., and the Galileo Science Teams (1996). *Science* **274,** 377–413.

Bergstralh, J. T., Miner, E. D., and Matthews, M. S., eds. (1991). "Uranus," Tucson, University of Arizona Press.

Burns, J., and Matthews, M., ed. (1986). "Satellites," Tucson, University of Arizona Press.

Cruikshank, D. P., Matthews, M. S., and Schumann, A. M. (eds.) (1996). *Neptune and Triton*. University of Arizona Press, Tucson.

Gehrels, T., ed. (1984). "Saturn," Tucson, University of Arizona Press.

Hartmann, W. K. (1988). "Moons and Planets," 4th ed., Brooks/Cole.

Morrison, D., ed. (1982). "The Satellites of Jupiter," Tucson, Univ. of Arizona Press.

Stone, E., and the Voyager Science Teams (1989). *Science* **246,** 1417–1501.

Primitive Solar System Objects: Asteroids and Comets

Lucy-Ann McFadden
University of Maryland

Daniel T. Britt
University of Tennessee

GLOSSARY

Albedo Measurement of the fraction of sunlight that a surface reflects.

Aphelion Point at which a solar system body is farthest from the sun in its orbit.

Apparition Time at which an object in the solar system can be viewed from the earth in the nighttime sky.

Astronomical unit (AU) The mean distance from the earth to the sun, 1.496×10^8 km.

Centaurs Asteroids with semimajor axis between Jupiter and Neptune. These objects are probably cometary in composition and a source for short-period comets.

Eccentricity Parameter of an elliptical orbit describing how much the orbit deviates from that of a circle.

Ecliptic Plane defined by the earth and the sun.

Edgeworth-Kuiper belt A vast reservoir of primitive objects that are the source of short-period comets. This belt extends from the orbit of Neptune out to approximately 1000 AU.

Inclination Orbital element that is the angle measured between the ecliptic and the orbital plane of an object.

Main asteroid belt A zone between the orbits of Jupiter and Mars where the vast majority of asteroids are located.

Meteorite Rock of extraterrestrial origin that has survived passage through the earth's atmosphere and reached the surface of the earth.

Perihelion Point in the orbit of a solar system body at which it is closest to the sun.

Plasma Electrically neutral gas consisting of charged particles such as ions, electrons, and neutrals. A plasma has a high temperature but its density is very low.

Semimajor axis Distance at which an orbiting body is halfway between its aphelion and perihelion.

Solar nebula Cloud of gas and dust that surrounded the sun during its early stages of formation. The planets formed from material in this cloud.

Solar wind Constant radial flow of energetic charged particles from the sun.

Trans-Neptunian objects Asteroids with semimajor axis outside the orbit of Neptune. This is the inner edge of the Edgeworth–Kuiper belt that is thought to extend out

as far as 1000 AU and is a major source of short-period comets.

ASTEROIDS AND COMETS are among the smallest objects in the solar system. They are defined as small, naturally formed solid bodies that orbit the sun. The word comet comes from the Greek *kometes*, meaning "long-haired," referring to the long streams or tails of gas and dust which flow off a comet during its passage into the warmer regions of the inner solar system. This activity makes comets often easily visible, while asteroids of the same size and distance are much harder to see. Asteroids were only discovered long after the invention of the telescope and are rarely bright enough to see with the naked eye. The outflow of gas and dust is the defining difference between asteroids and comets. Comets have a detectable outflow some time during their orbit and asteroids do not.

Asteroids and comets have a range of compositions that include materials that have remained relatively unaltered since accretion from the solar nebula as well as materials that have been extensively heated, melted, and processed. Scientists refer to unaltered material as "primitive" because it closely resembles the starting composition of the solar system. The goals of the study of asteroids and comets are to determine how these objects formed, what they can tell us about the early composition of the solar system, and how those compositions have changed over the age of the solar system. From the study of asteroids and comets, as well as the study of the larger planets, we expect to piece together the history of our solar system. In this article the current state of our knowledge of the asteroids and comets is presented as well as some issues to be resolved in the coming decades.

I. INTRODUCTION

The asteroids and comets are planetesimals (literally, "small planets") that failed to condense into a full-sized planet. As of November 2000 the number of asteroids with known and catalogued orbits was approaching 20,000 and there are over 105,000 objects that have been observed at least once. During the 1990s we had our first closeup views of asteroids with a series of spacecraft missions, observations by the Hubble Space Telescope (HST), and high-resolution imaging with ground-based radar. A landmark in asteroid exploration was reached in 1999 when the NEAR-Shoemaker spacecraft became the first to enter orbit around an asteroid. This mission greatly expanded our knowledge and understanding of these "small planets." Shown in Fig. 1 are the four asteroids visited so far by spacecraft.

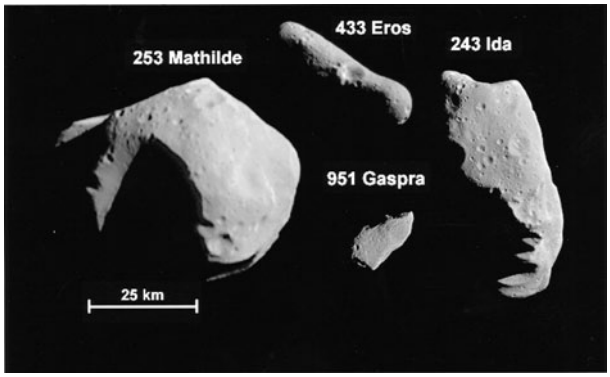


FIGURE 1 The family portrait of asteroids. Four of the asteroids visited by spacecraft viewed at the same scale. Note the irregular shapes and rough surfaces of these objects. [Courtesy NASA/JHUAPL/JPL.]

The total mass of the asteroid belt is approximately 1.5×10^{24} g, about one billionth the mass of the sun. About 30% of this mass is contained in one asteroid, the largest, named 1 Ceres. Asteroids are given a number and a name. The numbers are assigned numerically as the orbits of the asteroids are determined precisely enough for their position to be predicted with enough certainty that it can be found at a later apparition. By tradition, the discoverer has the right to choose a name. Until an asteroid is named, they are given a temporary designation that contains the year in which it was discovered and a two-letter code (sometimes followed by additional numbers). The first letter marks when the asteroid was discovered during the year, with the year divided into sequential half-month intervals (the letters I and Z are not used). The second letter denotes the order in that particular 2 weeks in which the asteroid was discovered. For example, an asteroid named 1999 AB was the second asteroid (referring to the B) discovered in the first half (referring to the A) of January 1999. If more than 25 asteroids are discovered during the 2-week period the sequence of second letters is repeated with a subscript "1" (the letter I is not used in this sequence to avoid confusion with a number). The next 25 discoveries use the sequence of second letters with the subscript "2," and so on for each batch of 25. An asteroid designated $1999AB_4$ would be the 102nd asteroid discovered during the first half of January 1999. Some designations hint at a frantic pace of discovery. Asteroid 2000 EE_{104}, for example, was closely tracked because it has an orbit that brings in into a series of near approaches with Earth. It was the 2605th asteroid assigned a preliminary designation in the first 2 weeks of March 2000, which must have been a very productive observing period! Once an asteroid receives a number designation, tradition allows the discover to name the asteroid. Wide latitude in choosing names is also a tradition and asteroid names include figures from

mythology (3 Juno), countries (2713 Luxembourg), musicians (4148 McCartney, 1815 Beethoven), authors (6984 Lewiscarroll), fictional characters (2309 Mr. Spock), and even asteroid scientists (3066 McFadden, 4395 Danbritt).

Comets are primarily icy bodies with highly elliptical orbits. With their large orbital periods, they stay at great distances from the sun for most of their lifetimes. They are typically observed only when they pass relatively close to earth. There are records of observations of approximately 1500 comets, but there are an estimated 10^{10}–10^{12} comets in the solar system. By convention, the comets are divided into three groups, long-period comets, with periods greater than 200 years, short-period comets, with periods less than 200 years, and Jupiter family comets, with periods less than 20 years. Comets are named after their discoverer (or discovers) and given a designation (much like asteroids) that includes the year of discovery, an uppercase code identifying the half-month of discovery, and a consecutive number indicating the order of discovery during the half-month. The designation is prefixed by either P/ for a short period comet, C/ for other comets, or D/ for a defunct comet. For example, C/1996 B2 (Hyakutake) was the second long-period comet discovered in the second half of January 1996. To avoid having a new designation for each return of a periodic comet, once they are observed to return the "P/" is preceded by a sequence number and the discovery designation is dropped. Comet Halley is known simply as 1P/Halley. Since some people are prolific comet discoverers, a number often follows names. For example, the periodic comet 135P/Shoemaker–Levy 8 was the eighth comet discovery by the Shoemaker–Levy team.

II. ASTEROIDS

A. Orbits

An orbit is defined by six parameters or orbital elements. Only three, the semimajor axis a, the eccentricity e, and the inclination i, are needed to describe the major aspects of the asteroid population. The bulk of asteroids have semimajor axis in the 2.2- to 4.3-AU range, which defines the main asteroid belt. Most of these asteroids have eccentricities between about 0.03 and 0.4 and inclinations between 0° and 30°.

1. Asteroid Belts and Planet-Crossing Asteroids

The vast majority of known asteroids orbit the sun in the main asteroid belt between Mars and Jupiter. However, this is not the only asteroid belt and asteroids do not all stay within their belts. A small fraction of main-belt asteroids have orbits that cross the orbits of the inner planets. These are called near-earth asteroids, and are a major source of the meteorites that occasionally fall to earth. Asteroids with orbits that approach the earth, passing inside the perihelion of Mars, are called Amors. Those that cross the orbit of the earth are called Apollo asteroids. A third group of near-earth asteroids (NEAs) that have orbits crossing that of the earth's but that do not travel further from the sun than the earth's aphelion are called Aten asteroids, after the name of the first asteroid discovered with this type of orbit. There are currently over 1200 known NEAs.

An estimated 100–1000 tons of extraterrestrial material, ranging from submicrometer-sized dust to boulders weighing many tons, reach the earth's surface every day. Some of this material is derived from the near-earth asteroids. Larger asteroids also collide with earth. It is estimated that an asteroid that is 1 km in diameter or larger will collide with the earth on the average about once every 1.5–2 million years. Most near-earth asteroids are not on collision courses with the earth, but their orbits may be near enough to the earth to be used for building and manufacturing in space in the future. Calculations of the orbits of near-earth asteroids indicate that they have lifetimes in their present orbits for 10–100 million years. They are "culled" from near-earth space either a collision with the inner planets, collisions with other asteroids, or ejection from the solar system by gravitational perturbations due to near misses with Mars, earth, or Venus. Since the age of the solar system as determined from radioactive dating of meteorites is at least 4.5 billion years, there must be a source of these asteroids to keep replenishing that part of the population that is lost. Some of the near-earth asteroids are probably the solid remains of comets that have exhausted all their gases and now orbit the sun in planet-crossing orbits. The major source of near-earth asteroids are fragments from large asteroids in the main asteroid belt that are perturbed into planet-crossing orbits by gravitational interactions with Jupiter (see Section II.A.2, on Kirkwood gaps). One of the major goals of near-earth asteroid studies is to use these objects as a source of information on the composition and processes of the larger asteroids in the main belt. Other equally important goals are to determine their importance as economic resources as well as hazards from near-earth asteroid impacts on earth.

The outer solar system also has its belts of asteroids and its planet-crossing objects. Asteroids with semimajor axis between Jupiter and Neptune are called Centaurs and these objects regularly interact gravitationally and collisionally with the massive gas giant planets. An additional belt of asteroids was first observed in the 1990s outside the orbit of Neptune and is called "trans-Neptunian." This is the inner edge of the Edgeworth–Kuiper belt that is thought

to extend out as far as 1000 AU. These outer solar system "asteroids" probably have cometary compositions and are almost certainly the sources of comets, but as long as they stay in the cold outer solar system, they show no outflow of gas and dust and thus look like primitive "asteroids." New members of these groups are being discovered at a rapid rate and as of November 2000 there are 63 Centaurs and 345 trans-Neptunians.

The asteroids of the outer solar system, the Centaurs and trans-Neptunian objects, also interact with the planets. Orbits with semimajor axis greater than 45 AU are thought to be stable, but within 45 AU, studies show a number of chaotic zones as well as some stable orbits. Perturbations on trans-Neptunian asteroids are thought move them into the solar system, supplying the Centaurs with fresh objects and acting as the reservoir for short-period comets (see Section II).

2. Kirkwood Gaps

Asteroid orbits in the main belt are not distributed randomly. There are regions in which asteroids are relatively numerous and regions in which they are nearly absent. This structure, shown in Fig. 2, is most apparent when the population of asteroids is plotted as a function of semimajor axis. Asteroid orbits were probably initially randomly distributed throughout the main-belt region, but the gravitational forces of Jupiter removed asteroids in certain orbits. These regions of relatively few asteroids are called the Kirkwood gaps, after the 19th-century American astronomer who first recognized their existence. The Kirkwood gaps are located at distances from the sun where the orbital period of the asteroid is an integer multiple of the orbital period of Jupiter. When these asteroids and Jupiter line up with respect to the sun, Jupiter's mass exerts a gravitational force on the asteroid. For example, one of the major gaps is at the 1:3 resonance with Jupiter. Asteroids with this orbit would have nonrandom close approaches with Jupiter every third orbit. These periodic perturbations would alter the orbit of these asteroids, typically increasing the eccentricity to the point where they cross the orbits of other asteroids and eventually the inner planets. These orbital changes increase the chance of a collision and decrease the asteroid's lifetime. The overall effect is to drain objects out of the resonances and create a gap in the distribution of asteroids in the main belt. Asteroids once located near the Kirkwood gaps are probably one of the major sources of near-earth asteroids. (See Sections 1.D and 1.E.)

Examination of Fig. 2 shows that there are two resonances that are populated by asteroids. These are the Hilda and Trojan regions. The Hildas at a semimajor axis of 4.0 AU are in a 3:2 resonance with Jupiter. However, orbital librations prevent these asteroids from making close approaches to Jupiter; thus, they are not systematically perturbed and their orbits remain stable. The

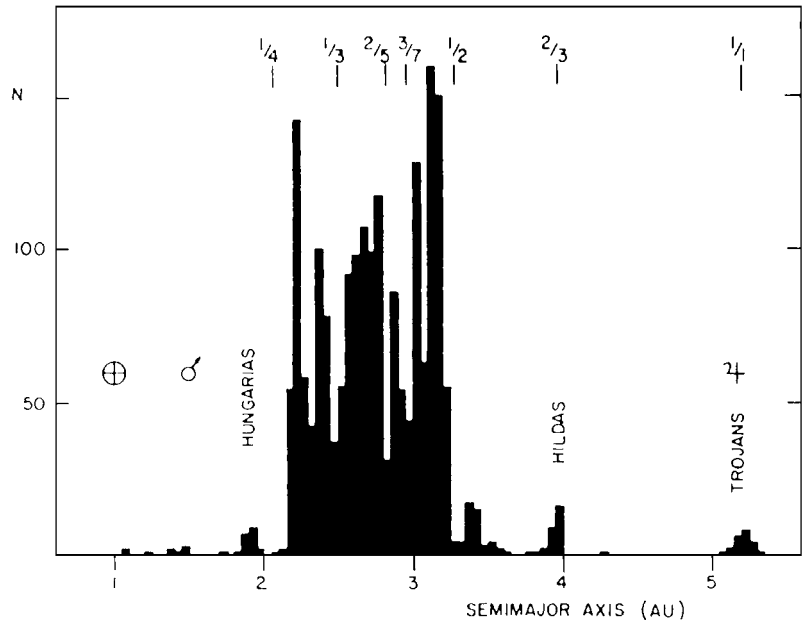


FIGURE 2 The distribution over heliocentric distance of the first 1978 numbered asteroids, in increments of 0.05 AU. Fractions indicate the ratio of orbital periods of the asteroids to that of Jupiter, which form the dynamical resonances called Kirkwood gaps. The orbits of the planets are indicated by symbols at 1, 1.5, and 5.2 AU. [Figure complied by B. Zellner and published in NASA Conference Publication 2053.]

Trojans actually are in the same orbit as Jupiter, but are located at the preceding and following Lagrangian points of the Jupiter–sun system. At these points, the gravitational forces on the asteroids due to Jupiter and the sun are equal to the radial forces of the asteroids' motion and the orbits are stable.

3. Hirayama Families

Some asteroids are found to have nearly identical orbital elements. It is thought that these groups are remnants of a collision or fragmentation of large (∼100-km-diameter) asteroids. The early 20th-century Japanese astronomer K. Hirayama identified these groups of asteroids, and they bear his name. If they are collisional remnants of a larger asteroid, then the study of their fragments will provide the equivalent of a stratigraphic view of the interior of the parent asteroid. Physical studies of a few asteroid families have shown that the members of some families have similar photometric properties that differ from the surrounding nonfamily population of asteroids. These results support the theory that some families at least resulted from the collision disruption of a single parent body and that their composition is homogeneous. There are indications from reflectance spectroscopy and photometric surveys that some other smaller families may be groupings of unrelated objects. We conclude that some alleged collisional families are actually dynamical groupings isolated from the rest of the asteroid belt by complex orbital resonances.

B. Size

The size of an asteroid is fundamental information from which the density and albedo of the object can be derived. In addition, the size distribution of asteroids in the main belt is related to the formation mechanism and subsequent evolution of the asteroid population. Asteroid sizes range from the 960-km diameter of 1 Ceres down to objects of a few meters in diameter. The minimum sizes of observed asteroids are not a function of the population, but are set by the constantly changing limits of observational technology. There are only three asteroids with diameters larger than 500 km. Asteroids with diameters between 50 and 500 km number more than 300. The number of asteroids of smaller diameter increases exponentially but not uniformly, which would indicate that the population was in equilibrium in terms of collisions. From these observations one must determine what portion of this distribution represents the initial distribution remnant from the formation of the solar system and which portion is the product of subsequent collisional destruction of the asteroid belt. Different processes may represent different size distributions; thus, knowledge of this size distribution is important to understanding the history of the main belt. Examining the size distribution of asteroids as a function of their orbital elements or taxonomic type (to be discussed) may also reveal the nature of the dynamical processes presently active in the asteroid belt. Indeed, studies have shown that the size distributions differ in different regions of the asteroid belt and for different taxonomic types and families of asteroids in the same region.

1. Thermophysical Models

Although we have directly imaged a few asteroids, most of the information we have on the size of asteroids is based on thermophysical models using measurements of asteroid "heat" at 10 and 20 μm and their reflectance in the visible or near-infrared. Assuming the surface is in equilibrium with the incident sunlight, the sum of the reflected and emitted radiation must equal the total solar flux on the asteroid surface. The reflected component, measured by its visible brightness, is proportional to the product of the geometric albedo p and cross section of the asteroid. The measured thermal emission is proportional to the product of the absorbed incident sunlight $(1 - A)$, where A is the total albedo integrated over all wavelengths [called the Bond albedo, $A = q(\lambda)p(\lambda)$], and the cross section. With measurements of the reflected and emitted radiation, and assumptions concerning the physical properties of the surface (i.e., values of the emissivity, its angular distribution, and the relationship between the geometric and total albedo) the size and albedo of an asteroid can be determined with an accuracy of 10–20%.

In 1983, the Infrared Astronomical Satellite (IRAS) was launched to make a survey of celestial objects at infrared wavelengths. Contained in this data set are the infrared fluxes of 1811 asteroids and 25 comets. Thermal flux densities at 12, 25, 60, and 100 μm and derived diameters and albedos make this the largest, most complete, and least-biased survey of asteroids to date. This survey also discovered many small, dark asteroids in the outer portions of the main belt.

2. Occultation Diameters

One of the most accurate methods of measuring the diameter of an asteroid, short of a space mission, is by observing stellar occultations, which occur when an asteroid passes across in front of a star as seen from the earth. Essentially observers are mapping the asteroid's "shadow" as cast on the earth by the star. A network of observers record the time the star "winks out" at different locations on the earth and it is possible for these data to reconstruct the space of the asteroid's shadow.

Although stellar occulations of asteroids are quite common, a number of challenges limit the chances for

observations. First, the angular diameters of all asteroids are fairly small, so the path that the star's shadow traverses as projected on the surface of the earth is also small. Even for large asteroids, the path is only a few hundred kilometers wide. Observers and telescopes need to be placed across the path of the occultation. But predicting where the shadow falls is challenging. Asteroid orbits always have small errors. These errors are trivial in celestial terms, but can amount to hundreds of kilometers in the predicted occulation path. The errors can be reduced by observations a few days prior to the occultation, but as a result, path predictions can change by hundreds of kilometers days or even hours before the event. With good coverage of the occultation path, both the diameter and projected shape of the asteroid can be determined to an accuracy of a few percent. Most of these observations are done by amateur astronomers under the aegis of the International Occultation and Timing Association (http://www.anomalies.com/iotaweb/). This is invaluable work since these occultation diameters are the "ground-truth" tests for the thermophysical models. Improving the accuracy of the models, which can then be applied to the much larger IRAS and thermal observations dataset, is critical to our understanding of asteroid shape, density, and internal structure.

C. Shapes and Rotation Rates

By monitoring the reflected sunlight from the surface of Ceres as a function of time as the asteroid rotates, we see about a 10% change in brightness called a light curve. This variation indicates one of three things: (1) that there is a change in the cross section or shape of the asteroid as it rotates, (2) that the surface reflectance properties vary across the surface, or (3) that the asteroid has a companion body orbiting around it, eclipsing the primary asteroid and changing the amount of reflected light seen when the companion is at different locations in its orbit. A spherical body with no color variations would produce no light curve. Any rotating, smooth spheroid (that is not perfectly spherical) with a homogeneous surface composition would have a singly periodic light curve (two peaks of equal magnitude). A spherical asteroid with color variations across its surface would also have a singly periodic light curve, its period being equal to the asteroid's rotation rate. In order to estimate the shapes and relative albedos of eclipsing asteroids, at least one of the components has to be elongated and/or have different albedos, otherwise there would be no change in the light curve. Doubly periodic light curves occur in cases of irregular shape and/or variations in albedo across the surface. These types of light curves are the most common in asteroids. It is believed that most of the asteroid brightness changes are due to their irregular shape because there is little change in both the degree of polarized light and overall color as observed with rotation. This view has generally been confirmed by the spacecraft close encounters and radar observations. The asteroids examined close up exhibit very little spectral variation across their surfaces, but show extremely rough and irregular surfaces. The asteroid "family portrait" in Fig. 1 shows the huge craters on Mathilde and the sharp scarps on Ida. However, as shown in Fig. 3, HST images of 4 Vesta and other observations of some of the larger asteroids have shown evidence of compositional

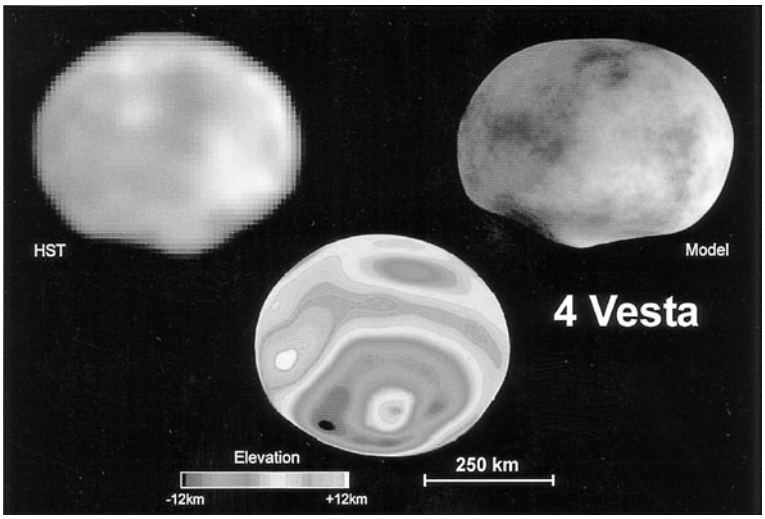


FIGURE 3 Asteroid 4 Vesta as viewed by the Hubble Space Telescope. Shown in the upper left is the raw HST image. A processed and modeled version is shown in the upper right and the lower image details is a color-encoded elevation map of the asteroid. [Hubble Space Telescope—Wide Field Planetary Camera 2 PPC97-27—STScI OPO—September 4, 1997—Courtesy of P. Thomas (Cornell University), B. Zellner (Georgia Southern University), and NASA.]

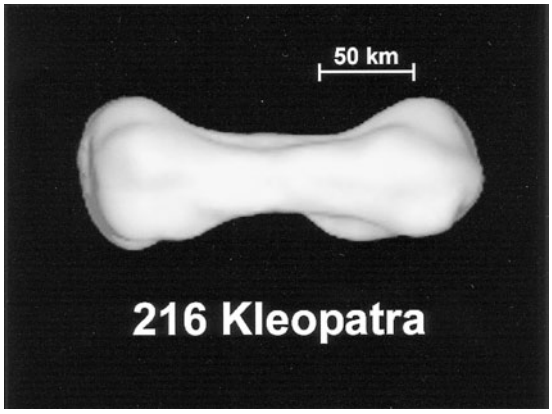


FIGURE 4 Asteroid 216 Kleopatra as reveled by radar. This asteroid is one of the most irregularly shaped objects in the solar system and resembles a 200-kilometer long dog bone. [Courtesy NASA/JPL.]

heterogeneity across their surfaces. Radar images of Asteroid 216 Kleopatra are shown in Fig. 4. This object is one of the most irregular asteroids seen so far and resembles a 200-kilometer-long dog bone.

Rotation rates of asteroids range from a few minutes to many days and reflect the processes of formation and subsequent collisional evolution. Because of the low gravity of asteroids, rotation periods of less than 2 hr will cause the centrifugal acceleration at the equator to be greater than the gravitational acceleration. The result would be that any loose material (rocks, boulders, or soil-sized particles) would be flung off the object. Very few asteroids rotate this fast and those that do are typically very small. These objects are probably single, solid fragments from previous collisions. The strength of the asteroid material plays an indirect role in determining the rotation rate. As an asteroid made of weak material increases in rotation rate from collisions, it tends to deform from a spherical body. This deformation results in a slower spin rate. The combined effects of collisions resulting in an increase in rotation period and the slowing down from loss of ejecta and deformation have probably produced the observed average rotation periods for asteroids (10 km diameter or larger) of between 5 and 12 hr. It is interesting that no asteroids larger than about 1 km in diameter spins fast enough to require some tensile strength to hold the body together. The implication of this observation is that asteroids are probably heavily fractured and likely do not have much, if any, internal strength. Most asteroids may be so-called "rubble piles" of shattered rock fragments held together by mutual gravity.

D. Taxonomy

Based on what we know from meteorites, asteroids can range compositionally from the equivalent of soggy dirt clods to almost solid iron. These objects have been classified into 18 different types according to their color and albedo properties. These types are summarized in Table I along with their suggested meteorite analogs based on compositions that are inferred from reflectance spectra (see Section 1.E). Note that 8 of the 18 asteroid types do not have meteorite analogs and their inferred surface mineralogy is highly speculative. The distribution of major asteroid types by orbital semimajor axis is displayed in Fig. 5. The asteroid types tend to be roughly "stratified" with distance from the sun. In the inner regions of the main belt, the E, R, and S types predominate. These types are thought to be composed of mostly high-temperature silicates that were probably either heated and differentiated or at least strongly metamorphosed. The C-type asteroids dominate between 2.5 and 3 AU. The outer portion of the main belt is composed of C-, P-, and D-type asteroids. The C types, with their low albedo, are optically most similar to carbonaceous chondrite meteorites, which are composed of hydrous silicates, iron- and magnesium-rich silicates, and oxides with varying amounts of elemental carbon. These meteorites are significant because the ratios of their nonvolatile elements are the same as those in the sun. The P- and D-type asteroids are probably composed of carbon-rich assemblages. The spectra of these types are consistent with carbon-based compounds called kerogens, which are a plausible analog to outer solar system material. These objects are probably compositionally similar to comets and may retain water ice and frozen volatiles because they orbit in relatively cold regions of the outer asteroid belt. The zonal structure of the asteroid has probably existed since the formation of the solar system. The inferred compositions of the asteroids are consistent with the observed trend in composition seen among the major planets. Those forming closer to the sun are enriched in iron and silicate minerals that are stable at

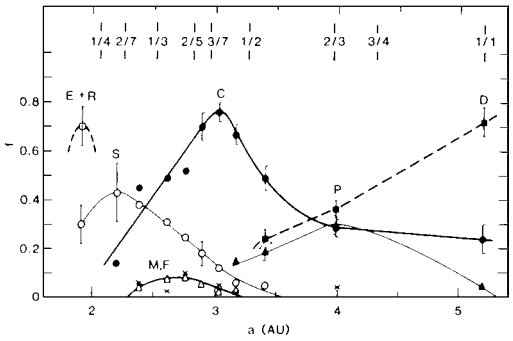


FIGURE 5 Distribution of asteroid types within the main asteroid belt based on photometric data for 656 asteroids. Definitions of the designations are in Table I. [Reprinted with permission from Gradie, J., and Tedesco, E. (1982). "Compositional structure of the asteroid belt," *Science* **216,** 1405–1407. Copyright 1982 AAAS.]

TABLE I Asteroid Taxonomy

Asteroid class	Inferred mineralogy	Possible meteorite analogs
Primative asteroids		
D	Organics + anhydrous silicates? +ice??	None (cosmic dust?)
P	Anhydrous silicates + organics? +ice??	None (cosmic dust?)
K	Olivine, orthopyroxene, opaques	CV3, CO3 chondrites
C (wet)	Clays, carbon, organics	CI, CM chondrites
Metamorphic asteroids		
B	Clays, carbon, organics	None (highly altered CI, CM??)
G	Clays, carbon, organics	None (highly altered CI, CM??)
F	Clays, opaques, organics	None (altered CI, CM??)
C (dry)	Olivine, pyroxene, carbon (+ice??)	"CM3" chondrites?
E (wet)	Clays, salts ????	None (opaque-poor CI, CM??)
W	Clays, salts ????	None (opaque-poor CI, CM??)
Q	Olivine, pyroxene, metal	H, L, LL chondrites
Igneous asteroids		
V	Pyroxene, fledspar	Basaltic achondrites
R	Olivine, pyroxene	None (olivine-rich achondrites?)
A	Olivine	Brachinites, pallasites
M (dry)	Metal, enstatite	Irons (+EH, EL chondrites?)
T	Troilite ?	Troilite-rich irons (mundrabilla)?
E	Mg-Pyroxene	Enstatite achondrites
S (7 subtypes)	Olivine, pyroxene, metal	Stony-irons, IAB irons, lodranites, winonites, siderophyres, ureilites, H, L, LL chondrites

higher temperatures, while those found in the outer regions of the solar system are enriched in volatiles and silicate minerals that are stable at lower temperatures.

E. Mineralogical Composition and Chemistry

Until there is a sample return mission from an asteroid, the composition of any asteroid has to be determined without the benefit of direct examination. A number of different techniques are used to provide information on the mineral composition and texture of asteroids using large telescopes and sensitive instruments from the ground. By measuring the intensity of reflected sunlight from the surface of an asteroid as a function of wavelength (color), certain rock-forming minerals and their chemistry can be determined. With this knowledge we can infer some of the formation processes of the asteroids. For example, Ceres reflects about 10% of the sunlight incident upon it. Also, as shown in Fig. 6, it has a featureless reflectance spectrum in the visible and near-infrared wavelength region and its reflectance is low in the blue and ultraviolet region. It is inferred that Ceres is composed of dark, claylike material similar to carbonaceous chondrite meteorites. In addition, at 3.0 μm an absorption band due to water in crystal structures and possibly free water is present, consistent with claylike mineralogy. Its density, 2.1 g/cm^3, is also consistent with a carbonaceous chondrite-like composition.

The asteroid 4 Vesta, on the other hand, has a different dominant mineralogy on its surface. Its albedo is about 38% and its spectrum has three strong absorption bands, one with an absorption edge beginning at about 0.7 μm and continuing into the ultraviolet, and two others at 0.93 and 2 μm. A fourth, weaker band is at 1.25 μm and is superimposed on the band at 0.93 μm (Fig. 6). The band centers are at wavelengths indicative of a combination of the minerals pyroxene and plagioclase, and both are minerals that commonly form from the cooling of silicate magma such as are prevalent on the moon and volcanic regions of the earth and the other inner planets. It is clear that the histories of Ceres and Vesta have been very different. In addition, observations of the surface of Vesta have been made as the asteroid rotates, and a compositional map of its surface has been derived based on the variation of the strength and position of the absorption bands. There are large regions of basaltic magma of varying iron compositions. The shape of 4 Vesta, as determined from Hubble Space Telescope observations, is shown in Fig. 3.

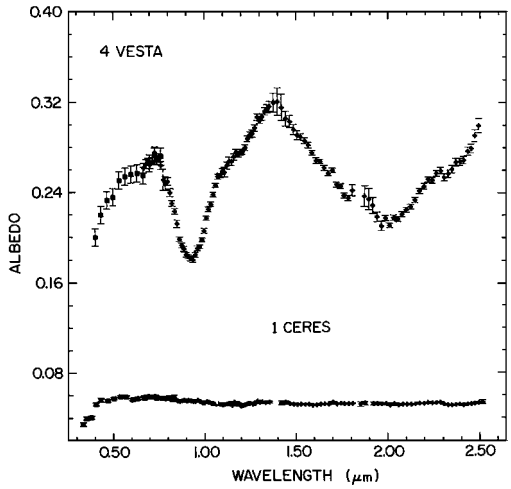


FIGURE 6 Reflectance spectra of asteroid 1 Ceres and 4 Vesta at 0.3–2.5 mm poltted as a function of albedo. The low albedo and flat spectrum of 1 Ceres indicate that the surface is probably dominated by claylike minerals. The strong absorptions in the spectrum of 4 Vesta below 0.6 mm and at 0.9 and 1.9 mm indicate the presence of a common mineral found in differentiated bodies, orthopyroxene.

Probably most of the meteorites that are studied in earth-based laboratories come from asteroids (some come from the moon and Mars) but we do not know which ones come from which asteroids. In determining the composition of an asteroid, one of the first steps in the analysis procedure is to compare the spectrum to that of the meteorite types studied with the same techniques. Ceres is about three times as bright as the carbonaceous chondrite meteorites as measured in reflectance in the laboratory. This difference is not explained, but the other aspects of its composition and density are similar to carbonaceous chondrite meteorites. Vesta has a composition analogous to a group of meteorites called basaltic achondrites. They have clearly been melted and cooled slowly so that crystals could form. They do not represent the original state of material from the initial condensation of the solar nebula, but have been "differentiated" by heating, melting, and recrystalization of the rocks. Why some asteroids apparently have not been significantly heated during their formation and others seem to have at least partially melted is an unsolved question.

F. Asteroid Surfaces

Now that a handful of asteroids have been visited by spacecraft and another handful have been observed either by the HST or by radar we are starting to understand the variety and complexity of asteroid surfaces. Many of the smaller asteroids are apparently fragments of collisions or rubble piles of collisional debris that have reassembled by self-gravity. The rubble-pile structure can result in some incredible shapes, as shown by the radar images of 216 Kleopatra in Fig. 4. Impacts dominate the evolution of surface features on asteroids. The typical asteroid is much like the moon with its surface covered with craters. Shown in Fig. 7 is the north pole of asteroid 433 Eros. Unlike the moon, impact processes also appear to erase craters, perhaps by shaking the loose material that covers most asteroids. The cratering process creates some very rough terrane, which include multikilometer scarps on 243 Ida, a large central valley on 433 Eros, and huge craters on 253 Mathilde. Mathilde has at least five craters whose diameters are significant fractions of the diameter of the asteroid. Early models of asteroid collisions suggested that the energy required to form any one of these craters should have been sufficient to destroy the entire asteroid, but Mathilde has survived to make us rethink our models.

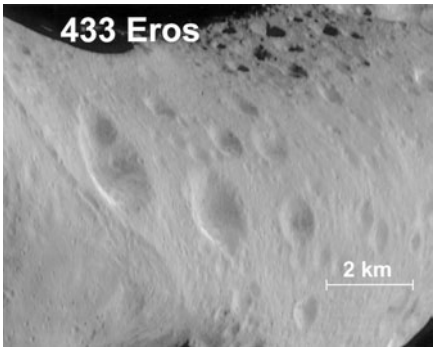


FIGURE 7 The north pole of asteroid 433 Eros as seen from the NEAR/Shoemaker spacecraft. The surface of 433 Eros is much like the Moon's surface in that it is covered by impact craters and blanketed by a loose mantle of debris called a regolith. [Courtesy of NASA/JHUAPL.]

Also like the moon, the surfaces of asteroids appear to "weather" from exposure to the space environment. With time, S-class asteroids tend to darken, probably because of micrometeorite bombardment and exposure to the solar wind.

III. COMETS

A. Introduction

Every so often, we read about the appearance of a comet in the sky. If we live away from city lights or have binoculars or a telescope, comets can be seen sitting in the nighttime sky as a diffuse glow that sometimes extends out to a tail. About once a decade, on the average, there is a comet that can be seen by just looking in the right part of the sky without the aid of binoculars or telescopes. Even more rarely, a comet is close enough to the earth

and the sun at the same time so that it can be seen in daylight. A comet is an awesome sight especially if it is bright and has a long tail. The opportunity to look at or even search for new comets should not be passed up. In ancient times comets were thought to be omens from the gods; Aristotle believed that they formed in the earth's atmosphere in response to gaseous emissions from the earth. Tycho Brahe tested Aristotle's hypothesis with observations of the comet of 1577. He reasoned that if the comet were an atmospheric phenomenon, there would be a sizable parallax observed. Brahe could measure no such parallax and concluded that the comet was much further from the earth than the moon. After Isaac Newton recorded his theory of gravitation in the "Principia," it was Edmond Halley who showed that, based on the laws of gravitation, comets orbit the sun. He hypothesized that the comets seen in 1531, 1607, and 1682 were actually the same one seen repeatedly as its orbit passed close to the sun every (approximately) 76 years. Furthermore, he predicted that it would return again in 1758. He thus successfully applied the basic principles of the scientific method, which included fitting observations to theory and then proposing a test to prove or disprove the theory. He was proven correct when, 16 years after his death, the comet was again seen and thereafter named comet Halley. Modern scholarship has uncovered records of 30 appearances of comet Halley in human history, culminating with its 1986 visit to the inner solar system. During this apparition, it became the first comet to have been viewed up close by spacecraft and well as being the subject of an intensive ground-based telescopic observing campaign.

B. Structure

Comets are seen as fuzzy balls of light sometimes with one or more wispy tails attached (Fig. 8). Comets are made of frozen volatiles, mostly water ice, and dust. As the comet gets closer to the sun and heats up, the ices near the surface turn to vapor, releasing clouds of gas and dust that seem to stream out behind, forming the characteristic "tail" of the comet. A comet's brightness comes from two processes. The first is simply the sunlight reflecting off the smoke-sized dust particles that form part of the tail. The second process occurs when the sunlight interacts with the gases in the tail and ionizes some of them creating a plasma of charged particles. As these charged particles interact with sunlight, the ions absorb and re-emit energy in a process called fluorescence.

Fluorescence occurs at specific wavelengths of light for specific ions and molecules. The fluorescing light seen from comets by the naked eye is due mostly to CO molecules, but other molecules such as OH (hydroxyl),

FIGURE 8 Comet West photographed March 9, 1976. The spectacular tail is the defining feature of a comet. [Courtesy of Dr. E. P. Moore, Joint Observatory for Comet Research, NASA/GSFC/NMIMT.]

H_2O (water), CN (cyanogen), NH (amine), S (sulfur), O (oxygen), H (hydrogen), and CS (carbon monosulfide) have been detected in regions of the spectrum outside of the sensitivity range of the human eye.

1. Nucleus

The gas and dust-rich cloud surrounding the solid core or nucleus is called the coma, which is responsible for the diffuse glow of the comet and is its most visible characteristic. The nucleus cannot be seen through the coma from earth because of its small size (typically 1-10 km in diameter). Cometary nuclei are thought to be composed predominantly of water ice, and ices of low-temperature volatiles like methane, with dust and other solids mixed. This is the "dirty ice-ball" model proposed by Fred Whipple in 1950. Using this model and the idea that comets lose volatiles and dust as they are heated up in the inner solar system, Whipple was able to explain the observed splitting of comets, meteor showers, and deviations in comets' orbits not attributable to gravitational forces. The study of the nucleus is important because its size, shape, and composition control all cometary phenomena that we observe. Presumably, the composition of the nucleus is controlled by the pressure, temperature, and composition of the material from which the comet formed. Since comets include ices that are not stable in relatively warm temperatures of the inner solar system, they must have formed far from the sun, where temperatures are much lower and a range of highly volatile ices are stable. Knowledge of the composition and structure of the nucleus of comets will tell us about the composition and thermodynamic conditions in the outer solar system.

A few comets have been studied when they are far from the sun and are not surrounded by gases. Other periodic comets have been studied using imaging processing techniques that can model and electronically subtract the coma. The results show that most comets are probably between 1 and 10 km in diameter and reflect 2–4% of the sunlight that reaches the surface, results that are in agreement with spacecraft data. Comet Hale–Bopp, which lit up the winter sky in 1997, was an unusually large comet with a diameter estimated at between 27 and 42 km. Another unusually large comet is comet Halley. The Giotto spacecraft flying past comet Halley in 1986 resolved the nucleus of a comet for the first time and showed an object 15 km long and 8 km wide that reflects 2–4% of the incident light. There are bright jets of dust on the sunward side of the nucleus and topographic features rising into sunlight. The nucleus was not sampled directly, so there is no direct information about the composition of the nucleus, only knowledge about the dust in the coma.

Comets are probably composed of three major types of materials: frozen volatiles, silicate minerals, and organics. The volatiles are primarily water ice (about 80% water ice based on Halley flyby results), with lesser amounts of carbon monoxide, carbon dioxide, and a wide range of trace gases. The silicates are made up of varying amounts of fine dust and larger rocks composed of common minerals such as olivine, pyroxene, and clays that are found on earth as well as in meteorites. Although little is known about the exact makeup of the organic components of comets, they include polymerized formaldehyde and well as a range of complex organic compounds.

2. Coma

The coma consists of gas and dust ejected from the nucleus. It is by studying the composition and chemical processes active in the coma surrounding the nucleus that we can infer the composition of the nucleus.

Unfortunately, not all of the comet evaporates to a gas, particularly the refractory silicates and complex organics, so the whole complement of cometary material cannot be known without *in situ* measurements of the nucleus. Most of the observed volatile species in comets are photodissociation fragments of more stable molecules ejected from the nucleus. Table I lists the chemical species detected in comets. From the study of comet Halley, it was learned that outgassing from this comet is nonuniform, with most gas being ejected from a few active vents on the sunlight side of the nucleus. With increasing distance from the nucleus, the gas distribution becomes more spherically symmetric. Dust jets however, retain their asymmetric distribution.

3. Tails

Comet tails come in two types. The large and often curved tails, an example of which is shown in Fig. 8, are type II tails and are composed of dust. Dust particles are typically ejected from the surface of a comet by the gas jets created by warming frozen volatiles. This gas-powered ejection, as well as forces from the sun's gravity, solar radiation pressure, and the differing masses and ejection directions of the dust particles, puts the dust into orbits of their own that steadily diverge from the parent comet. The orbital/velocity differences between the comet and it ejected dust tend to produce a curved tail relative to the comet–sun line. Comet tails will always point away from the sun because of the radiation pressure of sunlight. The force from sunlight on the small dust particles pushing them away from the sun is greater than the force of gravity acting in the direction toward the sun. As a result, during its inbound passage a comet's tail streams behind the nucleus, but on its outbound passage

back to the outer solar system the tail is in front of the nucleus.

The dust analyzers on the Soviet VEGA spacecraft revealed the existence of at least three classes of dust grains. One class is composed of low-atomic-number elements, primarily carbon, hydrogen, oxygen, and nitrogen, called CHON particles. A second is similar to the composition of CI carbonaceous chondrite meteorites but enriched in carbon. The third type is equivalent to a hydrogen-enriched CHON particle. A significant finding from the spacecraft explorations of comet Halley was that the "parents" of some of the gaseous species might be small dust grains rather than larger, stable molecules.

Additionally, the dust detectors revealed large quantities of very small particles. The smallest size detected was 0.01 g, assuming a density of 1.0 g cm^{-3}. Their distribution was unanticipated, too. The smallest particles were detected much farther from the nucleus than expected based on classical theory where radiation pressure decelerates the particles. Possible explanations of this phenomenon include (1) the grains do not absorb radiation, (2) larger grains break up after leaving the nucleus, and (3) the grains are electrically accelerated by virtue of their electric charge and their presence in the magnetized plasma environment.

Some comets also have tails pointing straight back from the comet–sun line, as can be seen in Fig. 8, illustrating type I tails. These are composed of molecules that have physically interacted with charged particles emitted from the sun called the solar wind. Such interactions result in charged molecules (called plasma) that emit light. Plasma tails follow the path of the solar wind (which travels radially away from the sun). Thus, plasma tails form straight lines behind the comet head following the direction of the solar wind. They eventually cease to receive enough light to be seen and dissipate after leaving the inner solar system. Sometimes these tails form kinks or ropey structures, or break up entirely. This is due to changes in the flux of particles leaving the sun in the direction of the comet or to jetting from the comet itself. Studying the dynamics of these tails tells us about the interaction of the sun with other particles in the solar system.

C. Dynamics and Origin

We have reviewed what is currently known about the structure and composition of comets and why we need to know such things. But where do these comets come from and how do they get here? By observing the motion of a comet, we can plot its orbital path through the solar system. Comet Halley travels out beyond the orbit of Neptune, but many other comets travel so far away from the sun that they are seen only once in tens of thousands of years. Comets are grouped into two classes: the short-period comets, which orbit the sun in 200 years or less, and the long-period comets, with periods greater than 200 years. A subset of the short-period comets is the Jupiter family of comets with orbits less than 20 years. Are the short-period comets related to the long-period comets? We believe they are because their chemistry appears to be the same. Analysis of the orbits of comets have suggested two major reservoirs for these objects. The largest is known as the Oort Cloud, named after J. Oort, the Dutch astronomer who proposed its existence in 1950. The Oort Cloud is made up of trillions of comets and is located 1000–100,000 AU from the sun. Over millions of years, gravitational tugs from the occasional passing star exert enough force to change the orbits of a few comets and send them toward the sun. The sun's gravity takes over from there and pulls the comet into the inner solar system, providing the occasional long-period comet. The inner comet reservoir is the Edgeworth–Kuiper belt, which extends from the orbit of Neptune out to approximately 1000 AU. The inner edge of this belt interacts gravitationally with the gas giants and provides the source for the short-period comets. The Centaurs, described in the asteroid section, are probably originally members of the Edgeworth–Kuiper belt that have been perturbed by the outer planets. The trans-Neptunian objects, also described in the asteroid section, really are the population of inner edge of the Edgeworth–Kuiper belt. Revolutionary advances in astronomy have made constructs like the Edgeworth–Kuiper belt, originally only a theoretical proposal, observable reality, with over 345 objects already discovered.

IV. FUTURE DIRECTIONS

A. Areas of Study

The problems currently pursued in studies of both asteroids and comets will remain the major areas of study for a number of years to come. Among the enigmas of the asteroids is the source and mechanisms by which apparently some of the asteroids are heated while others are not. The nature of this heating mechanism and its distribution, which probably occurred shortly after the formation of the asteroids, represents a gap in the time line of solar system history.

While we use the meteorites as a low-cost "returned" *in situ* samples of asteroids, the continuing challenge is to characterize their source asteroids and determine solid links between meteorite and asteroid types. Several asteroid types, as shown in Table I, are apparently not represented in the meteorite collection. Of the meteorites we have, there are several strong differences between the reflectance spectra of common meteorite and asteroid types.

This may be due to alteration of asteroid surfaces by "space weathering" similar to effects observed on the lunar surface.

A major challenge is to find and characterize the large population of earth-crossing asteroids. Some of these objects potentially pose an impact hazard to earth. Asteroid search programs have so far found about half of the earth-crossing asteroids larger than about 1 km in diameter. However, very little is known about this population of relatively small asteroids. Some of these asteroids are, in energy terms, very accessible from earth and represent appealing targets for future sample return missions. These objects may be a resource for minerals, metals, and fuel for explorers in the distant future.

B. Future Opportunities

At the time of this writing we have visited five asteroids and one comet with spacecraft. The Giotto spacecraft flew by the comet 1P/Halley. On its way to Jupiter, the Galileo spacecraft flew by 951 Gaspra and 243 Ida. The NEAR spacecraft flew by 234 Mathilde and is currently orbiting 433 Eros. Finally, the Deep Space 1 spacecraft flew by 9969 Braille. This represents enormous progress in the understanding of small bodies over just 15 years, however many more missions are currently flying or on the horizon.

Deep Space 1 is on its way to fly by the comet 19/P Borrelly in late 2001. The Stardust spacecraft is cruising to collect and return dust from the comet 81P/Wild-2. The CONTOUR mission is scheduled for launch in 2002 and will conduct a "tour" with multiple comet flybys using Earth gravity assists. Targets include 2/P Encke (2003), 73/P Schwassman–Wachmann III (2006), and 6/P d'Arest (2008). The European Space Agency will launch its Rosetta mission in 2003 for a rendezvous with Comet 46 P/Wirtanen (2011), with flybys of asteroids 4979 Otawara (2006) and 114 Siwa (2008) on the way. The Japanese Institute of Space and Astronomical Science (ISAS) plans the launch of an ambitious rendezvous and sample return mission to a near-earth asteroids with a launch in 2002 or 2003. Finally, the Deep Impact Mission plans to fly by comet 9P/Tempel 1 and impact it with a 1000-kg copper projectile to study its structure and composition. This constellation of spacecraft and missions plan visits, either rendezvous or flybys, of 10 small bodies over the next 10 years.

Asteroids and comets have ceased to be just points of light on photographic plates and become worlds in their own right, each with a complex story to tell of geological and chemical evolution as part of the larger picture of the solar system's history. This story continues to come into focus as we learn more about the composition and history of these amazingly diverse objects. The future prospects for studies of small bodies are extremely exciting. Spacecraft missions, along with continuing observations by telescopes and radar, will greatly expand our knowledge of every aspect of these small worlds.

SEE ALSO THE FOLLOWING ARTICLES

ASTEROID IMPACTS AND EXTINCTIONS • CELESTIAL MECHANICS • COMETARY PHYSICS • IMPACT CRATERING • METEORITES, COSMIC RAY RECORD • SPACE PLASMA PHYSICS

BIBLIOGRAPHY

Binzel, R. P., Gehrels, T., and Matthews, M. S. (1989). "Asteroids II," University of Arizona Press, Tucson, AZ.

Gehrels, T. (ed.) (1994). "Hazards due to Comets and Asteroids," University of Arizona Press, Tucson, AZ.

Wilkening, L. (1982). "Comets," University of Arizona Press, Tucson, AZ.

Solar Physics

Therese Kucera
Emergent Information Technologies Inc.

Carol Jo Crannell
NASA Goddard Space Flight Center[1]

GLOSSARY

Active region Area on the Sun with especially strong magnetic field. The stronger ones are associated with sunspots. These areas are bright in ultraviolet and X-ray images.

Arcade Series of arches (magnetic loops) formed by magnetic fields extending into the outer layers of the solar atmosphere. The closed magnetic loops that form an arcade confine hot gas or plasma.

Arcsecond Unit of angular measure that is $1/60$ of an arcminute, or $1/3600$ of a degree. An arcsecond corresponds to a spatial distance on the Sun, as viewed from the Earth's orbit, of 725 km.

Astronomical unit (AU) Parameter describing the Earth's orbit around the Sun, corresponding to the mean distance between the Sun and the Earth, approximately 1.5×10^8 km. Because the Earth's orbit is elliptical, the actual Earth–Sun distance is not constant.

Chromosphere Thin layer of the solar atmosphere that lies above the photosphere and below the corona. Light from this layer is predominantly Hα emission with additional contributions from other strong atomic transitions.

Corona Outermost and hottest layer of the solar atmosphere. The corona is visible to the naked eye during eclipses when light from the inner layers is blocked by the moon.

Coronal hole Region of the corona that appears to be dark when observed in the ultraviolet and soft X-ray portion of the electromagnetic spectrum. These features were discovered with instruments carried on space satellites and rockets. They are thought to be the source of high-speed solar wind streams.

Coronal mass ejection Magnetically confined plasma structure which erupts from the Sun. Abbreviated CME.

Faculae Regions in the upper photosphere, frequently in the vicinity of sunspots, that are brighter than the surrounding medium due to their higher temperatures.

Filament Dense, massive structure that lies above the chromosphere, generally along a line separating regions of opposite magnetic polarity. Filaments appear as dark, irregular lines when observed on the solar disk. When observed over the limb of the Sun filaments appear bright and are referred to as prominences.

Flare Rapid release of energy from a localized region on the Sun in the form of electromagnetic radiation and, usually, energetic particles. The spectral range of the radiation may extend from meter waves through γ-rays. The time-scales of such events range from fractions of a second to hours.

Footpoint Lowest visible portion of a magnetic loop, corresponding to the vicinity on the solar disk in which the loop intersects the photosphere.

Gamma-ray Highest energy electromagnetic radiation. Usually considered to be at energies above 100 keV.

Granulation Irregular light-and-dark structures, visible in the lower photosphere, indicating the tops of convection cells. Individual granules appear and disappear on time-scales of order 10 min and exhibit a range of sizes, with 1000 km being a typical diameter.

Hα Light emitted with a wavelength of 656 nm (6563 Å) from an atomic transition in hydrogen, the lowest energy transition in its Balmer series. This wavelength is in the red portion of the visible spectrum and is the dominant emission from the solar chromosphere.

Hard X-rays High-energy X-rays, generally considered to be X-rays with energies above about 10 keV.

Helioseismology The study of the interior of the Sun using observations of solar vibrations.

Magnetic dipole The simplest magnetic field—that produced by a single current loop or by a bar magnet with south polarity at one end and north at the other.

Optical depth Measure of how far one can "see" into a semi-transparent medium, such as the solar atmosphere. At the surface of such a medium, none of the photons emitted in the direction of an observer are absorbed, and the optical depth is zero. Beneath the surface, only a fraction of the number of photons that originate within the medium traverse the distance to the surface without being absorbed. That fraction decreases exponentially with increasing optical depth.

Photosphere Innermost portion of the solar atmosphere. The base of the photosphere is defined to be the lowest level in the Sun that can be observed directly in the visible portion of the electromagnetic spectrum.

Plage Portion of a magnetic active region that appears much brighter in Hα than the surrounding chromosphere. Plages usually are visible before the sunspots with which they are associated and persist after the sunspots disappear.

Plasma Gas in which some or all of the constituents are partially or fully ionized. The plasma state is generally associated with temperatures in excess of 10^4 K.

Prominence Filament viewed on the limb of the Sun that extends above the chromosphere into the corona. A prominence, like the chromosphere itself, radiates primarily in Hα.

Soft X-rays Low-energy X-rays, generally considered to be X-rays with energies below about 10 keV.

Solar cycle The approximately 11-year cycle over which the Sun's magnetic field varies in complexity. Also known as the sunspot cycle.

Spicule Small, filamentary magnetic structure containing material at chromospheric temperatures that extends into the corona. Spicules are dynamic features that rise and then dissipate on time-scales of minutes.

Streamer Large-scale structure formed by closed magnetic field lines in the Sun's corona.

Sunspot Region in which strong magnetic fields emerge into the solar atmosphere from below the solar surface. Visible as dark spots in both Hα and in the white-light continuum, sunspots are cooler than the surrounding medium because the magnetic fields from which they are formed suppress the temperature of the plasma they contain.

Surge Great eruption of hot material that originates below the chromosphere. A surge may accompany a flare and is considered to be a type of eruptive prominence.

Tachocline The boundary region between the radiative and convective zones.

THE SUN is important to us for many reasons. From our perspective as Earth-dwelling mortals, the Sun's primary importance is its role as the source of energy that sustains life on Earth. From an anthropological perspective, the Sun serves as a source of human wonderment, artistic inspiration, and even religious devotion. From the point of view of a technological civilization reaching into space, it is a source of variations in space weather that can affect our spacecraft, astronauts, and communications.

From the perspective of space science, the Sun is both the closest and most accessible star and a laboratory with unique facilities for studies of plasmas and magnetohydrodynamics. The solar laboratory contains matter at temperatures, pressures, and densities attainable nowhere else currently accessible from Earth. The information presented here, a description of the Sun, its activity, and the thrust of future investigations, was developed from this latter perspective.

I. DESCRIPTION OF THE SUN AS A STAR

A. General Physical Characteristics

Table I presents numerical parameters characterizing the size and energy output of the Sun. These values are the standard yardstick by which other stars are measured. The large number of significant digits tabulated here serves mainly to illustrate the precision to which these parameters

TABLE I Dimensions, Distances, and Luminosity[a]

Radius of the Sun, $R_\odot$	6.9599×10^5 km
Mean distance from Sun to Earth (astronomical unit, AU)	$1.495979(1) \times 10^8$ km
Sun-to-Earth distance	
At perihelion	1.4710×10^8 km
At aphelion	1.5210×10^8 km
Mass of the Sun, $M_\odot$	$1.989(1) \times 10^{30}$ kg
Luminosity of the Sun, $L_\odot$	$3.826(8) \times 10^{26}$ W
Intensity of the Sun's radiation at I AU	1.360 kW m^{-2}

[a] The numbers in parentheses represent the measurement uncertainty in the last digit of the associated parameter.

are known. Also listed are parameters characterizing the Earth's orbit around the Sun and the intensity of the Sun's radiation at the mean orbital distance.

The appearance of the Sun depends critically on how it is observed. Each type of radiation observed carries specific information about the physical processes at work on the Sun. Special types of instruments reveal aspects otherwise invisible. Coronagraphs show the dimmer outer regions of the Sun's atmosphere otherwise visible only during total solar eclipses. Spectroscopy can reveal motions, magnetic field strengths, temperatures, and densities. *In situ* measurements of the characteristics of the solar wind have extended our knowledge of the solar magnetic field both near the Earth and beyond the orbits of the planets.

As an example, Fig. 1 shows the Sun's disk observed on a single day in six different wavelengths of light. In visible light we can see the white disk of the Sun with the dark spots known as sunspots (Fig. 1a). By analyzing the spectral lines produced by the Sun we can measure the strength of the Sun's magnetic field at its surface, producing a magnetogram (Fig. 1b). This magnetogram reveals that the sunspots are regions of intense magnetic field. Further images of the Sun (Figs. 1c, e, f) reveal that the sunspot regions are at the bases of systems of hot loops which emit radiowaves, ultraviolet light, and X-rays. Figure 1d shows the Sun imaged in a spectral line of hydrogen known as Hα. In this line we also see the long dark filaments. These filaments form in long channels between areas of opposing magnetic fields. Such channels can be seen in the ultraviolet image in Fig. 1e.

Data concerning the Sun are obtained with many different kinds of instruments and from many different vantage points, both on the ground and in space. Techniques for observing the Sun's various emissions throughout the electromagnetic spectrum are illustrated in Fig. 2. Optical and most radio emissions are able to penetrate the Earth's atmosphere, so observations in these wavelength bands have usually been carried out with ground-based facilities. The atmosphere and its fluctuations do distort the highest and lowest frequency radio emission and, thus, limit the spectral range that can be measured from Earth. In the optical domain, fluctuations in the Earth's atmosphere limit the angular resolution to approximately 1 arcsec. At the Sun, this corresponds to a linear dimension of 725 km, approximately the distance between Washington, DC, and Cincinnati, OH.

For light with wavelengths shorter than that of visible light (ultraviolet (UV), X-rays and γ-rays) the Earth's atmosphere is very strongly attenuating. In the range of wavelengths from UV through soft X-rays the Sun's emission is very intense, and useful observations can be obtained for short times with instruments carried on board rockets, which rise to altitudes of a few hundred kilometers before they fall back to Earth. Hard X-rays and γ-rays, the highest-energy, shortest wavelength electromagnetic radiations, can penetrate deeper into the Earth's atmosphere. These emissions, therefore, can be observed from balloon-borne platforms at altitudes of 30–10 km. High-altitude scientific balloons can observe the Sun continuously for up to 30 days if launched in the polar regions.

Continuous measurements and long-term observations of the Sun, however, require spacecraft, which carry instrument payloads above the Earth's atmosphere and sometimes out of Earth orbit for missions that may last many years. Spacecraft beyond the Earth's magnetic field can also measure the properties of the solar wind directly as it flows through the solar system.

B. Structure and Composition

Unlike the Earth, the Sun is gaseous throughout and so does not have any solid surface. The structure of the Sun can be thought of as a series of concentric spherical shells or layers, each characterized by a unique combination of physical processes. At the center of the Sun is the nuclear-burning core, as illustrated in Fig. 3. Traveling outward, one encounters first the radiative zone, then the convection zone, then the photosphere, the chromosphere, and finally, the corona. All of these regions are powered by the nuclear-burning core from which energy is transported outward through successive layers by radiation and convection. The temperature is 15×10^6 K in the core and decreases monotonically outward to a minimum of approximately 4×10^4 K in the chromosphere. The transition from radiative to convective energy propagation occurs in the region in which the temperature drops below 2×10^6 K, so that convection becomes a more efficient transport mechanism than radiation. In this boundary between the radiative and convective zones is also an abrupt change in the rotational profile of the Sun known as the tachocline. Inside the tachocline the Sun essentially rotates

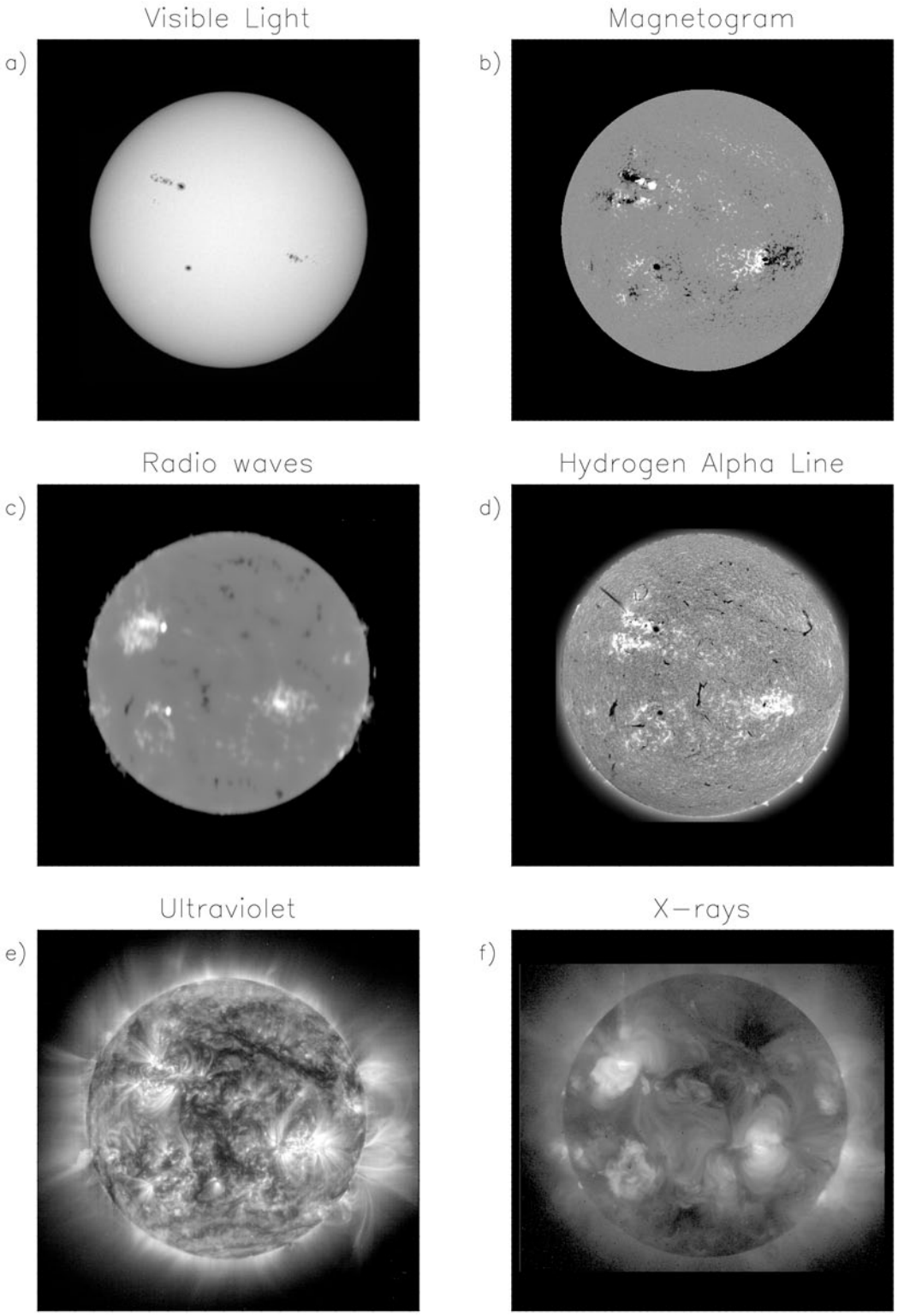


FIGURE 1 The Sun as seen in different wavelengths on Dec. 21, 1999: (a) Visible light over a broad band of wavelengths (Big Bear Solar Observatory). (b) Data showing the magnetic field flux along the line of sight. Black signifies a strong south pole and white a strong north pole (SOHO/MDI). (c) Radio waves at 17 GHz (Nobeyama Solar Radio Observatory). (d) The Hα line at 656 nm (Big Bear Solar Observatory). (e) Ultraviolet at 17.1 nm (SOHO/EIT). (f) Soft X-rays (Yohkoh/SXT).

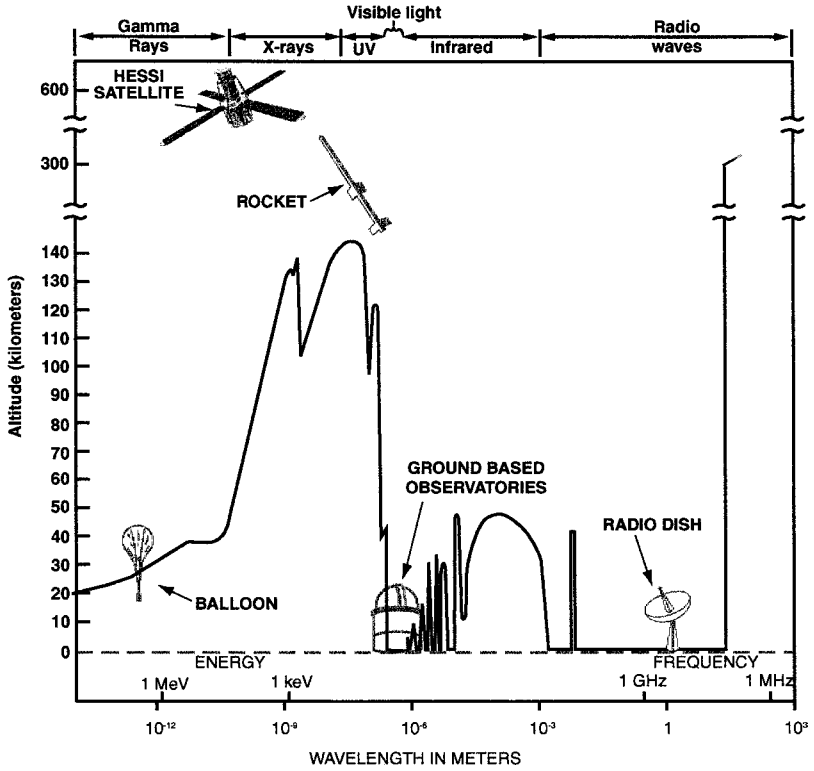


FIGURE 2 The varying transparency of the atmosphere as a function of wavelength. The curve shows the altitude at which incoming radiation at a given wavelength is reduced by half. Instrumentation used to observe at different wavelengths is also shown. Visible light and radio waves reach the ground and can be observed from there, but for other wavelengths we must send instruments to higher altitudes. (Transparency data courtesy of J. Pasachoff "Astronomy: From the Earth to the Universe.")

as a solid body. Outside the tachocline, in the convection zone, the Sun rotates differentially, with the equatorial areas rotating more quickly than the poles (see Section II).

The only radiation that carries information out of the Sun directly from these inner regions is a flux of neutrinos produced in nuclear burning. The ability of these massless (or almost massless) particles to penetrate so much solar matter is a result of the fact that they interact very weakly with matter. Consequently they are also very difficult to detect. Until recently, there was a discrepancy between the

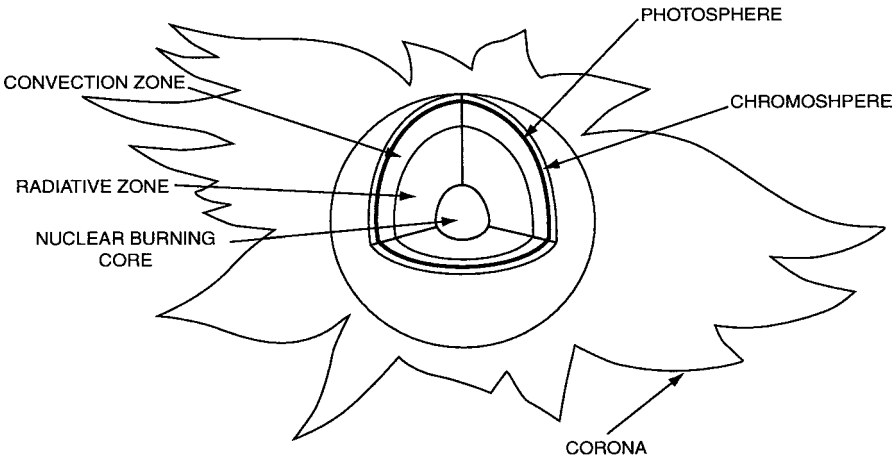


FIGURE 3 Structure of the Sun.

predicted number of neutrinos produced by the Sun and the actual number observed. From this it was clear that there was a problem with our fundamental understanding of either the Sun or neutrino physics. Advances in particle physics in the 1990s revealed that the apparent discrepancy was due to an incomplete understanding of the properties of neutrinos.

Most of our knowledge of the solar interior regions does not come from direct observations but rather from models checked indirectly by neutrino observations and, more recently, by the study of the Sun's vibrations (helioseismology, see Section II).

The innermost layer of the Sun that can be observed in visible light is called the photosphere. The base of the photosphere is defined as the depth in the solar atmosphere at which a photon of wavelength 500 nm has a 37% probability of escaping without scattering or being absorbed. This condition is referred to as optical-depth unity for a photon of the specified wavelength. From lower depths, even greater quantities of material are encountered along any outward trajectory, thus decreasing the probability of an emitted photon escaping without interacting. The photosphere is, as a result, more luminous than the optically thin outer portions of the solar atmosphere and defines the size of the Sun as observed in visible light. The height, temperature, and density of the various layers of the solar atmosphere are given in Table II.

The atmospheric layer that lies just outside the photosphere is the chromosphere. The visible emissions from the chromosphere are generally overpowered by the full light of the photosphere except when the Sun is observed with special filters that block most of the visible light. One exception occurs during a total eclipse when the chromosphere is visible as red light at the edge of the Sun's disk just before totality. During totality itself, however, light from the chromosphere, as well as from the photosphere, is blocked by the moon. This red light is Hα emission, emitted at a wavelength of 656 nm due to an atomic transition in neutral hydrogen. Unlike the simple representation in Fig. 3, the chromosphere is not a smooth spherical shell but, instead, exhibits many large- and small-scale features that may extend well into the corona. Their identity with the chromosphere is based on their temperature which supports Hα emission. Prominences are large-scale chromospheric features with an arch-like structure, visible in Hα at the edge of the solar disk against dark sky. Because these structures absorb light from the underlying photosphere, they appear as dark features when observed on the solar disk. Spicules are fine-scale chromospheric features with an even more flame-like appearance than the corona.

TABLE II Characteristics of the Solar Atmosphere

Region	Height above base of photosphere (km)	Temperature (K)	Density (atoms m^{-3})
Photosphere	0–320	6500–4500	10^{23}–10^{22}
Chromosphere	320–1990	4500–28,000	10^{21}–10^{16}
Corona	at 7×10^5	1.8×10^6	10^{12}

Perhaps the most puzzling feature of the outer solar atmosphere is its temperature structure. Instead of a continued decrease in temperature with distance outward, an increase is observed. The gradual increase observed in the chromosphere steepens with altitude so that the temperature in the corona is over a million degrees K. While this temperature increase appears to violate the elementary thermodynamic principle that a body cannot supply heat to a hotter body without external work being done, the paradox is resolved with the understanding that the photosphere heats the corona from the nonthermal source of energy stored in its magnetic fields. The exact mechanism by which this happens, however, is not yet determined. Recent work has emphasized the importance of small-scale processes occurring all over the solar disk. The most popular theories of coronal heating involve either currents generated by changing magnetic fields or heating by hydromagnetic waves. Other less conventional theories receiving consideration are heating by other types of waves (specifically ion-cyclotron) or high-energy particles escaping up from the photosphere. The relative importance of different possible mechanisms is the subject of intensive investigation.

The area of steeply rising temperature from chromosphere to corona is referred to as the transition zone. Material in transition zone temperatures between the chromosphere and corona (about 10^5 K) are even more discrete and fragmented than the features of the chromosphere, and so it is argued that it is misleading to describe it as an actual layer in the Sun's atmosphere. That is why the transition "zone" is not listed in Table II.

When specifying the chemical composition of the Sun, one can be at least 98% correct by saying "hydrogen and helium." In spite of the fact that these are the two lightest elements in the periodic table, this statement is true not only for abundance by number of atoms but also for abundance by weight. The next most abundant elements, in decreasing order, are oxygen and carbon. The rest comprise less than 1% of the Sun by weight and less than 0.03% by number.

The photospheric abundances are quite similar to those determined for local galactic abundances, but in the corona, it is found that the ratio of elements with low first ionization potentials ($\lesssim$10 eV, e.g., iron) to high ones (e.g., oxygen) are higher than those ratios in the photosphere by factors of 3–5. There is still some argument about

whether this is due to a depletion of elements with high first ionization potentials (FIPs), an enhancement of low FIP elements, or some combination of the two. Most models of the FIP effect utilize diffusion effects in the chromosphere or transition zone which affect neutral and ions differently, but the exact cause for this abundance difference is not yet determined. This effect exists to varying levels in the solar wind (see Section III.C) and in flares.

C. The Sun among Stars

Stars can be classified in many different ways. Among the most useful schemes are classifications by luminosity, or intrinsic brightness, and by spectral class, which is a measure of a star's surface temperature. The plot in Fig. 4 illustrates a particularly informative combination of these classification schemes, called a Hertzsprung–Russell (H–R) diagram after its originators. In the H–R diagram surface temperature increases from right to left. Normal stars populate the main sequence, the band that runs diagonally across the diagram. These stars are in the early stages of their evolution and are known as dwarfs to distinguish them from the much larger, more luminous Stars in the upper right-hand portion of the diagram.

As indicated by its location in the diagram, the Sun is a G dwarf, an average star both with respect to its luminosity and with respect to its spectral class. Contrary to what one might assume, the lack of superlatives in its description adds to the attractiveness of the Sun as an object of astronomical interest. The normalcy enables most of what is learned about the Sun to be applied to other stars, all of which are much less accessible. Distance makes spatially resolved observations of stellar dynamics and activity, described for the Sun in the following sections, directly observable only on the Sun. For the same reason, development of instrumentation for observations of astrophysically significant phenomena on the Sun has pioneered the development of instrumentation for observations of similar phenomena outside the solar system, such as γ-ray bursts.

D. The Solar Magnetic Field

1. Magnetic Field Motions

Understanding the Sun's magnetic field is key to understanding the Sun in general. The Sun's appearance and behavior are strongly affected by the magnetic field, which exists and changes on every scale we can observe.

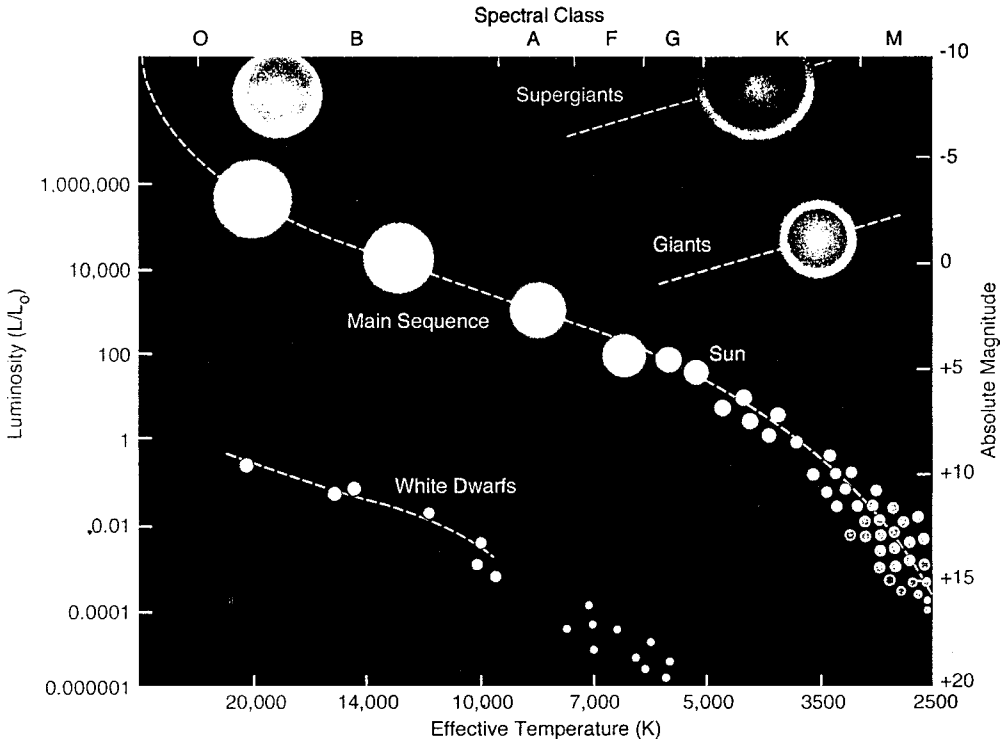


FIGURE 4 Plot of luminosity as a function of surface temperature and spectral type for different types of stars. Stars are plotted schematically to show colors and relative sizes, but not to true scale (which would require much larger red giants). The Sun is part of the Main-sequence, the band of middle-aged, hydrogen-fusing stars stars running diagonally across the plot. Based on a figure by Hartmann, W., and Impey, C. "Astronomy: The Cosmic Journey," Brooks-Cole, Pacific Grove, CA.

Variations of the magnetic field at any given point are governed by two things: the motion of the plasma and the diffusion of the magnetic field through the plasma. On large scales on the Sun the diffusion is much less important than the plasma motions. Motion perpendicular to field lines is very slow, so all plasma motion must be either along field lines or else move as the field lines move. Under these circumstances the plasma is said to be "frozen in" to the magnetic field.

This is quantified in the Magnetic Reynolds number,

$$R_M = \frac{\tau_{\text{dif}}}{\tau_{\text{trans}}} = \frac{l^2/\eta}{l/v},$$

where l is the length scale of the motion, v is the velocity of bulk plasma motions, and η is the diffusion coefficient, $\eta = (4\pi\sigma)^{-1}$, where σ is the conductivity. The diffusive time scale, τ_{dif}, is the time it takes for a magnetic field to diffuse out of a conductor of length l, while τ_{trans}, the translation time scale, is the time it takes for the plasma to move over the distance l. The conductivity of solar plasma is not very high. The conductivity can be approximated by $\sigma = 2 \times 10^{-14} T^{3/2}$ emu where T is the temperature in Kelvin. However, even the smallest spatial scales we observe on the Sun are fairly large, about 300 km. The diffusion scale over such lengths is over a year. In comparison, we often see solar plasma moving at velocities of hundreds of Kilometers per second, so the bulk motion time scale is very short. Thus on scales which we can currently resolve in the Sun $R_M \gg 1$; the magnetic field moves through the plasma so much more slowly than it moves with the plasma that the diffusion can be ignored.

This does not, however, mean that diffusive processes are not important on smaller scales. Theories of energy release on the Sun needed to explain coronal heating, flares, coronal mass ejections, the dynamo, and other solar phenomena rely on small scale dissipative processes such as reconnection (see Section III.D), heating via waves, and current heating, all occurring below the resolution that can be observed at this time.

2. Measurements of the Magnetic Field

a. Photosphere and chromosphere. The magnetic field in the photosphere and chromosphere can be quantitatively measured by way of the Zeeman Effect. In the presence of a magnetic field certain electron energy levels split into values corresponding to discrete values of the electron angular momentum. In classical Zeeman splitting these cause the spectral line resulting from transitions to such energy levels to split into three, with the two new components shifted to either side of the original line by

$$\Delta\lambda = \frac{\pi e \lambda^2}{m_e c} g B = 4.7 \times 10^{-13} \lambda^2 g B,$$

where λ is the wavelength in Ångstroms, B is the magnetic field in Gauss, and g is the Lande g-factor, a parameter whose value depends on the details of the atomic energy levels, and which is usually between 0 and 3 for lines in the visible spectrum. There are cases for which the atomic levels and spectral line structure are more complicated than for the classical case, but with correct analysis these can also be used for Zeeman measurements.

The Zeeman effect can be used for more than just measurements of magnetic field strength, however. The outer components of the line are circularly polarized in response to magnetic fields along the line of sight, while measurements of linear polarization of the lines can reveal the transverse component of the line as well. A resulting "vector" magnetograph image of a magnetic active region is shown in Fig. 5. Obtaining accurate measurements of the linear polarization is especially difficult, requiring both high count rates and good spectral resolution. As a result, current "vector" magnetograms have fields of view encompassing only part of the solar disk, but full disk versions should become available in the future.

b. The corona. Quantitative measurements of the magnetic field are much more difficult in the corona. Because of its wavelength dependence the Zeeman effect it is very difficult to measure in the UV and X-rays where most coronal spectral lines are found. Furthermore, the high temperatures of the corona lead to thermally broadened spectral lines which are usually too wide for the line splitting to be measured. As a result, most analysis of the coronal magnetic field is purely morphological or based on

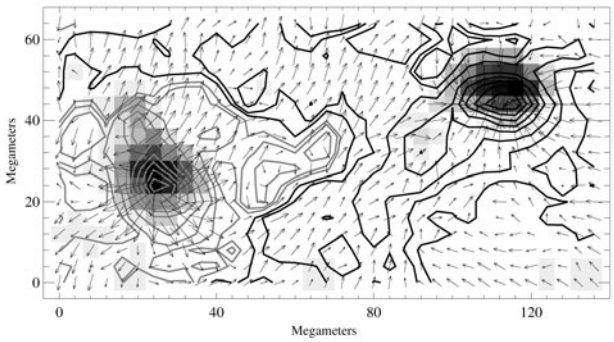


FIGURE 5 Vector magnetogram of a simple bipolar active region observed on June 4, 1998, by the Mees Solar Observatory Imaging Vector Magnetograph. The grey scale shows the visible light intensity. Contours show the strength of the field parallel to the light of sight (black is negative polarity, grey positive polarity. Contours are at ±50, 100, 200, 400, 600, 800, 1000, 1200, and 1400 G). The arrows show the strength and direction of the field perpendicular to the line of sight. (Courtesy K. D. Leka.)

models of the field extrapolated from the values measured in the chromosphere and corona.

Coronal magnetic field can be directly measured using radio observations of active regions. Microwave emissions from areas of strong magnetic field are partially produced by gyro-emission radiated as electrons gyrate around magnetic field lines. However, in order to determine the magnetic field strength, temperature, emission measure (the number of particles in the emitting volume) and the proportions of different radio emission mechanisms must also be determined, so it is not a routinely used method.

The Hanle Effect can be used to measure the magnetic field of sources on the solar limb by measuring the depolarization of light scattered by the Sun's atmosphere. This technique is useful only in certain circumstances such that it is chiefly used to measure magnetic fields in solar prominences, structures of chromospheric temperature plasma magnetically suspended in the corona.

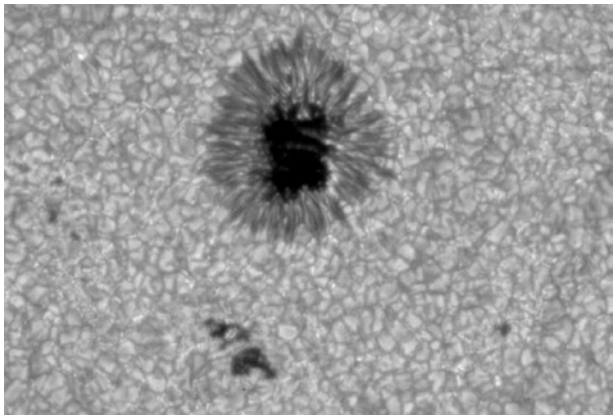


FIGURE 6 Sunspot and granulation seen in visible light. The dark center of the sunspot is known as the *umbra* while the lighter outer part is the *penumbra*. Average granule size is about 1000 km across, roughly the size of Bolivia. (Courtesy Brandt, Scharmer, Shine, Simon, Swedish Vacuum Solar Telescope, La Palma.)

II. MOTIONS IN THE SOLAR INTERIOR AND PHOTOSPHERE

Even the "quiet," nonactive Sun exhibits motion on many different scales. In this section the focus is on continuous and regular types of activity which characterize the Sun's general behavior and condition.

The motion of the Sun's surface that has been known for the longest time is solar rotation. Because of its gaseous nature, the Sun does not rotate as a solid body. The Sun does rotate about an axis, but the regions near the equator rotate more rapidly than the polar regions, with the equator rotating once every 25 days and the poles taking about 33 days.

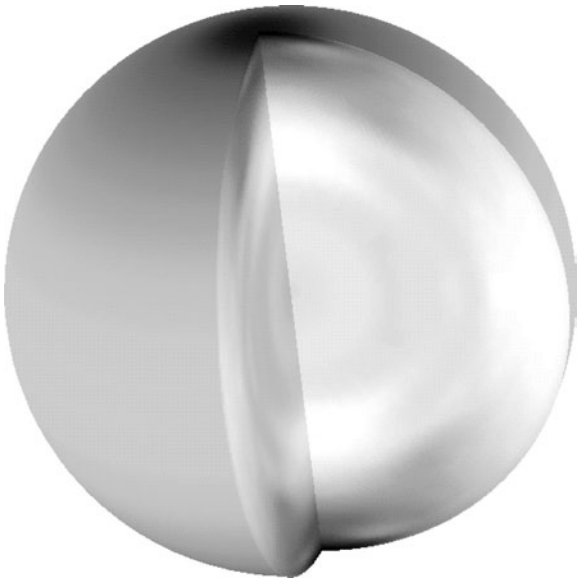
This differential rotation, as it is known, is most evident from observations of sunspots. A sunspot (see Fig. 6) is an outcropping of magnetic field through the visible surface of the Sun. Sunspots are characterized by a dark core called the umbra surrounded by a less dark region called the penumbra. Magnetic fields in the umbra range from 1000 to 4000 G. The motions of charged particles are constrained by magnetic fields so that plasma cannot easily flow across field lines. This results in a suppression of convection in regions of high magnetic field. As a result, the plasma in a sunspot is cooler (at about 3800 K) and, hence, darker than the surrounding photosphere. Sunspots form at the edges of bands of faster rotating material, indicating that they are related to the shear caused by parts of the Sun's magnetized atmosphere moving at different speeds.

Recent studies of the Sun's interior via helioseismology (see the following) have found that differential rotation extends below the surface of the Sun down to the bottom of the convection zone. This is shown in Fig. 7a. The region below the convection zone, the radiative zone, rotates more as a solid body. The shear occurring in the tachocline in which neighboring plasmas move at different speeds is probably the ultimate source region of the solar magnetic cycle (see Section III).

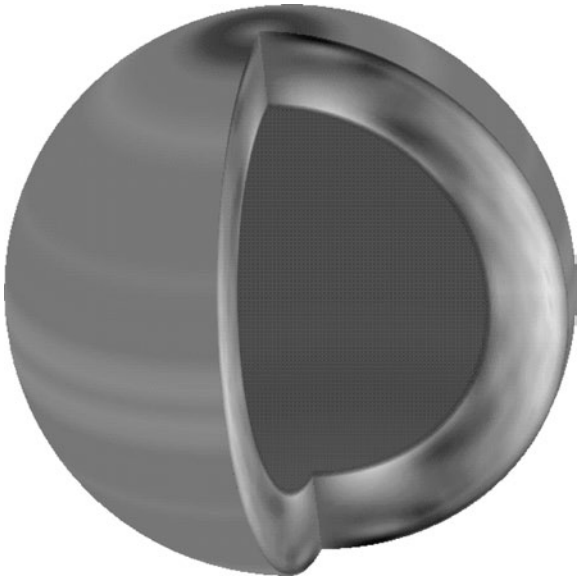
Superimposed on this pattern of differential rotation in the convection zone and solid body rotation below is another pattern of zones of material rotating faster or slower than the areas around them, rather like the jet stream in the Earth's atmosphere (Fig. 7b). These patterns of slower and faster moving material vary with time and long-term variations are still being investigated.

At the surface, another type of motion can be observed as a slowly changing grainy pattern. When the photosphere is seen with good angular resolution, 1 arcsec or better, its appearance is similar to that of a simmering pot of grits with individual grits (analogous to granular cells) appearing and disappearing on a time-scale of about 10 min (see Fig. 6). The linear dimensions of the granules are typically about 1000 km. They are actually the tops of a pattern established in the convection zone and visible in the photosphere. Plasma rises and expends in the cell centers. The heat is radiated away and the cooled material sinks down again at the cell boundaries.

Slower and larger convection patterns, called supergranules, are visible on a scale size of 30,000 km and on a time scale of days. The magnetic field is affected by these motions, emerging with material in the cell interiors and being swept toward the cell boundaries. Many of the field elements cancel with elements of opposite polarity along the way and the remainder collect in the "network," the supergranule boundary areas seen in ultraviolet. The energy produced by the cancelation of magnetic field elements on



(a)



(b)

FIGURE 7 Cut-away images showing the speed of solar rotation at the surface and in the interior. Quickly moving areas are shown in white, slower areas in black. The first image (a) shows the basic differential rotation, with the convection zone rotating more quickly at the equator than the poles and the interior radiation zone rotating more as a solid body. (b) In this image the average rotational patterns shown in (a) have been subtracted to reveal smaller variations, currents which flow more slowly or quickly than the plasma around them. The currents change with time. These flows were measured with SOHO's MDI instrument in 1996 and 1997.

these relatively small scales may be the ultimate energy source of the high temperatures in the corona, although the exact mechanism which transports the energy into the corona has not been determined.

There is a class of granulation between granules and supergranules known as mesogranules. These range in size between about 5000 and 10,000 km. However, there is some dispute as to whether these cells are a distinct class of granulation or simply part of a continuum of structures. Typical dimensions of the different scales of granulation are given in Table III.

The supergranules penetrate only a few percent of the depth of the convection zone, and the small-scale granules penetrate even less. The remainder of the convection zone cannot be probed by observations of these processes.

An even more regular motion of the quiet Sun, and one that allows us to probe more deeply into the interior, is its oscillation. This rhythmic rise and fall of the solar surface, with a dominant 5-min period, corresponds to the resonant frequency of the convection zone for sound waves. The speed at which a wave travels is affected by the temperature and the composition of the medium in which it is traveling. The longer the wavelength of a mode, the deeper it can penetrate into the Sun. Modes of different wavelengths, therefore, provide information on different regions within the Sun's interior. While a single mode could be observed and analyzed quite simply, in fact, an enormous number of modes are actuated simultaneously. A study of the Sun's interior using such observations requires coincident analysis of oscillations with periods ranging from 2.5 to 16 min, or longer. Using mathematical techniques it is possible to separate out the different vibrational modes and use them as probes of the solar interior. This study of solar vibrations is called *helioseismology*.

Starting in the 1990s, large-scale projects have been undertaken to understand the Sun's interior via the study of its oscillations. These have been undertaken both from the ground (most notably by the Global Oscillation Network Group or GONG) and in space via the Solar and Heliospheric Observatory (SOHO) spacecraft's MDI, GOLF and VIRGO instruments. Through them we have expanded our understanding of the solar interior structure as described previously. Because of these studies we were able to study the Sun's interior rotation for the first time and gain a much better understanding of the structure of the tachocline. These will be important in the study of the dynamo which powers the Sun's changing large-scale magnetic field. Helioseismology can allow us to map the structure of sunspot groups below the surface, and can even provide more information about the Sun's surface, for instance, allowing the detection of sunspots on the far side of the Sun invisible from Earth.

III. SOLAR ATMOSPHERE AND ACTIVITY

The solar magnetic field varies in a cycle known as the sunspot cycle, or, more generally, as the solar cycle. The

TABLE III Typical Parameters for Different Scale Solar Granulations Patterns

Structure	Size	Velocities		Lifetime
		Horizontal	Vertical	
Granules	300–2000 km	1.4 km/sec	1.8 km/sec	2–20 min
Mesogranules	5000–10,000 km		0.06 km/sec	30 min–6 hr
Supergranules	20,000–35,000 km	0.5 km/sec	<0.02 km/sec	1–2 days

cycle is best known for the increase and decrease of the dark areas in the photosphere known as sunspots, but is actually an ongoing transformation of the Sun's global magnetic field.

The increase in sunspot number is accompanied by increases in solar activity. This term can refer to the sunspot regions themselves (more generally known as *active regions*), but more specifically refers to phenomena which occur on time-scales of a few hours or less. The most energetic of these are *solar flares* and *coronal mass ejections*.

Numerical parameters characterizing various dynamic aspects of the Sun are presented in Table IV.

A. Sunspot Cycle

The most widely known parameter describing solar activity is the approximately 11-year duration, or period, of the cyclic increase and decrease in the occurrence of sunspots, as illustrated in Figs. 8 and 9. Figure 8 shows the sunspot number since 1749, clearly illustrating that different cycles vary in amplitude, length, and other details. Figure 9 shows the sunspot number for the most recent few cycles compared to the variations in two of the many other quantities which vary with the solar activity cycle, the total solar irradiance and the number of solar flares. The total solar irradiance, or light-energy output, also increases with the number of sunspots. At first it is somewhat surprising that this quantity, which is dominated by the Sun's output in visible light, should increase as the number of dark sunspots increases. However, sunspots are accompanied by increases in bright features known as plages and faculae. Plage, visible in the chromosphere in Hα, and faculae, visible in the photosphere, are made up of small magnetic flux tubes in active regions and along the boundaries of supergranules cells (the *network*). They appear in active regions before sunspots are formed and remain after the sunspots have have disappeared. Their brightness offsets the decrease in brightness of the sunspots, leading to a small net increase in the Sun's light output at solar maximum. Other quantities which increase are the number of solar flares, brightenings which are commonly observed in X-rays, and coronal mass ejections.

TABLE IV Time Scales Characterizing Solar Activity

Sunspot cycle	9–13 years
Full magnetic cycle	18–26 years
Solar rotation period	
At solar equator	27 days
At 60° solar latitude	31 days
Lifetime of a solar granule	10 min
Period of solar oscillations	5 min
Duration of a solar flare	
Impulsive phase	Milliseconds to 10 min
Impulsive plus gradual phases	Minutes to hours

The 11-year increases and decreases are only half the story, however. The full pageant covers 22 years, with one-half of the full cycle being characterized by a magnetic dipole field that is the reverse of the field that characterized the other half.

The effect of the global variation in the Sun's magnetic field over the solar cycle can be seen in solar eclipse images in Fig. 10. During solar minimum (Fig. 10a) the field bears a resemblance to that of a dipole (or bar magnet) aligned with its axis of rotation (although in reality the field is a little more complex than that). During this time it has an easily measurable north and south pole and relatively simple large-scale structures with coronal holes at the poles, streamers near the equators, and few sunspots. The first spots that appear during an 11-year cycle generally occur at solar latitudes 20–30° above or below the solar equator.

As a cycle progresses, both the number of sunspots and the observable magnetic complexity of the groups or active regions that they form increase. During this same time, the Sun's dipole field weakens, and the zones of sunspot activity migrate toward the solar equator so that by solar maximum the sunspot bands have shifted to 5–20° north and south of the solar equator. When this migration phenomenon is represented as a plot of sunspot latitude versus time of occurrence, the result is commonly referred to as a butterfly diagram, with symmetric wings above and below the solar equator. By the time of sunspot maximum (Fig. 10b), the large-scale solar magnetic field is very complex. Streamers and coronal holes can be seen at any and all latitudes. After solar maximum the Sun returns to solar minimum conditions, but now with the north and south magnetic poles reversed. The old activity bands shrink

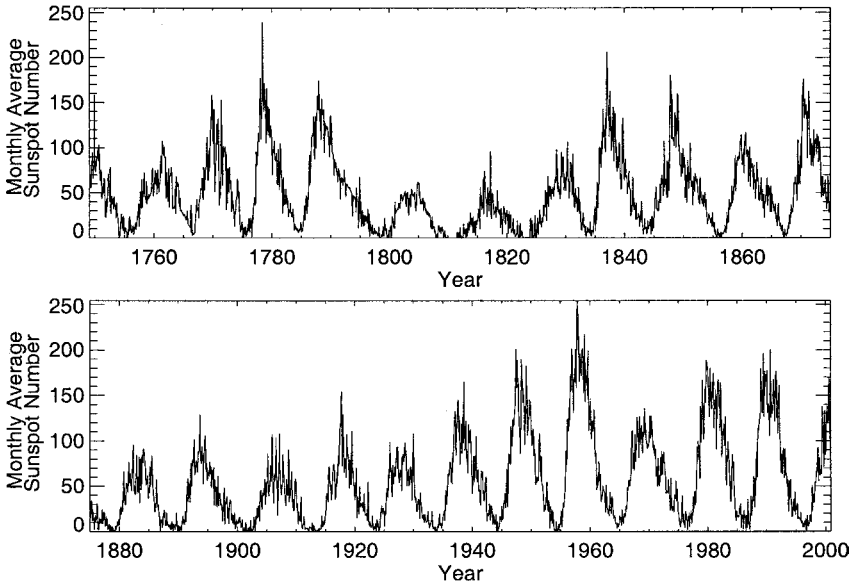


FIGURE 8 The Wolf Sunspot Number 1749–2000 AD. The sunspot number $= k(10g + s)$ where g is the number of sunspot regions, s is the number of spots, and k is a variable quantifying observing conditions and observer biases. Note that the sunspot cycle varies in both intensity and length.

toward the equator, eventually forming bands at latitudes from 5 to 10°.

Variations in the sunspot number have been observed on much longer time scales, as well. A variation with a 90-year period, known as the Gleissberg cycle, has been reported for the number observed at sunspot maximum. Researchers studying these phenomena have suggested that the solar dynamo may cause related changes in the solar luminosity and, consequently, changes in the Earth's climate as well. For instance, between 1645 and 1715 there was

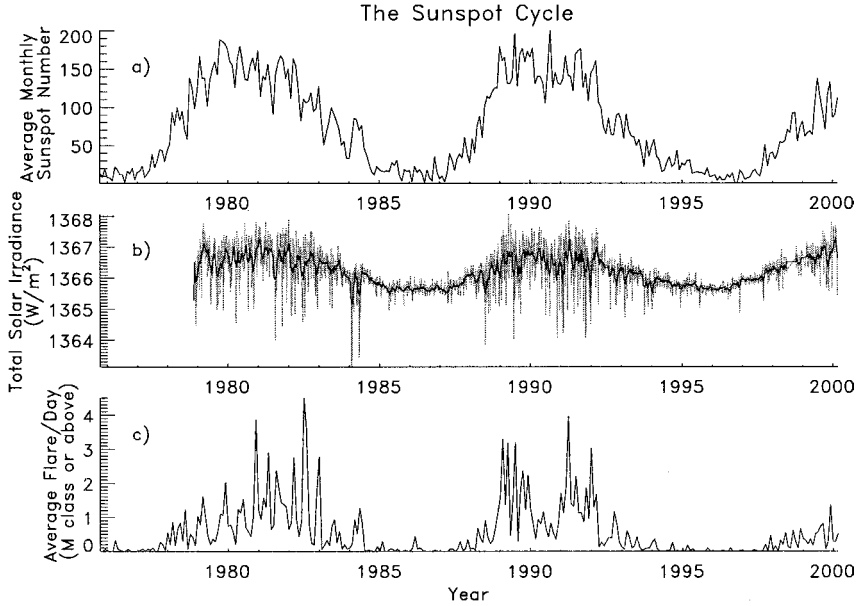


FIGURE 9 (a) Plot of the sunspot number for 1975–2000. Details of the cycles vary considerably from one cycle to another, but the approximately 11-year periodicity is known to extend back more than 2 centuries. Sunspot numbers were collected by the Royal Greenwich Observatory and Solar Optical Observing Network. (b) Plot of total solar irradiance. This is a composite of data from instruments on four spacecraft: Nimbus-7/ERB, SMM/ACRIM I, UARS/ACRIM II, SOHO/VIRGO. The gray curve shows daily values, the dark one the result of a running 30-day average. Data provided by C. Fröhlich. (c) Plot of average solar flares per day over the same time period. Flares counted had peak fluxes of over 10^{-5} W/m^2 in the 1–8 Å band of the GOES satellite X-ray flux detectors.

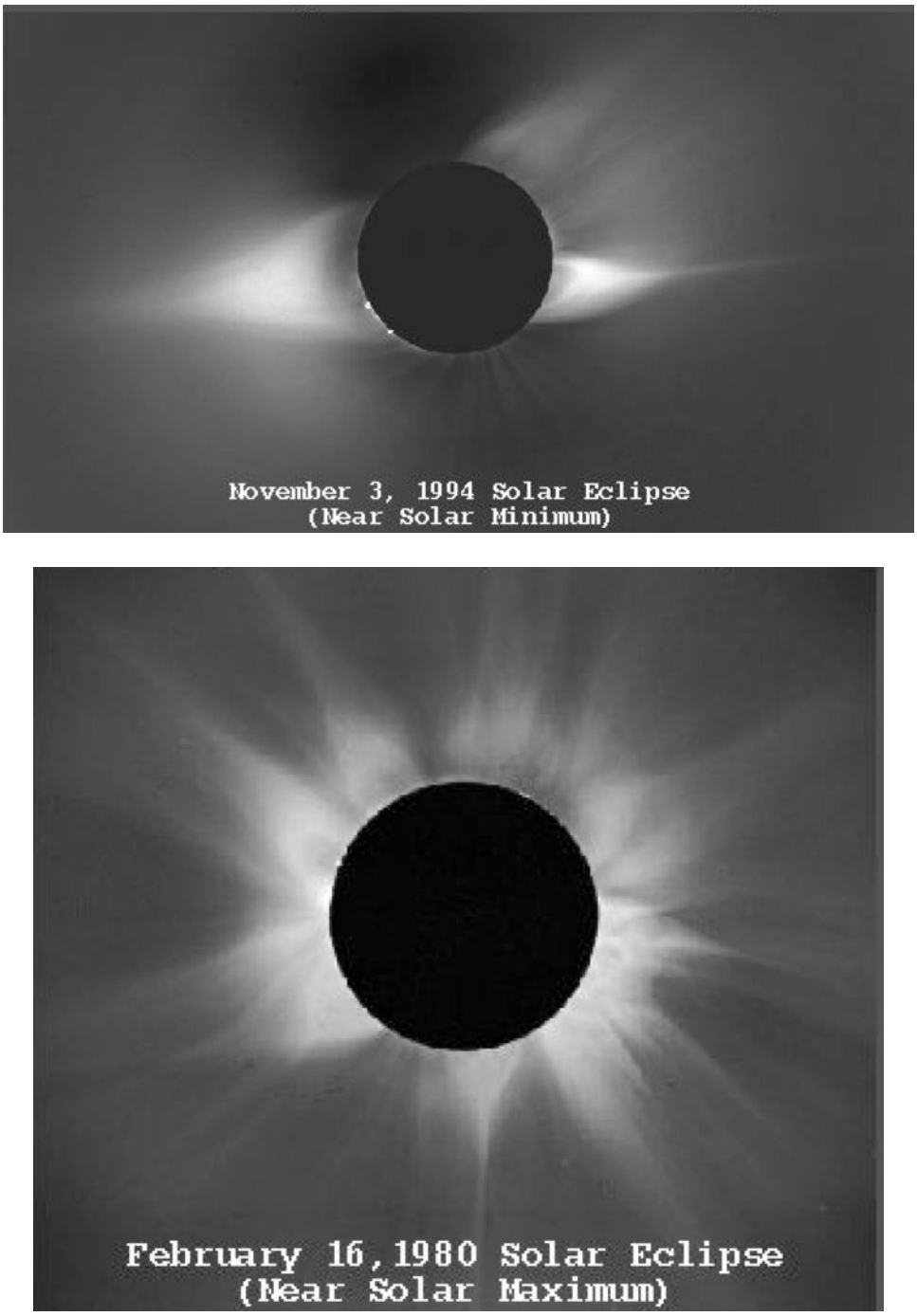


FIGURE 10 The solar corona observed during eclipses at solar minimum (1994) and solar maximum (1980). Note the change in shape from something similar to a dipole at solar minimum to a more complex field at solar maximum. 1980 eclipse image courtesy Rhodes College, Memphis, TN, and High Altitude Observatory (HAO), University Corporation for Atmospheric Research, Boulder, CO. 1994 eclipse image courtesy HAO. UCAR is sponsored by the National Science Foundation.

a well-documented near-disappearance of sunspots from the Sun. This apparent lull in solar activity, known as the Maunder Minimum, coincided with colder temperatures in the Northern Hemisphere. Other indications also exist of connections between Earth's climate and variations in solar emissions. However, the exact nature of the links and the mechanisms behind them has not been determined. This is an important area of current and future study.

B. The Role of Internal Magnetic Fields

The visible structure of sunspots and their periodic migrations are manifestations of magnetic dynamo activity in the outer layers of the Sun. This physical process is illustrated in Fig. 11. To understand these diagrams and the significance of the magnetic polarities, in particular, the reader should keep in mind that magnetic fields always form closed loops. The large-scale dynamo of the solar magnetic field is one in which the poloidal (dipole-like) field of the Sun is converted by differential rotation into a toroidal field and then back into a poloidal field of opposite polarity by the twisting of buoyant magnetic flux tubes.

In Fig. 11a, the field lines shown are poloidal ones connecting the north and south poles of the Sun inside the surface. Differential rotation stretches these field lines to the right, as shown in Fig. 11b, giving them a longitudinal component. This is known as the ω-effect. Wherever these field lines emerge from the surface, they form a new pole with polarity opposite to that of the pole from which they originated. Where they reenter, they form a pole with the same polarity as the original pole, as illustrated in Fig. 11c. These regions of both emerging and reentering flux are observed as sunspots.

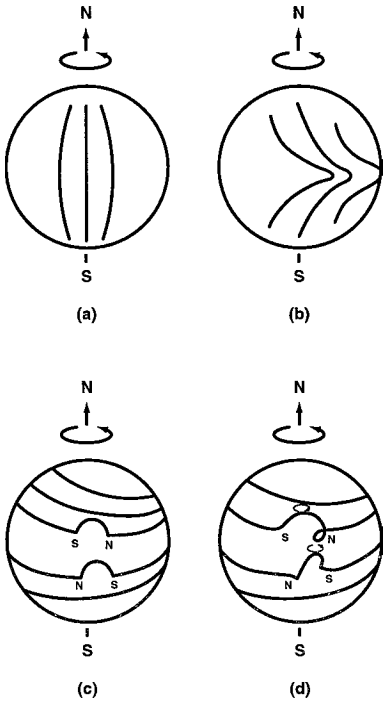


FIGURE 11 Magnetic fields in the surface of the Sun under the influence of a differentially rotating atmosphere. The material that comprises the Sun's convection zone rotates at a rate that is more rapid near the solar equator and decreases with increasing solar latitude. Field lines that are initially parallel to meridians, as in (a), are stretched and distorted, as in (b), by this differential rotation. The resultant increase in the magnetic field perpendicular to the Sun's rotational axis causes the ionized material in the vicinity of the field lines to become less dense and, hence, more buoyant, as in (c). The convective rising and expanding motion of magnetic flux tubes in the rotating reference frame of the Sun causes them to twist, an effect known as the Coriolis force, as in (d). This then provides a mechanism for the field lines perpendicular to the Sun's rotation axis to delete themselves by reconnecting, leaving behind a magnetic field parallel to the rotation axis but with a reversed north–south polarity.

As can be seen in Fig. 11c, a directivity is imparted to the polarity of these sunspot pairs relative to the direction of solar rotation. In each hemisphere of the Sun, the polarity of a leading spot is the same as that of the dominant dipole field in that hemisphere. The polarity of a spot trailing relative to the direction of rotation is the opposite. The effect of the Sun's rotation on the rising magnetic flux tube causes it to twist so that the newly emerging poles separate laterally. A mechanism is thereby provided by which the longitudinal fields can reform themselves along the north–south direction, but with a polarity opposite to that of the original solar dipole field. This is the α-effect portion of the dynamo. This dynamo activity explains both the cyclic 11-year variations in sunspot number and the fundamental 22-year cycle that includes reversals of the Sun's dipole field.

The "α–ω" dynamo described here is of course simplified, but most dynamo models use the basic concepts described. In detail, the theories are more complex. For instance, the model depends on reconnection of magnetic fields (see Section III.D), for which the effective diffusion must be sufficiently large. However, as mentioned in Section I.D.1, magnetic diffusion is a relatively slow process under solar conditions. With a time scale of 22 years and a length scale comparable to the size of the convection zone the relation $\tau_{\text{dif}} \sim l^2/\eta$ requires a diffusion coefficient, η, on the order of 10^{11} cm^2/sec. However, for typical convective layer temperatures of about 10^6 K, $\eta \sim 4 \times 10^3$ cm^2/sec. Because of this discrepancy, it is commonly assumed that turbulent fluid motions cause magnetic fields to reconnect more rapidly and are therefore a key contributor to the dynamo.

In the last decade helioseismology studies of the flows and the turbulence characteristics of the solar interior have provided important new constraints on dynamo models. One key discovery has been that there is a great amount of radial shear in rotation rate near the tachocline at border of the radiation and convection zones, but very little in the convection zone itself. This shear is needed for the ω-effect reconnection. Furthermore, the relative stability of this border area allows the necessary time needed for this effect to take place. As a result, most models now assume that at least the ω portion of the dynamo occurs in this area. The location of the α-effect process is still

a matter of debate. Dynamo modelers grapple with this and other issues as they continue their progress toward models which fully describe the observed characteristics of exterior solar activity and the solar interior.

C. Solar Wind and Coronal Holes

Electrons and ions stream out from the solar corona along trajectories defined by the Sun's magnetic field in interplanetary space. The Sun's rotation causes the magnetic field lines, and hence the particle streams, to follow a spiral pattern. This flow, known as the solar wind, was first deduced from theoretical considerations and subsequently understood to be the force that orients comet ion tails. Because the flow is actually an expansion of the solar corona and is continuously moving outward, its effect on comets is to sweep their tails away from the Sun.

The solar wind is not uniform. Magnetically confined eruptions of various sizes known as coronal mass ejections are regularly seen. These are discussed in more detail later in this article (Section III.E). In addition, the wind seems to divide between areas of relatively slow speeds and high densities (300–600 km/sec, 10 protons/cm^3) and high speeds and low densities (700–800 km/sec, 3 protons/cm^3).

The areas of high-speed wind can be traced back to areas in the Sun's corona known as *coronal holes*. These are dark regions which can be seen in images of the Sun taken in ultraviolet and X-ray. Coronal holes were first observed by NASA's Orbiting Solar Observatory (OSO)-6 satellite, and the holes and their connections to the solar wind have now been extensively studied by UV and X-ray imagers and *in-situ* solar wind sensors on board many spacecraft. A large coronal hole is seen to the north in Fig. 12's image from the Yohkoh satellite.

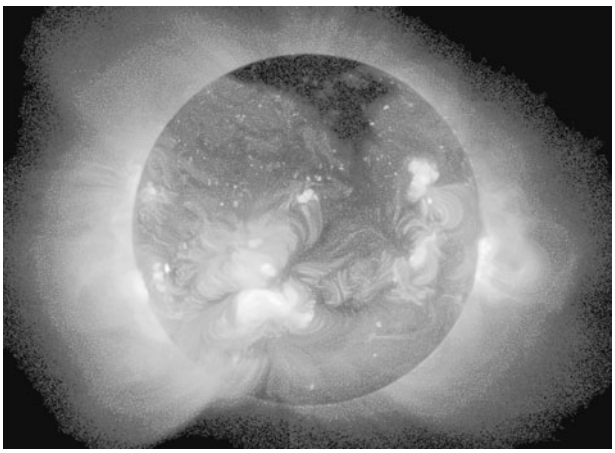


FIGURE 12 A large coronal hole is observed in the north in this image taken from space by the Yohkoh Soft X-ray Telescope on May 8, 1992. (Courtesy ISAS and NASA.)

These features, like the magnetic streamers, are often visible during a solar eclipse. Because the magnetic field of the Sun is continuously evolving, coronal holes are continuously changing and moving. During sunspot minimum when the solar dipole fields are strongest, coronal holes are invariably found at both of the solar poles. The connection of the coronal holes with fast solar wind has been further confirmed by the Ulysses spacecraft which has flown over the Sun's poles and measured the wind coming from the polar coronal holes.

A coronal hole is formed by magnetic field lines that stream out of the Sun to a connection point in interplanetary space rather than to a magnetic reconnection point near the Sun's surface. Solar wind plasma can flow easily out into the solar system from such regions, following the magnetic field lines.

This difference in magnetic configuration probably leads to another interesting difference between high and low speed solar wind—that of elemental abundances. As described in Section I.B, when compared to photospheric abundances, elements with low first ionization potentials (FIPs) are relatively more abundant in the corona than elements with high FIPs. The FIP effect is at a similar level in the slow speed solar wind and in close magnetic field structures in the corona, but is less pronounced in the high speed solar wind.

D. Solar Flares

The role of magnetic fields in solar flares is less well understood, but what is known leaves no doubt that magnetic fields provide the sources of energy and define the settings in which flares occur. Flares are transient events in which energetic particles and electromagnetic radiations, ranging from meter waves to γ-rays, are produced. Their rate of occurrence follows the sunspot cycle, with the most energetic flares occurring in the most complex magnetic active regions. The total energy released in a flare is greater the closer the flare site is to a sunspot.

Flares produce both increases in the emissions that can be observed continuously, even from the quiet Sun, and great bursts of high-energy emissions that are associated only with flare events. The amount of energy released from the Sun in a large flare, the size observed approximately once a month during the 3 years following sunspot maximum, is 10^{32} erg. This huge amount of energy (about 25,000 times global human energy consumption in the year 2000), represents only a few thousandths of a percentage increase in the total solar luminosity.

Our understanding of solar flares is based on a combination of temporal, imaging, and spectral data at almost all wavelengths as well as *in situ* solar wind data. Common properties of different wavebands of flare emission

TABLE V The Fast and Slow Solar Winds

Component	Speed (km/sec)	Density (particles/cm^3)	Abundances $\left(\frac{Fe}{O}\right)_{wind} / \left(\frac{Fe}{O}\right)_{phot}$	Source
Slow	300–500	3–20	3–5	Streamers (closed field regions)
Fast	700–800	2–3	1.5–2	Coronal holes (open field regions)

are listed in Table VI. It is not possible here to detail all the different observations that make up our understanding of flares, but we will discuss some of the important ones.

A key observation in understanding the particle behavior of flares is the comparative timing of microwaves, hard ($\gtrsim$10 keV) X-rays, and soft ($\lesssim$10 keV) X-rays. An example of such time profiles is shown in Fig. 13. Commonly, the time behavior of hard X-rays, γ-rays and microwaves are well correlated in time, coming in short ($\lesssim$1 min) bursts in the initial impulsive phase of the flare. Soft X-rays, on the other hand, increase approximately in proportion to the time integral of the hard X-rays and can last for tens of minutes or even hours in what is called the gradual phase of the flare.

Images indicate that flaring loops exhibit prominent hard X-ray emissions at their footpoints, while microwaves and soft X-rays generally come from the loops themselves. Some flare observations also show hard X-rays at the loop tops during the impulsive phase, as is shown in Fig. 14.

From this a scenario for a typical impulsive type flare, illustrated in Fig. 15, may be deduced. Some acceleration mechanism(s) injects relativistic electrons, protons, and heavier particles into the coronal loops. This may be occurring at the site of the hard X-ray loop-top sources seen in hard X-rays. Electron motion in the loops is a combination of gyration around the field lines, which produces microwave gyrosynchrotron radiation, and motion along the field lines, which causes energetic particles to collide with the dense chromosphere at the loop footpoints. This results in emission of X-ray bremsstrahlung, or breaking radiation, at the foot-points. Gamma-rays are produced at the same time by a number of processes involving interaction of accelerated protons and ions with the chromospheric plasma. The impact of the energetic particles leads to heating of the chromosphere so that hot plasma flows up into the loops where it can remain for hours, emitting soft X-rays in addition to radiation at ultraviolet, Hα, and other wavelengths. Energetic particles can also be accelerated into magnetic field structures which open out into interplanetary space and are measured by *in situ* solar wind monitors (see Table VII and Section III.E).

Not all flares fit this picture, although it seems to be the most common type. Some flares exhibit X-rays and γ-rays for over an hour. These long-duration events, referred to as gradual events, are often associated with CMEs (see Section III.E), and presumably involve a continuing acceleration mechanism beyond that which is operating during the impulsive phase of shorter flares.

Spectra can tell us even more about the different populations of high energy particles involved in flares.

TABLE VI Commonly Observed Flare Characteristics in Different Wavebands

Waveband	Emission mechanism	Physical source
Decimeter and meter waves	Coherent plasma wave emission	Electron beams in open or closed magnetic structures
Millimeter and centimeter waves	Gyrosynchrotron emission	Gyrating electrons in flaring loops
Broadband visible (uncommon)	Undetermined. Perhaps line excitation and thermal emission	Mostly at footpoints, probably due to electron bombardment or X-ray radiation
Hα	Line emission	Footpoint emission due to beamed electrons during impulsive phase loop emission during gradual phase
Ultraviolet	Line and continuum emission	Impulsive footpoint emission and/or gradual phase hot loop emission depending on the formation temperature of the emitting ion
Soft X-rays	Thermal bremsstrahlung	Gradual phase loop emission
Hard X-rays	Nonthermal and thermal bremsstrahlung	Footpoints and sometimes loop-tops during impulsive phase
γ-rays	Bremsstrahlung, Atomic bound-free emission nuclear line emission, electron–positron annihilation, pion decay, Compton backscatter	Impact of electron, proton, and ion beams at footpoints and possibly in the corona (not yet actually imaged)

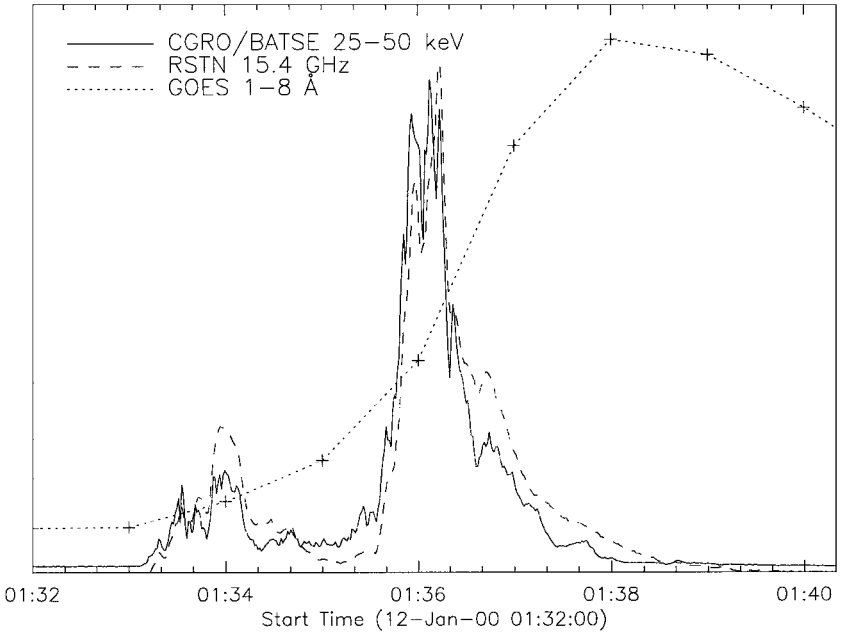


FIGURE 13 Time histories of the hard X-ray, microwave, and soft X-ray emissions observed from a solar flare on Jan. 12, 2000. Hard X-ray data are from the BATSE instruments on the Compton Gamma Ray Observatory, microwave from Radio Solar Telescope Network, and soft X-ray from a GOES spacecraft. The hard X-ray and microwave time profiles, although not identical, are very similar, suggesting they are produced by the same, or closely related populations of particles. Soft X-rays are proportional to a rough time integral of the other emissions. Although not universal, this pattern is evident in most impulsive type flares.

Figure 16 shows a model high-energy photon spectrum for a flare. At X-ray wavelengths emission is due to the X-ray bremsstrahlung mentioned previously. The evaporated thermal electrons dominate the lower energies, while impulsive X-rays due to nonthermal and high-temperature thermal sources lead to higher-energy X-rays. At γ-ray energies, accelerated photons and even heavier particles lead to line and continuum emission through a number of different mechanisms (see Table VI). For these cases it is often useful to compare X-ray data to radio spectra produced by the same populations of electrons as they gyrate about field lines or collide with ambient electrons.

The properties of flare spectra vary from flare to flare and over the course of a given flare. Some flares seem to be relatively richer in protons than others. In some the electrons seem more consistent with a super-hot thermal interpretation, while others seem to have a greater proportion of nonthermal electrons.

Much work has been done to try to understand the processes by which energy is transferred from the magnetic field to these energetic particles. Models involving shocks, electric, fields, turbulence, and/or transport of energy through waves are all being considered, although none is able to explain all the data in detail. In general it is not possible to deduce directly from a photon spectrum a unique particle spectrum. Instead modelers can only confirm that a particular model of the particle energy distribution could cause a particular photon spectrum. This makes it difficult to determine which processes energize primary flare particles.

One process which seems very likely to be important, however, is reconnection. In this process magnetic field lines reconnect to form a lower energy configuration, releasing the difference in energy. Figure 17 shows a simple example of reconnecting magnetic flux tubes. When plasmas containing oppositely directed magnetic field lines come into contact strong currents can result. With any resistivity at all this results in a break down of the "frozen in" condition mentioned in Section I.D, a reconfiguration of the field and energy release. Reconnection can happen at many scales, more than one of which may be important in flares.

Future progress on understanding solar flares will depend on combinations of higher temporal, spatial, and spectral resolution which can help us learn more accurately the characteristics and locations of the different particle populations involved.

E. Coronal Mass Ejections

Coronal mass ejections (or CMEs) are giant eruptions of plasma, confined by magnetic fields, which blow out of the solar corona and into space. CMEs move at speeds from

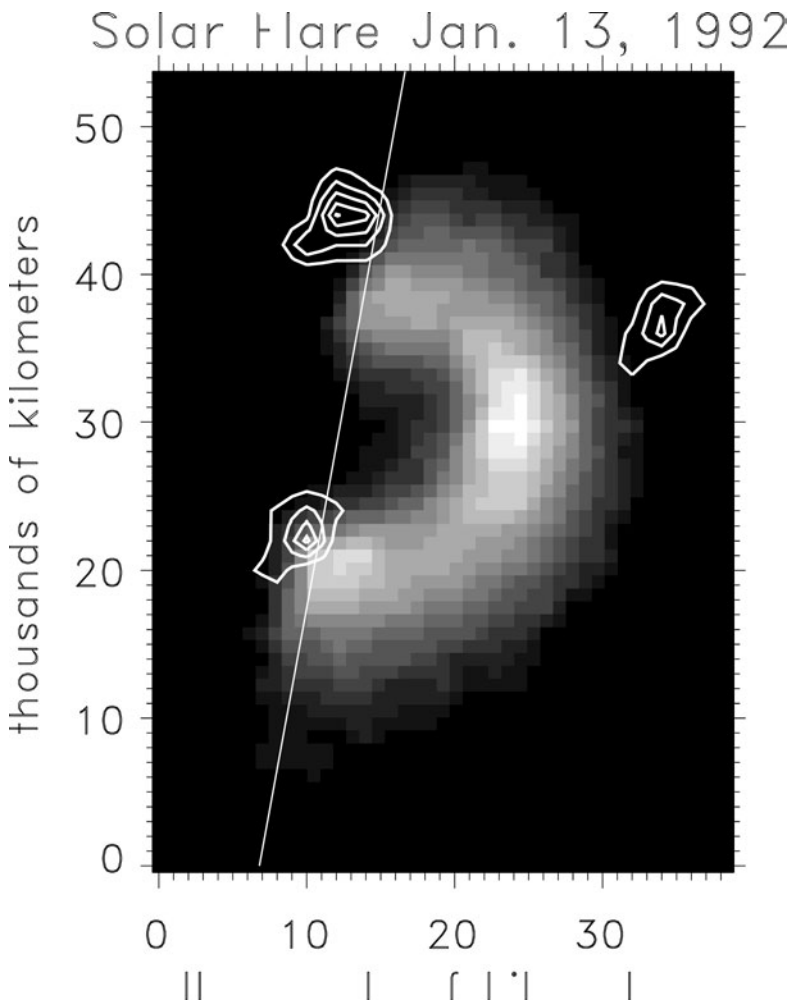


FIGURE 14 Image of a flaring magnetic loop in X-rays at Sun's limb. The grey-scale image shows relatively low-energy X-rays (0.25–4.0 keV) observed by the Yohkoh Soft X-ray Telescope. The contours show the location of higher energy X-rays (32.7–52.7 keV) observed by the Yohkoh Hard X-ray Telescope. The diagonal line shows the edge of the Sun, which is to the left. The loop top is oriented toward the right. (Courtesy ISAS and NASA.)

100 to over 1000 km/sec (averaging ≈425 km/sec) and can expand to over 10 times the size of the Sun in a matter of hours. Magnetic field and particle density variations associated with CME's have been measured as far away from the Sun as 4.5 AU with instruments on board the Ulysses spacecraft. Mass estimates for large, fast CME's can be over 10^{16} g with energies over 10^{32} ergs. The number of CME's ranges from less than 1 a day at solar minimum to over 3.5 a day at the maximum of the solar cycle. They chiefly occur in equatorial regions during solar minimum, but have a broader latitudinal distribution as solar activity and the complexity of the Sun's magnetic field increases during solar maximum.

Although it is possible to observe them with other instruments, CMEs are most commonly observed using visible-light coronagraphs which block out the light from the Sun's disk so that the faint outer corona is visible. This is most easily done from space, where the scattering of light by the Earth's atmosphere need not be considered. As a result, CMEs were not discovered until 1973, when they were observed with a coronagraph on board OSO-7. They have since been observed by many spacecraft including Skylab, SMM and SOHO, to name some prominent examples.

Figure 18 shows coronagraph observations from the LASCO instrument aboard SOHO. The dark disk in the center is the occulting disk, covering the disk of the Sun which is shown by the white circle. In this case the leading edge of the CME is followed by the bright core of an erupting prominence.

CME's have also been observed using UV and X-ray imagers of the corona, which have shown blast waves

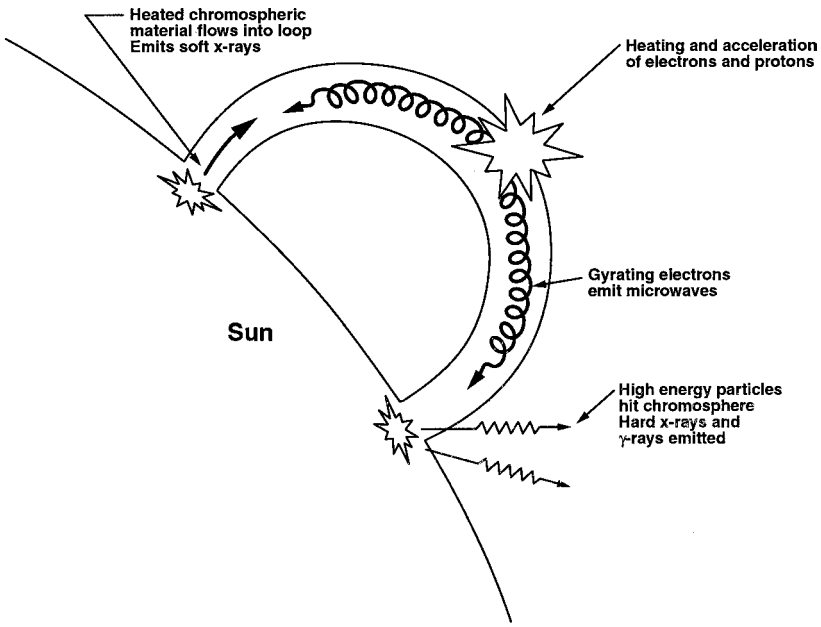


FIGURE 15 Schematic of an impulsive type flare. Electrons, protons, and ions are accelerated into a coronal loop or, more commonly, loops. These particles gyrate down magnetic field lines, producing microwaves as they go. When they hit the dense loop footpoints hard X-rays and γ-rays are emitted. The impact heats the chromospheric plasma at the footpoints, which expands up into the loop where it emits soft X-rays.

which radiate across most of the visible surface of the Sun from the location of CME lift off. Radio instruments observe the shock waves they produce in the solar wind, and they are detected as changes in the solar wind and accompanying interplanetary magnetic field by *in situ* spacecraft instruments outside the Earth's magnetic fields.

CMEs are often accompanied by some amount of brightening on the disk (i.e., at least a small flare). CME-associated flares are usually "gradual" flares, with longer-lived X-rays, and sometimes even γ-rays, than the shorter impulsive flares. CMEs are also associated in many cases, but not all, with erupting prominences, discussed in Section III.F).

CMEs lead to acceleration of energetic particles in the solar wind. CME particle events show different characteristics than those associated with impulsive flare events, as is summarized in Table VII. They are longer lasting, cover a broader area, and have lower electron to proton and ^{3}He/^{4}He ratios and fewer heavy ions, such as iron. It is thought that these streams originate in broad shocks associated with the CMEs rather than in flares.

TABLE VII Solar Energetic Particle Events

	CME associated	Flare associated
Time scale	Days	Hours
Size (solar longitude)	60–180°	30–45°
e/p	Proton rich	Electron rich
Fe/O	∼0.1	∼1
^{3}He/^{4}He	∼0.0005	$\gtrsim$0.1
Fe charge state	∼+14	∼+20
X-rays	Gradual	Impulsive

Starting in the 1990s, new observations have also allowed us to regularly observe CMEs that are directed toward the Earth and to monitor their effects on our planet and our space environment. CMEs directed toward the Earth or away from it expand so that they can be seen to form a halo around the Sun and are thus called halo CMEs. Combined observations from many spacecraft, including those of the International Solar and Terrestrial Physics Program (see Fig. 19), have allowed us to track these events and other variations in the solar wind from the Sun all the way to the Earth's magnetosphere, a journey that takes 2–4 days. Disturbances in the solar wind can cause many problems, including damage to spacecraft and power outages on Earth.

Like flares, the exact causes of CMEs have not yet been determined. It appears that energy stored in the magnetic field is being transformed into kinetic energy causing a large magnetic structure to lift off from the Sun. The occurrence of a CMEs at the Sun cannot be predicted at this time, although the approximately 3-day travel time of a CME from the Sun to the Earth makes it possible to prepare for some of its possible effects.

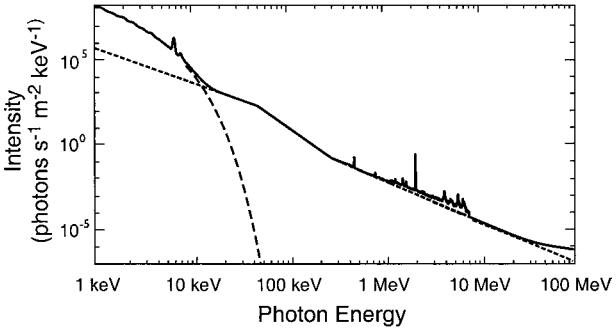


FIGURE 16 A model flare spectrum showing emissions in soft X-rays, hard X-rays and γ-rays. The line with longer dashes shows a soft X-ray thermal spectrum. The line with shorter dashes is a typical hard X-ray bremsstrahlung spectrum. The solid line is the sum of these plus the contribution of nuclear line emissions and pion decay.

F. Erupting Prominences

Another type of eruptive behavior generally associated with flares and CMEs is the ejection of structures knows as prominences. Prominences are long rope like structures of relatively cool (mostly 5000–8000 K) material suspended in the corona between areas of opposite magnetic polarity. They can remain generally stable for months, but can suddenly disconnect and lift off with billions of tons of material accelerated to speeds of many tens of kilometers per second. This usually, but not always, occurs in conjunction with a flare and/or CME. In many cases they reform within hours after eruption.

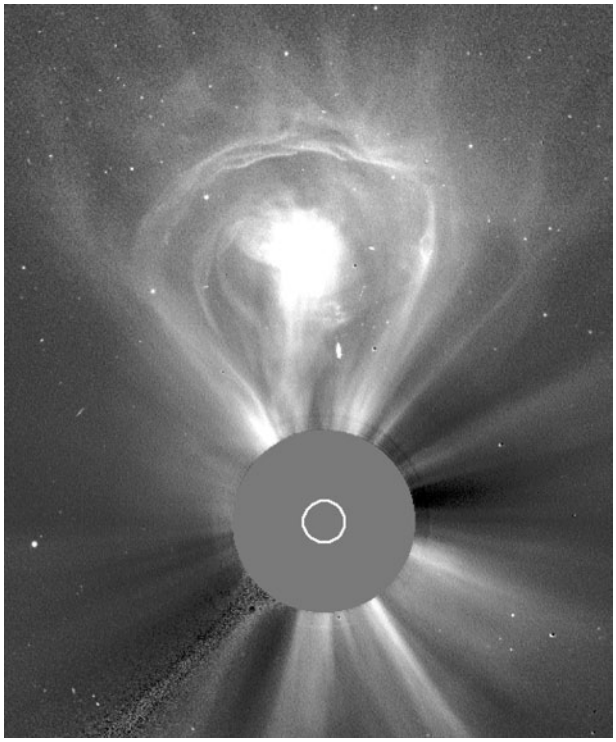


FIGURE 18 A huge coronal mass ejection (CME) shoots almost straight north from the Sun in an image taken by the LASCO coronagraph aboard the SOHO spacecraft. This event occured on Feb. 27, 2000. The grey disk is the occulting disk which blocks the bright disk of the Sun so that we can observe the faint corona. The white circle in the center shows the location of the Sun's disk. The pylon supporting the occulter is on the bottom left. Image courtesy of ESA and NASA.

IV. OUTSTANDING QUESTIONS

A. Heating of the Corona

The question of why the corona is hot has been central to solar physics for decades. It now seems well established that the mechanism involves the Sun's magnetic field, but the exact mechanism has still not been determined—waves, magnetic field reformation, or something else? In recent years there has been a new emphasis on heating due to ubiquitous small-scale processes which occur all over the Sun. Actual solar heating is probably taking place on relatively small spatial scales of less than 10 km, over an order of magnitude smaller than our current resolution. It may take instrumentation which can resolve such scales before we can be sure of the heating mechanism.

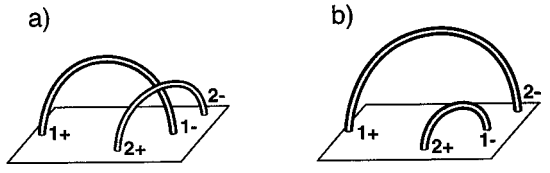


FIGURE 17 Simple schematic showing reconnection in two magnetic loops. In the first state, shown in (a), loops connect areas of opposite polarity, shown by the $+$ and $-$ signs. In the later state in (b) the field has reconfigured itself so that the areas of opposite polarity are connected differently and the energy stored in the magnetic field is lower, the energy difference having been given off in the reconnection.

B. Source and Structure of the Solar Wind

A related question concerns the source and structure of the solar wind. The solar wind can be seen as the expansion of the hot corona. However, the wind is very structured in ways that seem to be related to the magnetic field. How does the magnetic field promote or retard the wind? How are the origins of the slow wind from closed magnetic field regions different from the fast wind from open regions (coronal holes)? At what altitudes is the wind accelerated and what related processes are occurring in the photosphere, chromosphere, and lower corona?

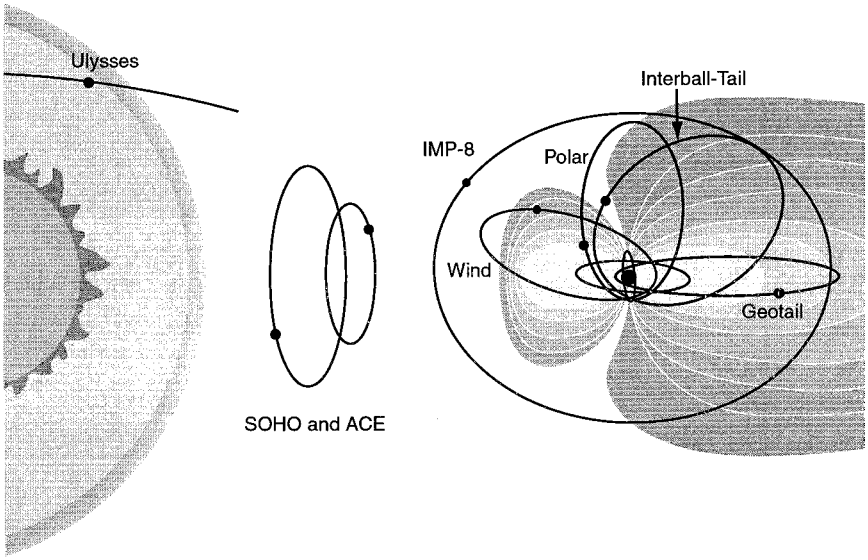


FIGURE 19 Schematic drawing of some of many the spacecraft and orbits used to study the Sun and its interactions with the Earth. The Earth's magnetic field is also shown.

C. Cause of the Solar Cycle

There are still key questions concerning why the Sun and other stars have magnetic cycles. What determines the length and intensity of individual cycles? What causes long noncyclic periods like the Maunder Minimum? Can such variations in the cycle be predicted?

With new helioseismology studies we are finally able to start studying the inner structure of the Sun and how it varies. As these studies continue over the next few solar cycles we should obtain the data needed to inspect the inner workings of the solar dynamo and actually understand the cause of the Sun's magnetic cycle.

D. Prediction of Solar Activity and Its Effects on Humans

Although there are a number of theories, the exact causes of flares, CMEs and other short time-scale solar activity have not been determined; nor have the relationships between these phenomena. Other important areas of study concern the effects of solar activity on humans. Will it be possible to predict solar activity? When will flares and CMEs happen and, once they have happened, which are likely to be most dangerous? What can we do to protect vulnerable technology and, especially, humans in space? With respect to longer time-scale variations in activity, we have yet to understand the effects of variations of the Sun on the Earth's climate.

SEE ALSO THE FOLLOWING ARTICLES

NEUTRINO ASTRONOMY • SOLAR SYSTEM, MAGNETIC AND ELECTRIC FIELDS • SOLAR TERRESTRIAL PHYSICS • STELLAR STRUCTURE AND EVOLUTION • ULTRAVIOLET SPACE ASTRONOMY • X-RAY ASTRONOMY

BIBLIOGRAPHY

Choudhuri, A. R. (1999). "The solar dynamo," *Curr. Sci.* **77**(11), 1475.

Foukal, P. V. (1990). "Solar Astrophysics," Wiley, New York.

Golub, L., and Pasachoff, J. M. (1997). "The Solar Corona," Cambridge University Press, New York.

Hoyt, D. V., and Schatten, K. H. (1997). "The Role of the Sun in Climate Change," Oxford University Press, New York.

Lang, K. R. (1997). "Sun, Earth, and Sky," Springer-Verlag, Berlin.

Parker, E. N. (2000). "The physics of the sun and the gateway to the stars," *Physics Today* **53,** 26.

Phillips, K. J. (1992). "Guide to the Sun," Cambridge University Press, New York.

Tayler, R. J. (1997). "The Sun as a Star," Cambridge University Press, New York.

St. Cyr, O. C., *et al.* (2000). "Properties of coronal mass ejections: SOHO LASCO observations from January 1996 to June 1998," *J. Geophys. Res.* **105,** 18169.

Strong, K. T., Saba, J. L. R., Haisch, B. M., and Schmeltz, J. T. (1999). "The Many Faces of the Sun: A Summary of Results from NASA's Solar Maximum Mission," Springer-Verlag, New York.

Zirin, H. (1989). "Astrophysics of the Sun," Cambridge University Press, New York.

Solar System, General

John S. Lewis
University of Arizona

GLOSSARY

Accretion The agglomeration of particles of condensed preplanetary materials to form larger bodies of asteroidal and planetary size.

Achondrite Any meteorite type that has melted and differentiated into a silicate-rich crust-like or mantle-like composition.

Asteroid A small body, less than 1000 km in diameter, primarily composed of rocky solids, in direct orbit around the Sun.

Asteroid belt A zone, mostly confined to the region between 2.2 and 3.3 times Earth's distance from the Sun (2.2-3.3 AU), within which most asteroids are found.

Astronomical unit (AU) The mean distance of Earth from the Sun (149,597,870 km).

Brown dwarf A low-mass star (0.013 to 0.07 of the mass of the Sun) that is capable of fusing deuterium but not massive enough and hot enough to fuse normal hydrogen.

Carbonaceous chondrite A type of primitive meteorite, rich in water, carbon compounds, and other volatile and oxidized materials, that shows little evidence of heating and has not melted since its formation 4.6 billion years ago.

Centaurs Objects of dirty ice, in moderately eccentric, moderately inclined orbits around the Sun that typically cross the orbits of two or more of the giant planets.

Chondrite Any of the most common stony meteorite types, dominated by silicates, sulfides, and native metal, that have never melted and differentiated according to density, and which generally contain small glassy beads called chondrules.

Comet A small body in orbit around the Sun that exhibits an envelope of glowing gases and dust when strongly heated during close approaches to the Sun.

Differentiation The process by which melted material separates into chemically distinct layers of different density, such as crust, mantle, and core, under the action of gravity.

Escape velocity The minimum speed with which an object must depart from the surface of a massive body in order to coast to infinite distance.

Galilean satellites The four large satellites of Jupiter (Io, Europa, Ganymede, and Callisto) that were discovered by Galileo in 1610.

Giant planet Any planet in which volatile materials, notably hydrogen, helium, water, ammonia, methane, and neon, are present in total masses at least comparable with the mass of rocky materials.

Kirkwood gaps Gaps in the statistical distribution of orbital semimajor axes and orbital periods in the asteroid belt, corresponding to orbital periods that are harmonically related to the orbital period of Jupiter.

Kuiper belt A broad flattened disk of asteroid-sized bodies, apparently with cometary dirty-ice composition, orbiting the Sun beyond the orbit of Neptune.

Libration Slow oscillation back and forth about a mean, named for the swinging of an ancient pan balance.

Meteor An atmospheric phenomenon, a flash of light at high altitudes, caused by the vaporization of a small particle entering at very high speeds. By common extension, the (unseen) body whose demise constitutes the meteor. (See meteoroid.)

Meteorite Extraterrestrial solid matter that survives entry into an atmosphere and reaches a planetary surface intact and in recoverable condition.

Meteoroid A small solid body, too small to be considered an asteroid or comet, in orbit around the Sun. The impact of a meteoroid on a planetary atmosphere generates the flash called a meteor.

Oort cloud A vast cloud of cometary bodies in randomly oriented orbits around the Sun at distances as great as tens of thousands of Astronomical Units, where perturbations of their motions by nearby stars can be very important.

Planet A planetary body in orbit around a star.

Planetary body Any object of less than stellar mass (incapable of producing energy by fusion) but larger than an asteroid (roughly, and somewhat arbitrarily, set at about 1000 km in diameter).

Planetesimal Any small solid body in the early solar system, typically from 10 to 1000 m in diameter, which constitutes part of a growing planet.

Protoplanetary disk A flattened disk of dusty gases out of which the Sun and its retinue of planets, satellites, asteroids, and comets form.

Regolith Solid surface material that has been shattered into rock fragments and dust by impacts.

Resonance A harmonic relationship between the periods of different solar system bodies, including orbit–orbit, spin–orbit, and sometimes spin–spin resonances.

Ring A thin, flattened disk of small particles in orbit around a larger body.

Roche limit The minimum distance at which a small orbiting body can resist being torn apart by the tidal stresses of its primary.

Star Any body that is luminous at any wavelength by reason of fusion reactions.

Terrestrial planet Any planetary body principally composed of rock (silicates, oxides, sulfides, and metals).

THE SOLAR SYSTEM consists of one star (the Sun), planets with their satellites and rings, asteroids, comets, and meteoroids. Studies of the chemistry, physics, and geology of solar system bodies combine with astronomical studies of star formation to give complementary perspectives on the origin, evolution, and eventual fate of the solar system. Supplemented by recent detections of over 50 planets of other stars, solar system studies give us a preview of the nature of planetary systems in general.

I. OVERVIEW OF THE SOLAR SYSTEM

The principal objects of studies of the solar system are to understand the origin, evolution, and fate of planetary systems. Among the most important specific issues are the conditions under which the solar system formed, how and how fast planets accreted from the protoplanetary cloud, how surface conditions on the planets changed with time in response to changes in planetary interiors and in the Sun itself, how and when life first appeared in the solar system, and whether these conditions and events are common in the galaxy at large.

It is known from astronomical studies that stars have formed and are still forming, at a rate of roughly 10 stars per year in our galaxy, as a result of the collapse of enormously massive interstellar clouds of gas and dust. During collapse, clouds with initial masses of as much as tens of thousands of Suns repeatedly become unstable and fragment again, producing many thousands of stars and stellar systems in a dense, highly dynamic cluster. Study of our stellar neighborhood and of relatively nearby clusters of young stars suggest that roughly 30% of all stellar systems contain solitary stars, about 20% contain 2 stars, 15% have 3 stars, and so on up to at least 10 stars. Overall, only about 10% of all stars are solitary, like our Sun. The theory of star formation strongly suggests that the formation of a star from an interstellar cloud fragment occurs through an intermediate stage in which the star forms at

the heart of a vast, flattened protoplanetary disk of gas and dust. Possibly many double stars form from a similar disk that happens through accidents of history to possess enough angular momentum to hinder its collapse into a single star. The implication of the available observations of dust disks around young solitary stars and the theory of protoplanetary disk formation and evolution is that the preconditions for origin of planetary systems like our own may be common, if not universal, around young stars.

The nine planets of our own solar system fall naturally into two major families, terrestrial planets and giant (Jovian) planets. The rocky terrestrial planets, Mercury, Venus, Earth, and Mars, contain massive cores of iron–nickel metal and sulfides, deep mantles principally composed of silicates of iron and magnesium, and thin crusts of low-density rocks generally enriched in sodium, potassium, aluminum, and volatile materials. The volatiles, notably water, carbon dioxide, nitrogen, and argon, also contribute thin superficial layers of oceans and atmospheres where temperatures were not so high as to cause their loss. The giant planets, in contrast, contain dense cores of rocky and icy materials surmounted by massive atmospheres principally composed of hydrogen and oxygen, with substantial traces of water, ammonia, methane, neon, and other highly volatile materials. Of the giant planets, Jupiter and Saturn have nearly the composition of the Sun (roughly 95% volatiles compared to about 99% for the Sun), whereas Uranus and Neptune more closely resemble the composition of dirty ice, supplemented by captured hydrogen and helium. The giant planets have over 50 known satellites. Those that have been closely studied generally are ice–rock mixtures, although Io, the closest large satellite to Jupiter, has more the properties of a terrestrial planetary body. The ninth and outermost planet, Pluto, bears closer resemblance to the ice-bearing large satellites of the giant planets. It may best be thought of as the largest known representative of the Kuiper Belt swarm, in which its eccentric orbit is embedded. Pluto, like the largest Saturnian satellite, Titan, and Neptune's largest satellite, Triton, has an atmosphere and surface dominated by methane and nitrogen gases and condensates. The sizes, orbits, rotational states, magnetic fields, and atmospheres of the planets are summarized in Tables I–IV. Among the terrestrial planets, the interior composition and physical structure of Earth are the best determined. The oblateness (polar flattening) of Earth, the detailed structure of its gravity field, and seismic profiling of its interior combine to provide a well-determined interior structure. Earth's core is found to consist of a solid, dense inner core, apparently made of nickel-rich iron alloy, surrounded by a liquid outer core consisting of a melt of iron, sulfur, nickel, and many other rarer elements that have a chemical affinity for liquid iron or sulfides. The upper and lower mantle, dominated by dense silicates of ferrous iron (Fe++) and magnesium, are distinguished by a density discontinuity that reflect high-pressure phase changes. The crust, surmounting the mantle, is composed of two major compositionally distinct units, basalt-rich dense oceanic crust and iron-poor low-density continental crustal blocks, rich in silica, alkali metal oxides, aluminum, and water-bearing minerals. The crust in turn is substantially covered by oceans (the hydrosphere), which lies beneath an atmosphere mainly composed of nitrogen, oxygen, argon, and highly variable amounts of water vapor.

The interior structures of the other terrestrial planets are not nearly so well known. Nonetheless, their densities testify to significant differences in composition, and hence in mineral content and physical structure. Mercury is startlingly dense, attesting to a core with about 60% of the mass of the planet (compared to 31% on Earth). The density of Venus, correcting for self-compression by high interior pressures, is slightly less than that of Earth.

TABLE I The Planets[a]

Planet	Mass (10^{24} kg)	Equatorial radius (km)	Density (10^3 kg m^{-3})	Escape velocity (km s^{-1})
Mercury	0.3303	2,439	5.43	4.25
Venus	4.870	6,051	5.25	10.4
Earth	5.976	6,378	5.518	11.2
Mars	0.6421	3,393	3.95	5.02
Jupiter	1898.8	71,398	1.332	59.6
Saturn	568.4	60,330	0.689	35.5
Uranus	86.87	26,200	1.18	21.3
Neptune	102.8	25,225	(1.54)	23.3
Pluto	(0.013)	1,145	1.84	1.3

[a] Values in parentheses are uncertain.

TABLE II Planetary Orbits

Planet	Semimajor axis AU[b]	Semimajor axis 10^6 km	Orbital period (yr)	Eccentricity	Inclination[a] (deg)
Mercury	0.387	57.9	0.24085	0.206	7.003
Venus	0.723	108.2	0.61521	0.0068	3.394
Earth	1.000	149.6	1.00004	0.0167	0.000
Mars	1.524	227.9	1.88089	0.0933	1.850
Jupiter	5.203	778.3	11.86223	0.0483	1.309
Saturn	9.539	1427.0	29.45774	0.0559	2.493
Uranus	19.182	2869.6	84.018	0.047	0.772
Neptune	30.058	4496.6	164.78	0.0087	1.779
Pluto	39.44	5900.1	248.4	0.247	17.146

[a] Inclination is measured with respect to the orbital plane of the earth.
[b] An astronomical unit (AU) is the mean distance of the earth from the sun, or 149.60 10^6 km.

TABLE III Planetary Spins and Magnetic Fields

Planet	Rotation period	Obliquity of spin axis (deg)[b]	Magnetic moment (G cm^3)	Surface field[a,c] (G)
Mercury	58.65 days	2 ± 3	2.4×10^{22}	0.002
Venus	243.01 days	177.3	$<4 \times 10^{21}$	<0.00002
Earth	23.9345 hr	23.45	7.98×10^{25}	0.3
Mars	24.6299 hr	23.98	2.5×10^{22}	0.0006
Jupiter	9.841 hr (equator) 9.925 hr (interior)	3.12	1.5×10^{30}	4
Saturn	10.233 hr (equator) 10.675 hr (interior)	26.73	4.6×10^{28}	0.2
Uranus	17.24 hr (interior)	97.86	4.1×10^{27}	0.2
Neptune	18.2 ± 0.4 hr	(29.56)	—	—
Pluto	6.387 days	122.5	—	—

[a] Measured at the magnetic equator.
[b] Values in parentheses are uncertain.
[c] G = 1 Gauss.

Mars has a distinctly lower density, suggesting extensive oxidation of metal in Mars to much less dense iron oxides.

The giant planets, despite their compositional similarity to the Sun, are not to be thought of as "failed stars." The smallest possible hydrogen-burning Main Sequence star has a mass of 0.07 Suns, and the smallest hydrogen–helium body capable of any form of fusion reactions (deuterium burning) has a mass of 0.013 Suns. Jupiter's mass is only 0.0013 Suns. Thus Jupiter fails by a factor of 10 to meet even the most generous definition of a star. Jupiter, Saturn, and Neptune all have significant internal heat sources derived from the conversion of their gravitational potential energy into heat through slow shrinkage. In a sense, we are seeing the very tail end of their process of accretion and collapse. All three of these planets, but not Uranus, show internal heat sources that are 1.7 to 2.1 times as large as the flux of heat they receive from the Sun.

The four Jovian planets all have extensive, complex satellite systems and distinctive systems of rings. All the close satellites of all four giant planets orbit nearly in the plane of their planet's equator. However, the small outermost satellites of Jupiter, Saturn, and Neptune have highly inclined and even retrograde orbits, circling their planet in a direction opposite to the planet's rotation. These outer satellites may have been captured from the interplanetary swarm of small Centaurs, asteroids, and comets. The physical and orbital properties of the satellites of the solar system are described in Tables V–VII.

There are seven "lunar-sized bodies" in the solar system, intermediate in mass between the terrestrial planets and "small bodies," the smaller satellites and asteroids and comets. These seven, Earth's Moon, Jupiter's Galilean satellites Io, Europa, Ganymede and Callisto, Titan's largest satellite Titan, Neptune's largest satellite Triton, and Pluto together make up only a tenth of the mass of Earth. Nonetheless, these bodies display a phenomenal variety of compositions, physical properties, and geological features that render them fascinating objects for study. The smallest bodies make up for their lack of geological evolution by preserving records of conditions at the earliest

TABLE IV Planetary Atmospheres

Planet	Surface pressure (bars)	Average surface temperature (K)	Major gases	Fractional abundance by number[a]
Mercury	$\sim 10^{-14}$	440	Na	0.97
			He	0.03
Venus	90	730	CO_2	0.96
			N_2	0.035
Earth	1	288	N_2	0.77
			O_2	0.21
			H_2O	0.01
Mars	0.007	218	CO_2	0.95
			N_2	0.027
			Ar	0.016
Jupiter	—	—	H_2	0.90
			He	0.10
Saturn	—	—	H_2	0.94
			He	0.06
Uranus	—	—	H_2	0.85
			He	0.15
Neptune	—	—	H_2	0.85?
			He	0.15?
Pluto	0.001	—	CH_4	—
			Ne?	—

[a] This quantity is also referred to as the volume mixing ratio.

TABLE V Satellite Atmospheres

Satellite	Surface pressure (bars)	Average surface temperature (K)	Major gases	Fractional abundance by number[a]
Moon	$\sim 2 \times 10^{-14}$	274	Ne	0.4
			Ar	0.4
			He	0.2
Io	$\sim 1 \times 10^{-10}$	~110	SO_2	1
Titan	1.6	95	N_2	0.73–0.99[b]
			Ar	0.00–0.28
			CH_4	0.01–0.12
Triton	~0.1	~57	N_2	—
			CH_4	—

[a] This quantity is also referred to as the volume mixing ratio.
[b] This is the range of uncertainty.

TABLE VI The Satellites

Planet		Satellite	Mass (10^{20} kg)	Radius (km)	Density (10^3 kg m^{-3})	Surface composition
Earth		Moon	734.9	1738	3.34	Rocks
Mars	MI	Phobos	1.26×10^{-4}	11[a]	2.2	Carbonaceous
	MII	Deimos	1.8×10^{-5}	6.3[a]	1.7	Carbonaceous
Asteroids[b]						
Jupiter	JXVI	Metis	—	20	—	Rock?
	JXV	Adrastea	—	10[a]	—	Rock?
	JV	Amalthea	—	97[a]	—	Rock with sulfur
	JIV	Thebe	—	50	—	Rock?
	JI	Io	894	1815	3.57	Rock with sulfur
	JII	Europa	480	1569	2.97	Ice over rock
	JIII	Ganymede	1482.3	2631	1.94	Water ice
	JIV	Callisto	1076.6	2400	1.86	Dirty water ice
	2000 J1	—	—	8	—	Carbonaceous?
	JXIII	Leda	—	8	—	Carbonaceous?
	JVI	Himalia	—	90	—	Carbonaceous?
	JX	Lysithea	—	20	—	Carbonaceous?
	JVII	Elara	—	40	—	Carbonaceous?
	2000 J11		—	2	—	Carbonaceous?
	2000 J10		—	2	—	Carbonaceous?
	2000 J3		—	3	—	Carbonaceous?
	2000 J7		—	4	—	Carbonaceous?
	JXII	Ananke	—	15	—	Carbonaceous?
	2000 J5		—	2	—	Carbonaceous?
	2000 J9		—	3	—	Carbonaceous?
	JXI	Carme	—	22	—	Carbonaceous?
	2000 J4		—	2	—	Carbonaceous?
	2000 J6		—	2	—	Carbonaceous?
	JVIII	Pasiphae	—	35	—	Carbonaceous?
	2000 J8		—	3	—	Carbonaceous?
	JIX	Sinope	—	20	—	Carbonaceous?
	2000 J2		—	4	—	Carbonaceous?
	1999 J1		—	5	—	Carbonaceous?
Saturn	SXVIII	Pan	—	15?	—	Water ice?
	SXV	Atlas	—	16[a]	—	Water ice?
	SX	Janus	—	93[a]	—	Water ice?
	SI	Mimas	0.38	201	1.137	Water ice?
	SII	Enceladus	0.8	251	1.2	Water ice?
	SIII	Tethys	7.6	524	1.26	Water ice?
	SXIII	Telesto	—	11[a]	—	Water ice?
	SXIV	Calypso	—	12[a]	—	Water ice?
	SIV	Dione	10.5	559	1.44	Water ice?
	SXII	Helene	—	16[a]	—	Water ice?
	SV	Rhea	24.9	764	1.33	Water ice?
	SVI	Titan	1345.7	2575	1.882	Ices (atmosphere)
	SVII	Hyperion	—	132[a]	—	Dirty water ice
	SVIII	Iapetus	18.8	718	1.21	Ice/carbonaceous?
	2000 S5		—	9	—	Ice/carbonaceous?
	2000 S6		—	7	—	Ice/carbonaceous?

continues

TABLE VI *(Continued)*

Planet		Satellite	Mass (10^{20} kg)	Radius (km)	Density (10^3 kg m^{-3})	Surface composition
	SIX	Phoebe	—	110[a]	—	Carbonaceous?
	2000 S2		—	12	—	Ice/carbonaceous?
	2000 S8		—	4	—	Ice/carbonaceous?
	2000 S3		—	22	—	Ice/carbonaceous?
	2000 S10		—	5	—	Ice/carbonaceous?
	2000 S11		—	15	—	Ice/carbonaceous?
	2000 S4		—	10	—	Ice/carbonaceous?
	2000 S9		—	4	—	Ice/carbonaceous?
	2000 S12		—	4	—	Ice/carbonaceous?
	2000 S7		—	4	—	Ice/carbonaceous?
	2000 S1		—	10	—	Ice/carbonaceous?
Uranus	UVI	Cordelia	—	20	—	Water ice?
	UVII	Ophelia	—	25	—	Water ice?
	UVIII	Bianca	—	25	—	Water ice?
	UIX	Cressida	—	30	—	Water ice?
	UX	Desdemona		30		Water ice?
	UXI	Juliet	—	40	—	Water ice?
	UXII	Portia	—	40	—	Water ice?
	UXIII	Rosalind	—	30	—	Water ice?
	UXIV	Belinda	—	30	—	Water ice?
	1986 U10		—	20	—	Water ice?
	UXV	Puck	—	85	—	Water ice?
	UV	Miranda	0.7	242	1.3	Dirty water ice
	UI	Ariel	13	580	1.6	Dirty water ice
	UII	Umbriel	13	595	1.4	Dirty water ice
	UIII	Titania	35	805	1.6	Dirty water ice
	UIV	Oberon	29	775	1.5	Dirty water ice
	UXVI	Caliban	—	10	—	Dirty ices?
	UXX	Stephano	—	10	—	Dirty ices?
	UXVII	Sycorax	—	10	—	Dirty ices?
	UXVIII	Prospero	—	10	—	Dirty ices?
	UXIX	Setebos	—	10	—	Dirty ices?
Neptune	NI	Triton	1300	1750	4?	Ices
	NII	Nereid	—	300	—	Ices?
Pluto	PI	Charon	—	640	1.84	Methane ice

[a] Mean radii are given for satellites that are markedly nonspherical. Most other radii less than 100 km are inferred from the brightness of an unresolved object.

[b] Several asteroids have or appear to have satellites. The best documented include the asteroids (243) Ida (satellite named Dactyl), (45) Eugenia, (87) Sylvia, (90) Antiope, and (762) Pulkova. All are probably rocky bodies.

stages of the formation and evolution of the solar system. Indeed, meteorites are a priceless source of data derived from a wide variety of very diverse small solar system bodies. Although most meteorites come from roughly 50 different parent asteroids, nearly 20 meteorites in our collections have been positively identified as fragments knocked off the Moon and Mars by comet and asteroid impacts. It is a tantalizing possibility that samples of Mercury or Venus may lurk unrecognized in our meteorite collections.

Planetary rings, composed as they are of very small particles, are likely to be the most evanescent, albeit beautiful and spectacular, features of the solar system. The ages of the ring systems of the giant planets are not known. Irrespective of their ages, however, their dynamic behavior is complex and intriguing. Processes acting in present-day swarms of small bodies in ring and belt systems may have played a critical role in the accretion of preplanetary solids from the original dust disk. Data on comets and asteroids

TABLE VII Satellite Orbits

Planet		Satellite	Semimajor axis (km)	Semimajor axis (Rpl)	Period[a] (days)	Eccentricity[b]	Inclination[c]
Earth		Moon	384.4	60.3	27.322	0.0549	5.15
Mars	MI	Phobos	9.378	2.76	0.319	0.015	1.02
	MII	Deimos	23.459	6.91	1.263	0.00052	1.82
Jupiter	JXVI	Metis	127.96	1.7922	0.2948	<0.004	0?
	JXV	Adrastea	128.98	1.8065	0.2983	0?	0?
	JV	Amalthea	181.3	2.539	0.4981	0.003	0.40
	JIV	Thebe	221.9	3.108	0.6745	0.015	0.8
	JI	Io	421.6	5.905	1.769	<.0041	0.04
	JII	Europa	670.9	9.397	3.551	<0.01	0.470
	JIII	Ganymede	1,070	14.99	7.155	<.0015	0.281
	JIV	Callisto	1,883	26.37	16.689	0.007	0.281
	2000 J1		7,507	105.1	129.71	0.204	46[d]
	JXIII	Leda	11,094	155.4	238.72	0.148	26[d]
	JVI	Himalia	11,480	160.8	250.57	0.158	28[d]
	JX	Lysithea	11,720	164.2	259.22	0.107	29[d]
	JVII	Elara	11,737	164.4	259.65	0.207	28[d]
	2000 J11		12,557	175.9	289	0.25	28.2[d]
	2000 J10		20,174	282.5	588(R)	0.14	166[d]
	2000 J3		20,210	283.1	584(R)	0.22	150[d]
	2000 J7		21,010	294.3	621(R)	0.22	149[d]
	JXII	Ananke	21,200	296.9	631(R)	0.169	147[d]
	2000 J5		21,336	298.8	632(R)	0.24	149[d]
	2000 J9		22,304	312.4	683(R)	0.26	165[d]
	JXI	Carme	22,600	316.5	692(R)	0.207	163[d]
	2000 J4		22,972	321.7	712(R)	0.28	165[d]
	2000 J6		23,074	323.2	720(R)	0.26	165[d]
	JVIII	Pasiphae	23,500	329.1	735(R)	0.378	148[d]
	2000 J8		23,618	330.8	741(R)	0.41	153[d]
	JIX	Sinope	23,700	331.9	758(R)	0.275	153[d]
	2000 J2		23,746	332.6	752(R)	0.24	165[d]
	1999 J1		24,235	339.4	768(R)	0.125	143[d]
Saturn	SXVIII	Pan	135.6	2.220	0.576	0?	0?
	SXV	Atlas	137.64	2.281	0.602	0?	0?
	SXVI	Prometheus	139.35	2.310	0.613	0.0024	0?
	SXVII	Pandora	141.70	2.349	0.629	0.00422	0?
	SXI	Epimetheus	151.472	2.510	0.694	0.009	0.34
	SX	Janus	151.472	2.511	0.695	0.007	0.14
	SI	Mimas	185.52	3.075	0.942	0.0202	1.53
	SII	Enceladus	238.02	3.945	1.370	(0.0045)	0.02
	SIII	Tethys	294.66	4.884	1.888	0.0000	1.09
	SXIII	Telesto	294.66	4.884	1.888	0?	0?
	SXIV	Calypso	294.66	4.884	1.888	0?	0?
	SIV	Dione	377.40	6.256	2.737	(0.0022)	0.02
	SXII	Helene	377.40	6.256	2.737	0.005	0.2
	SV	Rhea	527.04	8.736	4.518	<0.001	0.35
	SVI	Titan	1,221.85	20.25	15.945	0.0292	0.33
	SVII	Hyperion	1,481.1	24.55	21.277	(0.1042)	0.43
	SVIII	Iapetus	3,561.3	59.03	79.331	0.0283	7.52

continues

TABLE VII *(Continued)*

Planet		Satellite	Semimajor axis (km)	Semimajor axis (Rpl)	Period[a] (days)	Eccentricity[b]	Inclination[c]
	2000 S5		11,339	185.6	449	0.33	46.2[d]
	2000 S6		11,465	187.7	453	0.32	46.6[d]
	SIX	Phoebe	12,944	214.5	55.5(R)	0.163	174.8[d]
	2000 S2		15,172	248.4	687	0.36	45.2[d]
	2000 S8		15,676	256.6	730(R)	0.27	153.0[d]
	2000 S3		17,251	282.4	825	0.27	45.5[d]
	2000 S10		17,452	285.7	858	4.47	34.7[d]
	2000 S11		17,874	292.6	888	0.38	33.1[d]
	2000 S4		18,231	298.5	924	0.54	33.5[d]
	2000 S9		18,486	302.6	939(R)	0.22	167.4[d]
	2000 S12		19,747	323.3	1037(R)	0.12	175.0[d]
	2000 S7		20,144	329.8	1067(R)	0.45	175.9[d]
	2000 S1		23,076	377.8	1311(R)	0.34	173[d]
Uranus	UVI	Cordelia	49.7	1.90	0.33	0?	0?
	UVII	Ophelia	53.8	2.05	0.38	0?	0?
	UVIII	Bianca	59.2	2.26	0.43	0?	0?
	UIX	Cressida	61.8	2.36	0.46	0?	0?
	UX	Desdemona	62.7	2.39	0.48	0?	0?
	UXI	Juliet	64.6	2.47	0.49	0?	0?
	UXII	Portia	66.1	2.52	0.51	0?	0?
	UXIII	Rosalind	69.9	2.67	0.56	0?	0?
	UXIV	Belinda	75.3	2.87	0.62	0?	0?
	1986 U10		76.42	2.91	0.43	0?	0?
	UXV	Puck	86.0	3.28	0.76	0?	0?
	UV	Miranda	129.783	4.95	1.413	0.0027?	4.22
	UI	Ariel	191.239	7.30	2.520	0.0034?	0.31
	UII	Umbriel	265.969	10.15	4.144	0.0050?	0.36
	UIII	Titania	435.844	16.64	8.706	0.0022?	0.14
	UIV	Oberon	582.596	22.24	13.463	0.0008?	0.10
	UXVI	Caliban	7,187	274.4	579(R)	0.082	139.7[d]
	UXX	Stephano	7,960	303.9	676(R)	0.146	141.5[d]
	UXVII	Sycorax	12,240	429.1	1289(R)	0.509	152.7[d]
	UXVIII	Prospero	16,150	616.5	1953(R)	0.327	146.3[d]
	UXIX	Setebos	18,250	696.7	2345(R)	0.494	148.8[d]
Neptune	NI	Triton	354.3	14.0	5.877(R)	<0.0005	159.0
	NII	Nereid	551.5	219	360.16	0.75	27.6[d]
Pluto	PI	Charon	19.1	16.7	6.387	0?	94.3

[a] R denotes a retrograde orbit.
[b] Eccentricities are forced eccentricities due to interactions with other satellites.
[c] Relative to the orbital plane of the planet.
[d] Orbits of distant satellites are strongly perturbed by the Sun.

are given in Table VIII and IX, respectively. Planetary rings are summarized in Table X.

II. DYNAMICS AND STRUCTURE

The orbits of most bodies in the solar system can be treated, to a very satisfactory level of approximation, as two-body problems, with the smaller body moving in the gravitational field of the more massive body, treating both as point masses. Such motions are very well described by Kepler's laws of motion. Kepler's first law (1609) postulates that planetary orbits are ellipses with the Sun at one focus (the "prime focus"). The second law (1609) states that the radius line from the Sun to the planet sweeps out area in the orbital plane at a constant rate. The third (1619) states that,

TABLE VIII Selected Asteroids

Asteroid	Semimajor axis (AU)	Eccentricity	Inclination (deg)	Diameter (km)	Type[a]	Rotational period (hr)
1 Ceres	2.768	0.077	10.598	1025	C	9.078
2 Pallas	2.773	0.233	34.800	565	C	7.811
3 Juno	2.671	0.255	13.002	244	S	7.211
4 Vesta	2.362	0.090	7.144	533	U	5.342
5 Astraea	2.577	0.189	5.349	122	S	16.812
16 Psyche	2.921	0.138	3.092	249	M	4.303
221 Eos[b]	3.012	0.071	10.02	98	S	10.450
320 Katherina[b]	3.013	0.074	10.08	—	—	—
590 Tomyris[b]	3.001	0.078	10.02	—	—	—
1580 Betulia	2.196	0.490	52.041	6	U	6.130
1862 Apollo	1.470	0.560	6.360	1	Q	3.065

[a] The major asteroidal classes are C (carbonaceous?), S (stony or silicate?), and M (metal?). A large number of asteroids are designated U (unclassified); these asteroids do not fall into any of the recognized classes.

[b] The asteroids Eos, Katherina, and Tomyris are members of a Hirayama family. For these asteroids we list their proper orbital elements. In all other cases we list the osculating orbital elements.

for all bodies that orbit the same primary body, the square of the orbital period (P) of each body is proportional to the cube of its mean distance (a) from the primary. Thus, for bodies orbiting the Sun, expressing the orbital period in units of Earth years and the distances in Astronomical Units, $P^2 = a^3$. Sir Isaac Newton (1665) demonstrated that this law is a consequence of universal gravitation, obeying the law that there exists an attractive force between any pair of bodies in the Universe, proportional to the product of their masses, and inversely proportional to the square of the distance between their centers. The force between two bodies of mass M and m whose centers are a distance R apart is therefore

$$F = GMm/R^2.$$

Here G is the universal gravitational constant, 6.673×10^{-11} m^3/kg s^2. Kepler's laws tell us nothing about the spacings of the orbits of planets or satellites. Johann Titius (1766) and Johann Bode (1772) proposed a rule to describe the spacings of the orbits of the planets, often overgenerously referred to as the "Titius-Bode law," but the rule has no physical basis, requires an irrational sequence of constants, predicts a planet in the asteroid belt, and falls down badly for the two most distant planets. This rule is seldom taken seriously by astronomers.

Nonetheless, the structure of the solar system is clearly not altogether random. There is only one case of two planetary orbits crossing (Neptune and Pluto), and resonant relationships are far more common than chance alone would suggest.

In celestial mechanics, the orbit of a body is described by the elements a, e, I, Ω, and $\tilde{\omega}$, where a is the semimajor axis, e the eccentricity, I the inclination, Ω the longitude

TABLE IX Selected Comets[a]

Comet	Name	Semimajor axis (AU)	Orbital period (yr)	Eccentricity	Inclination (deg)
1977	IX West	30,000 (exit)	5×10^6 (no return)	0.999815	116.9
1937	IV Whipple	16,000 (714)	2×10^6 (2×10^4)	0.999892	41.6
1910	I Great Comet	7,400 (1500)	6×10^5 (6×10^4)	0.999983	138.8
	P Giacobini–Zinner	3.51	6.59	0.7076	31.88
	P Halley	17.9	76.0(R)[b]	0.9673	162.23
	P Enke	2.21	3.29	0.8499	11.93
	P Tempel 2	3.04	5.29	0.5444	12.43

[a] The values quoted are estimates of the orbital parameters the comet had before it reached the inner solar system; the values in parentheses are the orbital parameters the comet will have when it is no longer perturbed by the planets. These comments do not apply to periodic comets.

[b] R denotes a retrograde orbit.

TABLE X Planetary Rings

Planet	Ring	Boundary radii (planetary radii)	Eccentricity	Inclination (deg)	Ringlet width[a] (km)
Jupiter	Halo	1.41–1.71	∼0	<10	—
	Main ring	1.71–1.81	∼0	0	—
	Gossamer ring	1.81–1.83	∼0	0	—
Saturn	D Ring	1.11–1.235	0	0	—
	C Ring	1.235–1.525	0	0	—
	B Ring	1.525–1.940	0	0	—
	A Ring	2.025–2.267	0	0	—
	F Ring	2.324	0.00026		16
	G Ring	2.82	0	0	—
	E Ring	3–8	0	0	—
Uranus	1986U2R	1.41–1.51	—	—	2500
	6	1.5973	0.0010	0.063	1–3
	5	1.6122	0.0019	0.052	2–3
	4	1.6252	0.0011	0.032	2–3
	α	1.7073	0.0008	0.014	7–12
	β	1.7431	0.0004	0.005	7–12
	η	1.8008	0	0.002	0–2
	γ	1.8179	0	0.011	1–4
	δ	1.8439	0	0.004	3–9
	1986U1R	1.9099	—	—	1–2
	ε	1.9527	0.0079	0	22–93
Neptune	Arc[b]?	2.145	—	—	23
	Arc?	2.51	—	—	8
	Arc[b]	2.66	—	—	15

[a] The width of a narrow ring varies from a minimum at pericenter to a maximum at apocenter.
[b] These partial rings have been confirmed by independent observers.

of the ascending node, and $\tilde{\omega}$ the longitude of the pericenter. The position in the orbit is given by the mean longitude,

$$\lambda = \int_0^t n\, dt + \varepsilon \simeq nt + \varepsilon,$$

where n, the mean motion of the body, is equal to $2\pi/P$, where P is the body's orbital period; ε is the longitude at epoch. A pair of planets or satellites are in resonance if the geometric configuration of their positions with respect to their orbits is repeated within a short period of time. This requires that the ratio of their orbital periods (or, equivalently, the ratio of their mean motions) be close to two small integers. For example, the ratio of the present orbital periods of Neptune and Pluto is $0.6634 \simeq 2{:}3$. The exact conditions that describe the resonant configuration of these planets is

$$\phi = 2\lambda - 3\lambda' + \tilde{\omega}' = \pi,$$

where the primed quantities refer to the orbital elements of Pluto. In fact, ϕ is not exactly π but librates (or oscillates) about π. Differentiating the previous equation with respect to time and rearranging, we obtain

$$(n' - \dot{\tilde{\omega}}')/(n - \dot{\tilde{\omega}}') = \frac{2}{3}.$$

Thus, the mean motions of the planets relative to the motion of Pluto's pericenter are exactly commensurate. The orbital paths of Neptune and Pluto intersect, leading one to expect that eventually the planets will collide. However, conjunctions of planets occur when $\lambda = \lambda'$, and the previously cited resonance conditions ensure that all conjunctions of Neptune and Pluto occur near Pluto's apocenter, which in turn ensures that these planets never have a close encounter.

Numerous other resonances of this type exist in the solar system, particularly among the satellites of Jupiter and Saturn, and it has been proved that the number of pairs of nearly commensurate mean motions in the solar system is too great to be ascribed to chance. We can conclude from this either that the mechanism of formation of the planets and satellites was such as to favor orbits

with commensurate mean motions or that the present distribution of orbits is the result of orbital evolution since the time of formation. It is likely that the resonances in the satellite systems of Jupiter and Saturn are the result of orbital evolution due to tidal friction.

In marked contrast to the planet and satellite systems, asteroids tend to avoid orbits that would be nearly commensurate with Jupiter. These gaps in the distribution of asteroidal mean motions (or semimajor axes) are known as the Kirkwood gaps after their discoverer, Daniel Kirkwood. Similar gaps occur in the ring system of Saturn. In particular, particles at the inner edge of Cassini's division (the prominent gap named after J. D. Cassini that separates the outer A ring from the inner B ring; see Fig. 1) are in a 2:1 resonance with the satellite Mimas. Resonances also appear to bound the edges of the Uranian ε ring. Bodies that orbit at the same mean distance from a primary, that is, that have a common orbit, are in 1:1 resonance. Examples include the Trojan asteroids, which coorbit with Jupiter and oscillate about the planet's Lagrangian equilibrium points; some move, on average, 60° ahead of the planet, whereas others trail the planet, on average, by 60°.

Resonances in the Solar System can also involve the spin and the orbit of a body. The rotational period of Mercury, for example, is exactly two-thirds of its orbital period. Spin–orbit resonances are not primordial but the result of spin–orbit evolution due to tidal friction. Tidal oscillations always result in the dissipation of mechanical energy. Since angular momentum is conserved, the decrease in the total energy of the system due to tidal interactions results in an exchange of angular momentum between the spin of a planet and the orbit of the tide-raising satellite. If the orbital period of the satellite is greater than the planet's rotational period, the planet is braked while the orbit of the satellite expands. Observations indicate that the Moon may have once orbited within 10 Earth radii from Earth, at which time Earth's day may have been less than 15 hr. When resonant states are encountered, resonance capture can occur, with the result that the resonant configuration is maintained despite the continued action of the tidal forces.

FIGURE 1 Saturn taken on August 4, 1981, by *Voyager 2*. Three of Saturn's icy satellites are evident at left. They are, in order of distance from the planet, Tethys, Dione, and Rhea. The shadow of Tethys appears on Saturn's southern hemisphere. A fourth satellite, Mimas, is less evident, appearing as a small bright spot between Tethys and the rings. The broad, inner B ring is separated from the outer A ring by Cassini's division. The dark region near the outer edge of the A ring is Encke's gap. Spokes, transient markings produced by electrical phenomena, are evident on the B ring. Note that the planet is markedly oblate with an equatorial radius 8.8% larger than the polar radius. (Courtesy of JPL/NASA.)

III. THE TERRESTRIAL PLANETS

The four terrestrial planets, Mercury, Venus, Earth and Mars, as well as two other planetary bodies in satellite orbits around planets (Earth's Moon and Jupiter's inner Galilean satellite, Io) share certain common features that probably characterize all terrestrial planets. They are composed almost entirely of ferrous native metals, sulfides, and silicates; they are large enough to have melted and differentiated into layers of different density and composition; and they have either sparse, anomalous satellite system, or no satellites at all. The two satellites of Mars, Phobos and Deimos, are small, dark, irregular bodies of asteroidal appearance. It is plausible that they were captured very early in the history of the solar system, when Mars still had an equatorial accretion disk or extended atmosphere. Earth's Moon is unusual in that it is very large relative to the size of its primary body (1/81.3 of Earth's mass). Neither Mercury nor Venus has a satellite, probably because their proximity to the Sun would have allowed any satellite to escape under the influence of solar tidal forces. The terrestrial planets also differ in their internal densities and structures, magnetic fields, surface geology, and atmospheric mass and composition.

The interiors of the terrestrial planets differ markedly, in that Mercury's metallic core has a mass roughly equal to 60% of the mass of the entire planet, compared to 30–31% for Earth and Venus. Such a high core fraction requires the loss, or failure to accrete, of some 70% of the silicate portion of Mercury's mass. The low density and high

rotational moment of inertial of Mars require a smaller, less dense core than Earth or Venus, and a more massive, denser mantle. These differences in turn suggest a large difference in oxidation state between the terrestrial planets, with the Martian core depleted in metallic iron and the mantle enriched in iron oxides. Such a trend toward higher oxidation states at greater distances from the Sun, which mirrors the behavior of asteroids, is explained in theory by progressively lower temperatures farther from the Sun. The temperature gradient required to produce such a large difference in oxidation state also predicts a higher initial content of chemically reactive volatiles in Mars than in Earth or Venus.

Earth's strong magnetic field is generated by convective overturn of the liquid outer core, driven by some combination of heat sources such as slow crystallization of the inner core and radioactive decay of trace constituents of the outer core. Venus, often loosely termed Earth's sister planet, has no detectable planetary dipole field. The absence of a field is probably as a consequence of the slow rotation of the planet, insufficient to organize convective core motions into large cells that can serve as a global dynamo. The Venus day is 243 Earth days long. The Moon also has no present planetary dipole field, but it has been known since the Apollo missions that remnant magnetization is widespread in the ancient lunar crust, implying the early existence of strong magnetic fields about 3.5–3.6 billion years ago, at the time when these rocks cooled through the Curie temperature, at which ferromagnetic materials retain permanent imprints of magnetization from ambient fields. These imprinting fields have been attributed both to external (impact and solar wind) and internal (core dynamo) sources. At present there is no proof that the Moon even has a metallic core: density, solar–wind interaction, and seismic data combine to suggest that the upper limit on the size of the putative lunar core is about 3% of the mass of the Moon.

The surfaces and tectonic styles of the terrestrial planets also diverge markedly, reflecting both differing conditions of origin and different evolutionary histories. Mercury and the Moon are heavily cratered, lightly marked by internal processes, and devoid of weathering by gases or liquids. Mars is divided roughly in half by a plane inclined about 30° to the Martian equator, south of which are lightly weathered cratered highlands reminiscent of the Moon, and north of which the surface has only a sparse sprinkling of relatively small craters amidst landforms dominated by extremely concentrated volcanism and the pervasive effects of both wind and water erosion (Fig. 2).

The surface of Jupiter's satellite Io is remarkable for its high degree of volcanic activity. Eruptions of sulfur dioxide, sulfur, and traces of several other sulfur-bearing gases send fountains of gas and dust as high as 250 km above the surface. The surface is coated with volcanic material at so great a rate that the entire mass of Io could be cycled through these volcanoes over the age of the solar system.

FIGURE 2 A network of valleys on the surface of Mars indicates the past existence of liquid water at the surface of the planet. Climatic conditions are now such that water cannot exist in the liquid state. However, the high density of craters on the valleys indicates that the channels are very old and that the warmer, wetter climate that produced the channels may have existed 3 billion (3×10^9) years ago. The scene shown is in Thaumasia Fossae (40°S, 90°W) and is 250 km across. (Courtesy of NASA.)

The red-hot surface of Venus is largely covered by rolling lowlands. Two continent-sized elevated areas, Maxwell Montes and Aphrodite, and several apparent volcanic piles stand well above the lowlands. The surface of Venus is lightly cratered, reflecting two important constraints on the cratering record. First, no part of Venus appears to be older than a few hundred million years, suggesting a major resurfacing event, or the end of an era of high surface renewal, about 0.5 Ga ago. Second, small craters are totally absent, reflecting the disruption of all but the very largest impactors during passage through Venus' dense atmosphere. That atmosphere is roughly 100 times as massive as Earth's. It is, like Mars' atmosphere, dominated by carbon dioxide, with a few percent of argon. Water is an extremely minor atmospheric constituent, about 50 parts per million. Clouds of sulfuric acid droplets and other poorly characterized materials shroud the surface from external view at visible wavelengths.

Earth's surface is most noteworthy for the extreme importance of continental drift and the broad prevalence of ocean basins. Some 72% of Earth's surface area is covered by oceans. The rocks underlying the oceans are generally dense, basaltic rocks coated by a thick layer of marine sediments. The ocean floor is almost without exception

younger than 200 million years, compared to the continental blocks, which have abundant rocks older than 1 Ga, and range in age up to nearly 4 Ga. The oceans, making up about 0.03% of the mass of the planet, contain many soluble materials, notably alkali metals, alkaline earths, halides, sulfates, and carbonates, leached from the continents by weathering. The atmosphere is extensively influenced by the presence of life. Oxygen, the second most abundant gas after nitrogen, is a byproduct of photosynthesis by plants. Human activities have recently achieved such a scale that they are now a factor in the global balances of many atmospheric gases.

IV. THE MOON

Before the landings of the first analytical instruments on the Moon during the American unmanned *Surveyor* program of the late 1960s it was widely expected that the Moon might be a primitive body, similar in composition to chondritic meteorites, that had never experienced melting and density-dependent differentiation. The *Surveyor* data showed a composition more similar to basaltic achondrite meteorites, a conclusion that was later confirmed and extended by the samples returned from the Moon between 1969 and 1976 by the American manned *Apollo* missions and the Soviet unmanned *Luna* 16, 20, and 24 missions. The returned rock and regolith samples show that the surface material of the Moon has been completely processed through internal igneous activity, which ceased about 3 billion years ago.

Theories of the origin of the Moon before the *Apollo* missions generally fell into one of three broad categories: the Moon was captured from an independent orbit around the Sun, or it accreted at the same time as Earth out of planetary raw material in a ring-like debris belt girdling Earth, or it was split off from Earth by rotational fission of a rapidly rotating proto-Earth. In light of the vast volume of new data, post-*Apollo* theories for the origin of the Moon have been extensively revised. The discovery that the bulk silicate composition of the Moon and the isotopic composition of the oxygen in these silicates are both closely similar to Earth's mantle has encouraged speculation that the Moon is derived from largely terrestrial material that was ejected into an orbiting debris belt by the impact of a very large, approximately Mars-sized, impactor. Dating of lunar rocks shows that the Moon was formed at least 4.4 billion years (Ga) ago, not long after the formation of the Sun and Earth at 4.55 Ga. The oldest known rocks on Earth are close to 4.0 Ga in age, and the oldest known detrital mineral grain, a zircon crystal from roughly 4-Ga-old sediments, has been dated at 4.3 Ga. The age distribution of lunar rocks reveals that there was a peak in igneous activity, correlated with a lengthy episode of intense bombardment by relatively large impactors, centered around 3.5 Ga ago. That episode faded out into essentially complete inactivity within 0.5 Ga of the peak of the bombardment, leaving numerous huge impact basins flooded by lava of a variety of different compositions. Although bombardment by smaller bodies continues to the present, these later events have failed to breach the lunar curst or generate fresh flooding by lava. The long impact bombardment of the Moon has left an intensely cratered surface covered with shattered rocks and dust ejected by these impacts. The great basins produced during the 3.5-Ga intense bombardment era of course are much younger and much more lightly cratered than the more ancient lunar highlands that survived the bombardment. This layer of pulverized and mixed impact debris, termed the "regolith," contains a fraction of 1% of meteorite material contributed by the impacting bodies themselves. This regolith, with a characteristic thickness of meters to hundreds of meters, blankets the entire lunar surface. Seismometers placed on the lunar surface during the *Apollo* landing missions have revealed an extremely quiescent lunar interior, in which seismic events of a given magnitude are roughly 100 million times less frequent than their counterparts on Earth. Further, the few events seen are concentrated at great depth, near 800 km, attesting that the outermost 800 km of the Moon is far colder and stronger than Earth's. Indeed, the principal role of these seismometers has been to monitor the explosion of the continuing rain of asteroidal and cometary impactors. The seismic waves from these impacts reverberate freely about the lunar interior, pointing to the absence of magma and soft rock, which would absorb and attenuate the seismic echoes. These same seismic data show a lunar crust that is 50–60 km thick on the basin-rich side toward Earth and about 100 km thick on the lunar far side, which is nearly all highlands. They also fail to reveal the presence of a metallic core in the center of the Moon. Magnetic and seismic evidence combine to limit the maximum size of any possible lunar metal core to a few percent of the mass of the Moon, far less than found in any of the terrestrial planets.

Lunar rocks, although similar to terrestrial rocks in a number of general ways, depart strikingly from Earth in several important particulars. Volatile materials, not only water and compounds of carbon and nitrogen but also many of the more volatile rock-forming elements such as sodium and potassium are strongly depleted in the Moon relative to terrestrial and meteoritic material. The Moon has both an elevated amount of FeO and a severely deficient amount of metallic iron relative to Earth. The other chalcophile (sulfur-loving) and siderophile (metal-loving) elements that are removed by differentiation and core

formation, such as nickel, cobalt, iridium, sulfur, phosphorus, selenium, lead, and the like, have abundance patterns in lunar igneous rocks similar to those on Earth. The abundances of core-forming trace metals in lunar regolith, in contrast, appear to reflect faithfully the admixture of a small amount of chondritic material from more recent small impacts. The Moon's very low bulk density, only 3.344 ($\pm$0.002) g/cm^3 confirms the rarity of free metal. These lines of evidence are often cited to suggest that the Moon is formed largely of shock-heated and devolatilized material derived from the mantles of Earth and the impacting planetary body, which, as attested by the Moon's elevated FeO content, may well have originated farther from the Sun than Earth. Computer models of the giant impact agree with the chemical and density data on the Moon in concluding that the material splashed off Earth in an oblique impact must have included almost no core material from the impacting body.

Curiously, both the Moon and Mercury have limited deposits of ice in the floors of deep, permanently shadowed craters near their poles. This ice is almost certainly impure water ice derived from impacting asteroidal and cometary bodies.

V. THE GIANT PLANETS

As was first pointed out by Harrison Brown a half-century ago, there are three general compositional classes of material in the solar system. These are permanent gases such as hydrogen, helium, and neon; ice-forming volatiles, such as water, ammonia, and methane; and rock-forming elements, most notably oxygen, silicon, magnesium, and iron. Taken together, these three classes add up to the same composition as the Sun, which consists of about 1% by mass of rock-forming elements, 1% by mass of ice-forming elements, and 98% by mass of hydrogen and helium. The terrestrial planets are composed, to excellent approximation, of rock-forming elements, whereas comets and icy satellites are mixtures of the rock-forming and ice-forming elements, often with roughly the same proportions that their component elements have in the Sun. The giant planets, Jupiter, Saturn, Uranus, and Neptune, contain very large amounts of the permanent gases in addition to the other two classes of materials. Of the giant planets, Jupiter and Saturn are closest to the Sun in overall composition. Based on bulk density and atmospheric composition data, it appears that Jupiter is enriched in the ice- and rock-forming elements by about a factor of five compared to the Sun, giving Jupiter, with a total mass of 318 Earth masses, a rocky core of about 15 Earth masses. Saturn is similarly enriched by about a factor of 10, giving it an endowment of about 9 Earth masses of rock out of 95.2 Earth masses total. Uranus and Neptune, in contrast, contain roughly half permanent gases and half condensable heavy elements. For many purposes it is desirable to distinguish between the two Jovian planets and the two Uranian planets. All four giant planets have low cloudtop temperatures because of their great distances from the Sun. However, Neptune and both Jovian planets have substantial internal heat sources that release more heat than these planets receive from the Sun. All four planets are fluid and convectively active, with deep interior temperatures that range from about 7000 K, the surface temperature of the Sun, in the Uranian planets to about 10,000 K in Saturn and 20,000 K in Jupiter.

Internal structure models constrained by thermal emission observations, bulk planetary density, planetary figure and rotational moment of inertial, atmospheric composition data, and theoretical equations of state for the major planetary materials have been calculated for all four giant planets. Jupiter (Fig. 3) has a deep, massive, and well-mixed atmosphere of hydrogen and helium and minor amounts of many other gases such as neon, water vapor, ammonia, methane, argon, phosphine, and hydrogen sulfide. About a quarter of the way from Jupiter's cloudtops to its center, pressures of about 3.5 Mbar are reached, sufficient to crush the electron shells of hydrogen molecules

FIGURE 3 Jupiter, its Great Red Spot, and two of its four largest satellites are visible in this photograph taken February 5, 1979, by *Voyager 1*. The innermost large satellite, Io, can be seen against the right-hand side of Jupiter's disk. The satellite Europa is seen to the right of the planet. Jupiter's colorfully banded atmosphere displays complex flow patterns highlighted by the Great Red Spot, a large, circulating atmospheric disturbance seen here in the bottom left portion of the disk. (Courtesy of JPL/NASA.)

to make a continuous lattice of protons surrounded by a mobile, unbound sea of electrons. This metallic hydrogen is an excellent electrical conductor, that can participate in generating planetary magnetic fields. In Saturn, the atmosphere is found to be dramatically depleted in helium, requiring a region deep in the atmosphere where helium and hydrogen unmix from each other (their higher mutual solubility in the hotter Jovian interior keeps them in solution with each other). The denser helium sinks to form a fluid helium "mantle" under the hydrogen-enriched atmosphere. Uranus and Neptune, with their ice- and rock-rich interiors, may contain massive oceans of extremely hot, dense water filled with a wide variety of solutes, from ammonia to rocks, that dissolve in water at high pressures and temperatures.

The most powerful observational constraint on the internal structures of a planet is derived from its optically determined polar flattening (oblateness) and its rotational moment of inertial about its principal axis. The optically derived oblateness f is the fractional difference between the measured equatorial radius R_{eq} and the polar radius R_p:

$$f = (R_{eq} - R_p)/R_{eq}.$$

The moment of inertia influences the motions of satellites of the planet, which feel a gravity field that differs significantly from that of a perfectly symmetrical (spherical) planet or a mass point. The external gravity potential of the planet is conveniently expressed as

$$V = -\frac{GM}{r}\left[1 - \sum_{n=1}^{\infty}\left(\frac{R_e}{r}\right)^{2n} J_{2n} P_{2n}(\cos\theta)\right],$$

where M is the mass of the planet and the given angle is the colatitude. The P_{2n} terms are the associated Legendre polynomials and the J_{2n} are the gravitational moments of the planet.

The oblateness and radial distribution of mass in the planet are manifested in the moment of inertial factor C, which is smaller for centrally condensed bodies with dense cores than it is for uniform spherical bodies:

$$C = I/MR_{eq}{}^2.$$

The value of C for a spherical body of uniform density (the Moon is quite a good example) is 0.400; that for a Jovian planet with a dense core is much less.

The polar flattening and the moment of inertia are related by

$$C - \frac{15}{8}\left(\frac{2}{5} - C\right)^2 = J_2/f.$$

Interior models constrained by observations of oblateness and the external gravity field are available for all four giant planets. The central pressures of Jupiter and Saturn are about 100 and 76 Mbar, respectively (1 bar is a pressure of 10^6 dynes/cm^2, or about 1.015 standard atmospheres; a "metric atmosphere"). These pressures are far above the metallization point of hydrogen at relevant temperatures; however, it should be kept in mind that the central regions of all these planets are occupied by massive rock-plus-ice cores which displace hydrogen well out from the center of the planet. The Uranian planets have central pressures of about 16 Mbar, with hydrogen displaced so far from the planetary center that it does not experience high enough pressures to be metallized. Nonetheless, Uranus, like both Jovian planets, has a strong magnetic field whose center is offset 0.3 Uranus radii from the center of the planet. The implication of this offset is that the field is rooted in convective motions in an electrically conducting fluid layer, probably a deep "ocean" of water with a wide range of ionized materials dissolved in it.

It is common for planetary physicists to speak of hypothetical "metallic iron cores" inside the giant planets. More likely would be cores of metallized iron oxides and sulfides, not metallic iron. Most likely in the Uranian planets would be a "rock soup" of oxidized iron compounds dissolved in hot, high-pressure oceans.

The atmospheres of the Jovian planets have been studied principally by spectroscopy and radiometry from Earth and from spacecraft missions. Spectroscopic studies are limited by the near-complete cloud coverage and intensely cold cloudtop conditions (140 K on Jupiter; 90 K on Saturn), which condense most components of the atmosphere and remove them from view. Nonetheless, hydrogen, helium, ammonia, methane, phosphine, germane, hydrogen cyanide, carbon monoxide, and a number of hydrocarbons made by the action of ultraviolet light on the simpler gases have been found on Jupiter. Saturn, colder, smaller, and farther from the Sun and Earth than Jupiter, has revealed a shorter list of atmospheric gases drawn from the same list. The Uranian planets show only hydrogen, helium, methane, and traces of a few heavier hydrocarbons in their infrared spectra. The *Galileo* orbiter, which was launched to Jupiter in 1989, was accompanied by a probe which was successfully dropped into Jupiter's atmosphere to measure atmospheric composition and cloud structure down to levels far below the cloudtops. By a highly improbably coincidence, the *Galileo* entry probe fell in a cloudless region of subsiding, dry stratospheric gases. It provided valuable data on the gases present, but failed to find and study the water cloud layer and its associated water vapor.

The *Cassini* mission now en route to Saturn also carries an entry probe, *Huygens*, for delivery not into Saturn, but into the atmosphere of Saturn's largest moon, Titan. The results of this mission are eagerly awaited.

VI. SATELLITES

The number of known satellites is about 60 and constantly increasing. Of these only 3, the Moon and Mars' two small satellites Phobos and Deimos, belong to the terrestrial planets. The Moon, treated in detail in Section IV, is a terrestrial-type body composed of differentiated material but missing a core. Phobos and Deimos are also rocky bodies, but appear more similar to undifferentiated asteroids in composition and structure. It appears that none of the satellites of the terrestrial planets may be native to the planet around which they presently orbit: the Moon is apparently a collision product, and Phobos and Deimos may have been captured by Mars early in the history of the solar system. Pluto has a large moon, Charon, which pursues a 6-day orbit around Pluto and is rotationally locked on to its primary. Pluto, in turn, is rotationally locked on Charon, making a 1:1:1 spin–spin–orbit resonance for the system. The giant planets account for all but these 4 of the known satellites.

Spacecraft missions to the outer solar system have revealed an astonishing diversity of satellites, many of which are or have been geologically active. Melting and differentiation, often abetted by continuing strong heat sources, have throughly reprocessed many of the larger satellites. The *Voyager* 1 *and* 2 flyby missions and the *Galileo* Jupiter probe and orbiter mission, soon to be supplemented by the *Cassini* Saturn orbiter and the *Huygens* Titan entry probe, have vastly surpassed all previous Earth-based astronomical studies of these bodies. Tidal interaction with their primary bodies have in several case strongly heated the larger satellites. Io, the innermost of the four large Galilean satellites of Jupiter, is wracked by powerful volcanic eruptions of sulfur dioxide and sulfur vapor. The mass flux observed from Io's volcanoes is sufficient to coat the entire surface of Io with about 1 mm of erupted material per year, or about 1 km of material per million years. The volcanic ejecta would amount to approximately the entire mass of Io over the age of the solar system. Not surprisingly, the very young surface of Io shows no record of bombardment cratering. Io is locked in a 2:1 orbit–orbit resonance with Europa. Europa's tidal forces pump up the eccentricity of Io's orbit to a value of 0.0041. This seemingly small eccentricity is important because it means that the instantaneous angular rates of rotation and orbital motion of Io cannot be equal. The huge tidal bulge raised in Io's thin, flexible crust by Jupiter is pumped up and down by about 100 m as Io's distance from Jupiter varies, heating the crust severely and keeping the silicate interior hot and active Fig. 4). Europa, the next Galilean satellite in order of distance from Jupiter, appears to have a thin ice crust overlying a deep ocean of liquid water. Although it is farther from Jupiter, tidal interactions with the other

FIGURE 4 Photograph of an active volcanic eruption on Jupiter's satellite Io taken by *Voyager 1* on March 4, 1979. On the limb of the satellite can be seen one of at least four simultaneous volcanic eruptions. The observed volcanism is extremely explosive, with initial velocities of more than 2000 mi/hr (~1 km/sec); the ejecta reach heights of 60 mi or more. Several eruptions have been identified with volcanic structures on the surface of Io, which have also been identified by *Voyager's* infrared instrument as being abnormally hot—several hundred degrees kelvin warmer than the surrounding terrain. (Courtesy of JPL/NASA.)

Galilean satellites also pump up its orbital eccentricity and heat it in a manner similar to Io. Impact scars are also rare on Europa, attesting to a mobile and rapidly resurfaced crust. Ganymede, a Mars-sized moon of Jupiter, has also evidently melted and differentiated in the distant past. Callisto, the outermost of the four, may never have differentiated and retains dark dust intermingled with its heavily cratered, ancient surface ices.

The dense atmosphere of Saturn's largest satellite, Titan, is principally composed of nitrogen and methane. The atmospheric pressure on the surface of Titan is roughly 1.5 bars, even higher than on Earth. Titan's gases are vulnerable to the chemical effects of solar ultraviolet radiation, which produces traces of higher hydrocarbons and nitriles such as CH_3CN. These gases in turn are less volatile than their lighter precursors, and readily condense from the atmosphere and rain out onto the surface at a rate sufficient to make several hundred meters of condensed materials over the age of the solar system. Titan must have widespread deep lakes or seas of liquid nitrogen and light hydrocarbons; however, radar studies of Titan from Earth find a rough and varied durface, ruling out a deep global ocean.

The Galilean satellites exhibit a strong density trend reminiscent of that seen among the terrestrial planets. Io, the densest, must be a rocky body throughout. Europa has tens of kilometers of ice and water on its surface, and Ganymede and Callisto contain roughly equal masses of rocky material and water ice. The simplest explanation for this density trend is that the Galilean satellites formed in a gas–dust accretion disk orbiting around Jupiter, with a strong radial temperature gradient in the disk that prevented ices from condensing close to Jupiter.

Neptune's largest satellite, Triton, and Pluto, both much farther from the Sun than Titan and therefore much colder, have much more tenuous atmospheres with compositions similar to Titan's.

The other satellites of Saturn, and all the satellites of Uranus, are many times smaller in mass than any of the Galilean group. Among them, Saturn's satellite Enceladus seems also to have been heated and dfferentiated by tidal interactions. Mimas (Fig. 5) also seems to be a differentiated body. All of these satellites have densities consistent with mixtures of rocky and icy solids, in some cases very ice-rich. At greater distances from the Sun, where lower temperatures prevail, ices other than water ice, including ammonia and its hydrates, methane and its hydrates, and carbon oxides, are expected to be increasingly important. The limiting example is cometary ice, which is known to contain large amounts of these other, more volatile, ices.

The small, outermost satellites of Jupiter present an interesting dynamical problem. The two families of bodies, orbiting the planet in opposite directions, are most easily inderstood as having been captured from independent orbits around the Sun. Capture requires some means of dissipating the excess energy of the body being captured, which may be done by gas drag near the outer edge of Jupiter's preplanetary accretion disk.

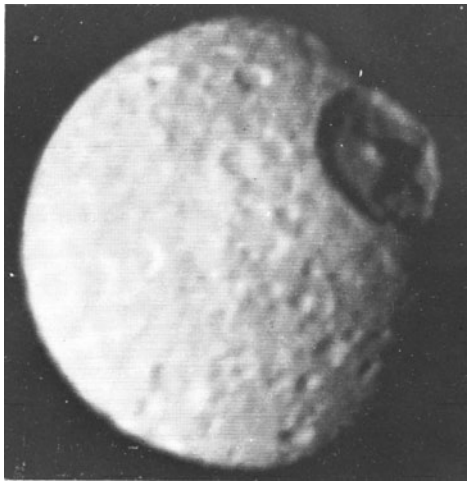


FIGURE 5 Image of the cratered surface of Saturn's satellite Mimas taken by *Voyager 1* on November 12, 1980. The prominent crater, Herschel, with a diameter of 140 km and a depth of 5 km, is the result of an impact that probably came close to shattering the entire satellite. (Courtesy of JPL/NASA.)

Triton, unique in being a massive satellite in a retrograde orbit, has very probably experienced severe tidal heating in its past. The tidal energy liberated by circularization of its orbit could have melted and differentiated Triton's interior even without assistance from the decay of long-lived radionuclides.

Many of the smaller satellites of the giant planets bear impact craters that approach the size of the body itself. These satellites are presumably collisional fragments from the erosion or disruption of larger ancient bodies.

VII. RINGS

Before 1979, Saturn was the only planet in the solar system known to have a ring system. Saturn's rings are both broad and bright, and if similar systems existed around any of the other planets, they would have been discovered a long time ago. However, we now know that a wide variety of ring systems can exist. All the planets have rings, but the rings have little in common, apart from the fact that most exist close to the planet, inside Roche's limit.

Roche's limit, the distance from a planet at which the tidal and rotational forces acting on a satellite would tend to disrupt it, depends on the properties of the satellite material and on the satellite's shape. For a satellite in synchronous rotation with no cohesive strength, that is, for a satellite with a rotational period equal to its orbital period that is held together by self-gravitation alone, the shape of the satellite is that of a triaxial ellipsoid and Roche's limit, R_1, is given by

$$R_1 = 2.46(\rho_p/\rho_s)^{1/3} R_p,$$

where ρ_p and ρ_s are the mean densities of the planet and satellite, respectively, and R_p is the radius of the planet.

The particles of Saturn's rings are primarily icy, but there is evidence of some albedo and therefore some compositional variations within the ring system. Most of the particles are in the 1-cm to 5-m size range, but wave structures detected by the *Voyager* spacecraft are strong evidence that small satellites with radii of $\sim$10 km also exist. The Jovian ring is optically thin and appears to contain little structure. However, because of smear motion, features smaller than $\sim$700 km are difficult to resolve in the *Voyager* images. The Jovian ring particles are micronsized and have short lifetimes ($<10^4$ yr) limited by erosion due to sputtering and meteoroid impacts. These particles must be replenished, probably by some source within the rings. This fact alone suggests the existence of a number of small,

unseen satellites within the rings, in addition, perhaps, to the two small, dark satellites that orbit near its outer edge.

The Uranian rings have a structure that is, to some extent, the opposite of that of the Saturnian ring system. The Saturnian system consists of thousands of ringlets and contains few clear gaps, whereas the Uranian rings are narrow and widely separated. All the Uranian rings, except the innermost ring discovered during the *Voyager* flyby, which is broad and diffuse, are optically thick. Some of the rings are eccentric, inclined to the planet's equator, and nonuniform in width and have very sharp edges. The most prominent ring is the outermost, ε ring. This eccentric ring increases in width from a minimum of 22 km at pericenter (the nearest point to the planet) to a maximum of 93 km at apocenter (the farthest point from the planet). Spectra of the rings in the wavelength range 0.89–3.9 μm show their geometric albedo (a measure of the reflectivity of the particles) to lie between 0.02 and 0.03, ruling out ice-covered particles: The ring particles are balck. *Voyager* observations indicate that particles in the ring have diameters of >10 cm.

The theory of shepherding satellites of P. Goldreich and S. Tremaine successfully accounts for the existence of narrow, eccentric rings with sharp edges. This theory postulates that a narrow ring is confined by the tidal torques exerted on it by a pair of nearby satellites, one orbiting interior to the ring and the other exterior to it. The discovery of two small satellites (Fig. 6) bounding the Uranian ring strongly supports this theory. The Neptunian rings are partial arcs and probably consist of particles trapped in resonances with unseen satellites. Partial arcs also appear to be evident in the *Voyager* Uranian data.

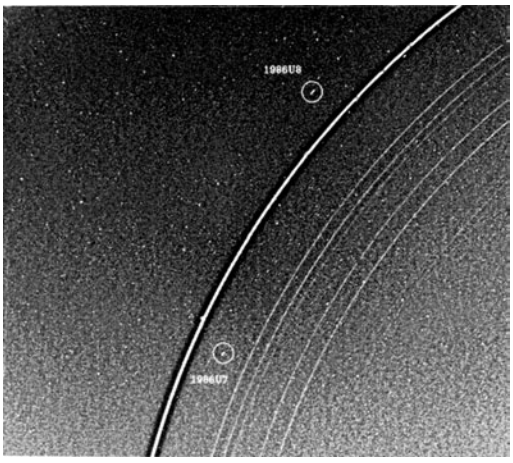


FIGURE 6 Image taken by *Voyager 2* on January 21, 1986. This is the first direct observation of the Uranian rings in reflected sunlight. Evident are the two "shepherding" satellites (discovered by *Voyager 2*) that bound and confine the bright, outermost ε ring. Lying inward from the ε ring are the δ, γ and η rings; then the β and α rings; and the barely visible 4, 5, and 6 rings. *Voyager 2* discovered two other faint rings that are not visible in this image. (Courtesy of JPL/NASA.)

Evidence is accumulating that many, if not all, the observed planetary rings are much younger than the age of the solar system. On the other hand, more rings will probably be created in the future—some by the cometary disruption of small satellites that orbit close to a planet, others by the disruption of small satellites by tidal forces. The Martian satellite Phobos has an orbit inside the planet's synchronous orbit (the orbit for which a satellite's orbital period equals the planet's rotational period), and due to tides raised on the planet by the satellite, the satellite's orbit is decaying. It is anticipated that in 10^8 yr the satellite will be pulled inside Roche's limit, where it may (its cohesive strength is unknown) disintegrate and form a ring system.

VIII. ASTEROIDS, COMETS, AND METEORITES

Most of the known asteroids have orbits in the main asteroid belt, between the orbits of Mars and Jupiter. Most of these belt asteroids have orbital semimajor axes between 2.2 and 3.3 AU, with eccentricities of order 0.2 and inclinations of order 15°. They are very diverse in albedo (reflectivity) and in their spectral properties. The period or semimajor axis distribution of asteroids is markedly clumpy, with several well-defined orbital families and several gaps in the period distribution corresponding to orbital resonances with Jupiter. Because of the eccentricities of asteroid orbits, however, there are no gaps discernible in the instantaneous distribution of asteroids over heliocentric distance.

The asteroids exhibit a steep size distribution, in which the largest single asteroid, 1 Ceres, accounts for the majority of the total mass of the population. There are approximately 500 known belt asteroids with diameters greater than 50 km, and approximately 50,000 with diameters larger than 1 km. Virtually all of the asteroids of diameter less than 50 km must be fragments from the collisional disruption of a much smaller initial population of large asteroids. The clumpiness of the spatial distribution of asteroids must at least in part be due to the breakup of large parent asteroids. In any case, the total mass of the asteroids is only a few percent of the mass of the Moon. Thus 19th century theories that attribute the asteroid belt to the explosion of a planet that once orbited in the belt have no physical basis. The chemical evidence supports the same conclusion: the only high-pressure minerals found in meteorites have obvious evidence of severe mechanical shock, and no evidence of prolonged crystallization at the high pressures characteristic of planetary interiors.

The belt exhibits clear compositional banding. Closest to the Sun, near 2.2 AU, are several classes of bodies whose multicolor visible and infrared spectra are best matched by laboratory spectra of differentiated planetary materials, such as metallic (M), enstatite achondritic (E), and possibly stony-iron meteorites. The heart of the belt is dominated by S (stony)-type asteroids, which have been variously identified by asteroid observers as stony iron or ordinary chondritic material. The outer half of the belt is in turn dominated by a sequence of low-temperature carbonaceous types, the C, S, and P classes, which range from similar to laboratory samples of carbonaceous chondrites (C) to super-carbonaceous (D, P) bodies that are presumably also rich in water. These classes bespeak a strong gradient in oxidation state and volatile-element content with distance from the Sun, probably attributable to the general decline of temperatures with heliocentric distance in the pre-solar nebula out of which the Sun and planets formed.

Over the past 100 years it has become evident that there is also a significant population of asteroids in orbits that range in far closer to the Sun than the edge of the belt proper. The Amor asteroids are defined as those that reach 1.3 AU or less at perihelion, but do not cross Earth's orbit. The Apollo asteroids have perihelion distances below 1.015 AU (Earth's aphelion distance, and are therefore either present-day Earth-crossers or episodic Earth-crossers. The Aten asteroids are those with orbital semimajor axes less than 1.000 AU, which therefore have periods less than one Earth year. There may be a small population of asteroids whose orbits are completely internal to Earth's and are not Earth-crossers at present. The Apollo and Aten asteroids are capable of colliding with Earth, and many Amors may from time to time be perturbed into Earth-crossing orbits. These three groups are collectively referred to as the Near-Earth Asteroids (NEAs).

Recent discoveries of NEAs have proceeded at a record pace. The best present estimate is that there are roughly 1000 NEAs larger than 1 km diameter. The mean time between consecutive impacts of NEAs larger than 1 km in diameter is 4×10^5 years. Comet and asteroid impacts capable of having global-scale climatological effects (those at least several hundred meters to about 2 km in diameter) are therefore common over geological time scales. Some of the most severe biological extinction events in Earth's history were caused by impacts, of which the best documented is the extinction event at the end of the Cretaceous Era, 65 million years ago. The end-Cretaceous impactor, which struck in shallow oceans on the north slope of Mexico's Yucatan Peninsula, was roughly 10 km in diameter. There is also recent evidence of a pervasive worldwide deposit of fullerene molecules, containing noble gases with nonterrestrial isotopic signatures, at the time of the most severe extinction event in the last 650 million years, the end of the Permian Era.

Comets are bodies that orbit the Sun, usually in distant and eccentric orbits, whose evaporation behavior reveals that they are composed of ice–rock mixtures. A typical comet, when observed close to the Sun, contains a small, dense solid nucleus, probably very dark in color, surrounded by a spheroidal coma of gases volatilized from the nucleus by solar heating, and often accompanied by two tails. One of these tails has a bright-line and band spectrum of tenuous gases stimulated by solar ultraviolet light, featuring simple molecules, radicals, and ions composed principally of volatile compounds of hydrogen, oxygen, carbon, nitrogen, and sulfur. This "plasma" (partially ionized gas) tail may also contain rock vapors when the comet is extremely close to the Sun. The plasma tail is "blown" radially outward from the Sun by the solar wind. The other tail is composed of dust, showing a spectrum the looks like continuous sunlight reflected from dark particles of dust. This dust tail, once free of interaction of the gases of the coma, follows ballistic orbits around the Sun perturbed, over the long term, by the interaction of grains with the solar radiation field (the Poynting–Robertson and Yarkovsky effects) into narrow, dense bands of dust that often can be traced back to their parent comet or to comets now extinct due to the exhaustion of their supply of volatile ices. Comet nuclei are typically a few km in diameter, although some faint comets may be much smaller, and rare Great Comets may be even larger. The size of a cometary coma is typically tens of thousands of kilometers in diameter, and tails can in some rare cases be as much as 100 million km long.

The vast majority of all comets follow randomly oriented orbits in a distant, grossly spherical cloud well outside the planetary system, in a region called the Oort Cloud. Perhaps 10^{12} comets, with an aggregate mass similar to the mass of Earth, orbit in this region until disturbed by the gravitational perturbation of a passing star into orbits that penetrate close enough to the Sun to begin rapid evaporation and development of a coma and tails. Such "fresh" comets have orbital periods of millions of years, perihelion distances of about 1 AU, aphelion distances of a few times 10^4 AU (in the Oort cloud), orbital ecentricities of about 0.9999, and random orbital inclinations: roughly half of these long-period comets are traveling in retrograde orbits around the Sun, opposite to the direction of motion of the planets. Long-period comets have, at most, visited the inner solar system once during the history of the human race. Smaller populations of comets are found in intermediate (100- to 1000-year) orbits and in short-period (<100-year) orbits. Many of the latter have been observed on several perihelion passages. Their orbits are generally of moderate eccentricity and low prograde inclination, many with nodes close to the orbit of Jupiter or Saturn. The

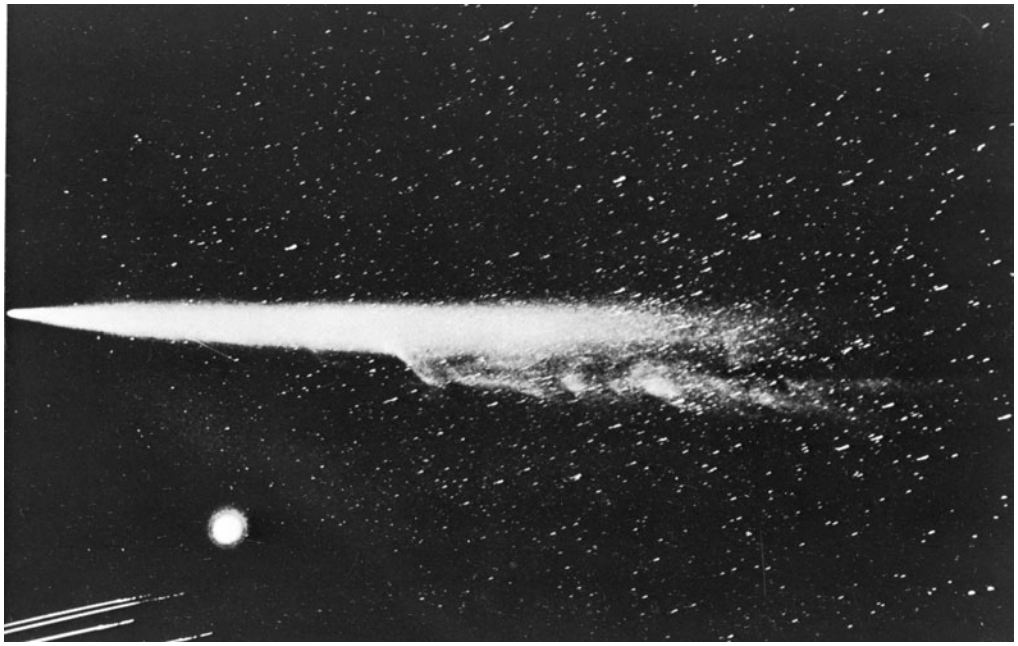


FIGURE 7 Halley's comet as photographed at Lowell Observatory in May 1910. The bright object beneath the comet is the greatly overexposed image of the planet Venus. The streaks at the lower left are the Flagstaff, Arizona, city street lights. The recent flyby of the comet by Soviet, Japanese, and European spacecraft revealed a potato-shaped body, $15 \times 8 \times 8$ km in size, that rotates with a period of 53 ± 3 hr. The surface is black. (Courtesy of Lowell Observatory.)

short-period ("periodic") comets have apparently been perturbed from long-period orbits into their present orbits as a result of close flybys of the Jovian planets.

The 1986 apparition of Halley's Comet was closely observed by several spacecraft launched from Earth, including the Russia *Vega* 1 and 2 (for "Venera-Galley" or Venus-Halley in Russian), the European Space Agency's *Giotto* spacecraft, and the Japanese *Sakigake* distant flyby. Close-up imaging of the nucleus of Halley by *Giotto* reveal a very dark, potato-shaped nucleus about 8×15 km in dimensions, marked by several violent jets of erupting gases and dust (Fig. 7).

The dust streams left behind by cometary activity usually cross Earth's orbit, and sometimes intersect the Earth's position in space. When they do, the tiny cometary dust particles strike the top of Earth's atmosphere at very high speeds, often over 30 km/s, and vaporize in a fraction of a second to produce a bright streak of sight in the sky. These luminous phenomena, first ascribed by James Joule in 1848 to the conversion of the kinetic energy of fast-moving solid particles into heat, are called a meteor shower. Meteor showers have never resulted in the fall of a single meteorite onto Earth's surface.

A meteorite, defined as an extraterrestrial rock that survives its passage through Earth's atmosphere to strike the surface, is a free sample of extraterrestrial material, usually of unknown provenance. On some rare occasions, the entry of a meteorite is observed photographically with sufficient detail and coverage to permit the determination of its preatmospheric orbit around the Sun. Also, sometimes the reflection spectrum of a fallen meteorite can be linked to the reflection spectrum of a particular asteroid in a compatible orbit, thus establishing a presumptive case for parentage from that asteroid. Any meteorite falling to Earth must come directly from an Earth-crossing orbit; however, such orbits are typically only temporary, with lifetimes against planetary impacts that are very much shorter than the age of the solar system. Asteroidal material must therefore be fed into the near-Earth zone continually. The most powerful source of such new material is gravitational perturbations of belt fragments by Jupiter. The ultimate place of origin of most meteorites is surely the asteroid belt, although dozens of meteorites out of the 20,000 known are fragments knocked off of the Moon (certain basalts and other igneous rocks) or Mars (the "SNC" meteorites; shergottites, nakhlites and chassignites) by large impacts. Some NEAs also have orbits and even spectral properties suggestive of derivation from extinct short-period comets. The absence of a detectable coma or tail can be understood if volatile ices have been exhausted in the outermost few meters of the nucleus by repeated heating at perihelion passages.

Most meteorites have "primitive" compositions and textures, meaning that they have very great ages of about

4.55 Ga, have never undergone thermal evolution to the point of melting or differentiating, and still retain their primordial content of volatiles plus all the helium, argon, and fission-produced krypton and xenon made inside them since the origin of the solar system. The asteroidal meteorites are effectively physicochemical probes of the very earliest history of the solar system, from times before the planets were accreted. The lunar and martian meteorites, in contrast, are much younger and highly differentiated, bearing priceless evidence regarding the ancient history of these two worlds. There is still hope that samples of Mercury and Venus may also be recovered and recognized among the meteorites found on Earth.

In early 2001 The NEAR (Near-Earth Asteroid Rendezvous) mission reached, orbited, mapped, and finally landed successfully on the surface of the NEA Eros. Many of the phenomena observed by NEAR, such as the very high boulder population and the startling dearth of small craters, are still under intensive study.

IX. RELATIONSHIPS TO OTHER PLANETARY SYSTEMS

Since 1995 approximately 60 planets have been discovered in orbits around other stars. Because of severe limitations on the search techniques now in use, only massive, close planets of relatively nearby low-mass stars can be found. All the orbiting bodies detected to date (by measurement of the radial component of the reflex orbital motion of their parent stars around their common center of mass) are comparable in size to Jupiter or Saturn, or larger. Many of the bodies found in this search appear to be brown dwarfs, bodies so large that they once passed through an era of fusing deuterium in their interiors (and hence are not planets), but which presently have exhausted their deuterium supply and maintain their luminosity by their slow collapse, converting gravitational potential energy into heat (and hence are not stars). Search techniques are currently close to detection of Uranus-sized planets; however, it will not be possible for many years to detect the presence of terrestrial-sized planets in these systems.

The most startling result of the search to date has been the surprisingly large number of systems in which Jovian or super-Jovian planets have been found in orbits very close to their parent stars; in some cases, so close that the "surface" temperature of the planet is over 1500 K. It is tempting to say that we see such odd systems because they are the only ones the present detection methods, with their several-year run of data, are capable of detecting. A Solar-System-like planetary system could not have been detected, because none of the giant planets have even come close to completing an orbit in five years, and because the large distance of Jupiter from the Sun and the Sun's relatively high mass make the reflex velocity of the Sun too small to detect. Theoretical studies have shown that inward migration of large planets can occur under some circumstances, but it is still far to early to come to any final judgments regarding the frequency and stability of other solar systems similar to our own.

X. ORIGIN OF THE SOLAR SYSTEM

A wide range of observations of star-forming regions in the Galaxy show that stars are formed in large numbers out of dense interstellar gas and dust clouds in quite short intervals of time, on the order of a million years. Once a few supermassive, superluminous stars form in this dense cluster of protostars, unaccreted gases are strongly heated by ultraviolet radiation from these largest stars, and flow outward into space. The resulting decrease in the total mass of the cluster causes it to cease to collapse. The differential motions of the young stars, now no longer gravitationally bound to each other, causes the cluster to expand outward into the surrounding galactic arm, where the cluster is smeared out and loses its identity as a result of differential rotation of the Galaxy. During this dissipation phase, which has a time scale of tens of millions of years, the most massive main sequence stars evolve past the end of hydrogen burning and either eject considerable mass quiescently, or, if massive enough, destroy themselves in a supernova explosion. Less massive hydrogen-burning stars dispel the excess luminosity of their T-Tauri phase and settle down onto the main sequence, blowing away the last remnants of their protoplanetary gas and dust disk. Students of the solar system identify the nebular phase as the time of formation of large solid bodies and gas-giant planets, during which the most ancient solids were accreted into asteroids and larger bodies. The chondritic meteorites are samples of those asteroidal bodies that were too small, or too far from the T-Tauri phase Sun, to be heated to the melting point. Astronomical observations of dense dust disks around young stars such as Vega and Beta Pictoris show that these disks may have radii of several hundred AU, about ten times the diameter of the documented flattened-disk population in the present solar system (the planets, asteroids, Chirons, and Kuiper belt bodies).

The high opacity and strong internal heat sources in the preplanetary nebula develop a strong radial temperature gradient, with temperatures high enough to vaporize rock close to the proto-Sun and cold enough to condense all except the permanent gases (hydrogen, helium, and neon) at tens of AU from the center.

Meanwhile, frictional forces between gases and dust and turbulent viscosity of the gas lead to rapid evolution of the nebular disk on a time scale of hundreds of thousands of years, pumping mass inward and angular momentum outward.

Two major mechanisms appear responsible for the formation of planets from such a preplanetary nebula: accretion of solid bodies via low-velocity collisions, and gravitational instability of the gas disk, leading to fragmentation into massive gas-rich planets.

SEE ALSO THE FOLLOWING ARTICLES

COSMOLOGY • METEORITES • MOON (ASTRONOMY) • PLANETARY ATMOSPHERES • PLANETARY RADAR ASTRONOMY • PLANETARY SATELLITES, NATURAL • PRIMITIVE SOLAR SYSTEM OBJECTS: ASTEROIDS AND COMETS • SOLAR SYSTEM, MAGNETIC AND ELECTRIC FIELDS • STELLAR STRUCTURE AND EVOLUTION

BIBLIOGRAPHY

Vilas, F., Chapman, C. R., and Matthews, M. S. (eds.) (1988). "Mercury," Univ. of Arizona Press, Tucson.

Bougher, S. W., Hunten, D. M., and Phillips, R. J. (eds.) (1997). "Venus II," Univ. of Arizona Press, Tucson.

Kieffer, H. H. (ed.) (1992). "Mars," Univ. of Arizona Press, Tucson.

Binzel, R. P., Gehrels, T., and Matthews, M. S. (eds.) (1989). "Asteroids II," Univ. of Arizona Press, Tucson.

Gehrels, T. (ed.) (1976). "Jupiter," Univ. of Arizona Press, Tucson.

Gehrels, T., and Matthews, M. S. (eds.) (1984). "Saturn," Univ. of Arizona Press, Tucson.

Berstralh, J. T., Miner, E. D., and Matthews, M. S. (eds.) (1991). "Uranus," Univ. of Arizona Press, Tucson.

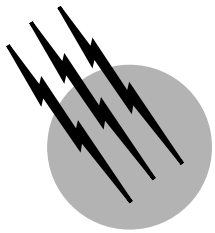
Miner, E. D. (1998). "Uranus: The Planet, Rings, and Satellites," 2nd edition, Wiley, New York.

Cruikshank, D. P. (ed.) (1995). "Neptune and Triton," Univ. of Arizona Press, Tucson.

Stern, S. A., and Tholen, D. J. (eds.) (1997). "Pluto and Charon," Univ. of Arizona Press, Tucson.

Burns, J. A., and Matthews, M. S. (eds.) (1986). "Satellites," Univ. of Arizona Press, Tucson.

Newburn, R. L., Jr., Neugebauer, M., and Rahe, J. (eds.) (1991). "Comets in the Post-Halley Era," Kluwer, Dordrecht, The Netherlands.



Solar System, Magnetic and Electric Fields

C. T. Russell
University of California, Los Angeles

GLOSSARY

Adiabatic invariants First, second, and third adiabatic invariants are conserved quantities associated with the three periodic motions (gyro, bounce, and drift) of charged particles trapped in a magnetic mirror configuration, such as the Earth's dipolelike field.

Bow shock Collisionless shock wave in the solar wind plasma that stands in the flow, slows and heats the flow, and deflects it around all planetary obstacles.

Coronal mass ejection (CME) A large-scale eruption of the solar corona that accelerates away from the sun and produces a large ($\sim 45^\circ$ across) disturbance in the solar wind. Best observed by coronagraphs viewing the limbs of the sun. Halo CMEs are CMEs moving directly at the observer and producing a disturbance in projection all around the sun.

Corotational electric field Electric field in a planetary magnetosphere associated with the rotation of the plasma at the same angular rate as the planet because of the high electrical conductivity of the plasma along magnetic field lines.

Debye length Electrical shielding length in a plasma. A test charge in a plasma cannot be sensed beyond this distance.

Field-aligned current Electric current flowing along planetary magnetic field lines connecting stresses in one part of a magnetosphere with those in another. Current system closes by flowing across the magnetic field.

Geomagnetic storm Period of several days in which currents circling the Earth in the equatorial plane of the

magnetosphere become enhanced. The energization of these currents is caused by changes in the solar wind and interplanetary magnetic field.

Gyrofrequency Number of times per second that a charged particle orbits a magnetic field line. Depends directly on the particle's charge and magnetic field strength and inversely as the mass of the particle.

Gyroradius Radius of orbit of charged particle in a magnetic field. Depends directly on particle mass and velocity perpendicular to the magnetic field and inversely as the charge and magnetic field strength.

Interplanetary coronal mass ejection (ICME) The disturbance in the solar wind associated with a coronal mass ejection from the sun. This disturbed region often is preceded by a collisionless bow shock and contains embedded within it multiple magnetic flux ropes.

Interplanetary magnetic field Magnetic field of the solar wind carried out from the Sun by the solar wind flow.

Ionopause Upper boundary of an ionosphere that interacts directly with the solar wind.

Ionosphere Ionized part of the atmosphere of a planet.

Magnetic cloud Interplanetary structure in which the magnetic field is enhanced, usually quiet and slowly rotating. Resembles a twisted magnetic flux tube and usually found within an ICME.

Magnetohydrodynamics Physics of magnetized electrically conducting fluids. Often applied to plasmas in situations in which they display fluidlike behavior.

Magnetopause Outer boundary of a magnetosphere confined by the solar wind.

Magnetosheath Shocked plasma behind a planetary bow shock flowing around the planetary magnetopause or ionopause.

Magnetosphere Magnetic cavity formed by the interaction of the solar wind with a planetary obstacle.

Magnetotail Long cylinder of magnetic field lines dragged out in two oppositely directed tail lobes behind a planetary obstacle.

Parallel electric field Electric field along a planetary magnetic field that is responsible for the acceleration of the electrons that cause the brightest aurora and decouple flows perpendicular to the magnetic field at different altitudes along the field line. Parallel fields also accelerate electrons into the magnetosphere from the ionosphere, and they enable reconnection by permitting the switching of magnetic partners.

Plasma Gas, fully or partially ionized, having equal densities of electrons and ions in which the energy of motion of the particles exceeds that associated with the electric potential of the charged particles. In an unmagnetized plasma, particle motions are nearly straight lines.

Plasma frequency Natural oscillation frequency of a plasma set into motion by a small separation of the electrons and ions.

Plasma parameter Number of electrons in a cube whose side is a Debye length. "Collective" plasma behavior occurs when this number is much greater than unity. It also is a rough measure of the number of plasma oscillations between interparticle collisions.

Reconnection Process in which the magnetic topology of the magnetic field in an element of plasma changes. For example, at the dayside magnetopause, the magnetic field lines in the postshock solar wind plasma become linked with the terrestrial field lines and accelerate the plasma on the newly joined field lines.

Solar wind Supersonically expanding upper atmosphere of the Sun that because of its high electrical conductivity carries the solar magnetic field with it.

Substorm Disturbance in the night-time current systems in the terrestrial magnetosphere usually lasting a couple of hours and accompanied by enhanced auroral activity.

Sunspot cycle Eleven-year cycle in which cool, dark, magnetized regions seen in the photosphere become first more frequent and then less frequent. Because the magnetic polarity pattern of sunspot groups changes every 11 years, the cycle is in actuality a 22-year cycle. This period is not exactly constant but varies slightly from cycle to cycle.

Tail lobes Two regions within a magnetotail of oppositely directed magnetic flux in which the magnetic energy density greatly exceeds the plasma pressure.

THE STUDY of solar system electric and magnetic fields includes the investigation of the magnetic fields generated by electrical currents flowing in the interior of the Sun and the planets as well as electric and magnetic fields and their associated current systems flowing in the solar wind and in the magnetospheres and the ionospheres of the planets. Magnetic fields are generated by electric currents that arise when more particles of one charge flow in a particular direction per unit time than particles of the opposite sign. Such currents can flow in the highly electrically conducting fluid cores of the planets, which are both differentially rotating and convecting. Such currents are also found in the various plasmas of the solar system. These plasmas, or electron–ion gases, are highly electrically conducting. The material inside the Sun, its atmosphere, the expanding solar wind, and ionized material in the upper atmospheres of all the planets are plasmas. One of the more important of these current systems is the one that generates the solar magnetic field. Finally, there are currents at the atomic level in solid materials. While our usual exposure to magnetic fields of this kind is from man-made

magnets, nature also produces magnetized materials. Such remanent magnetization often preserves a record of the magnetic field at the time of the formation of the material and thus the study of magnetized rocks has led to an understanding of the magnetic history of both the Earth and the Moon.

Magnetic fields have both magnitude and direction. The most common instrument for sensing this direction is the compass. The scientific instrument for measuring magnetic fields is called a magnetometer. There are several different types of magnetometers in common use. Proton precession magnetometers measure total field strength and are frequently used in terrestrial studies. Fluxgate magnetometers measure components of the magnetic field in a particular direction and are most commonly used on spacecraft. Solar magnetic fields are sensed remotely through the splitting of spectral lines by the Zeeman effect. Magnetic fields are measured in nanotesla, gauss, and teslas ($1\ \mathrm{T} = 10^4\ \mathrm{G} = 10^9$ nT). The magnetic field on the surface of the Earth at the equator is about 0.31 G. In the outer reaches of the Earth's magnetosphere it is about 0.001 G or 100 nT. Outside the magnetosphere in the solar wind it is about 10 nT. The magnetic field on the surface of the Sun is highly variable but on average it has a magnitude of several gauss.

Electric fields are generated by the separation of positive and negative charges. Perhaps the most spectacular naturally occurring separation of charges is produced in convecting clouds in the Earth's troposphere leading to lightning discharges. Since electrons and ions differ greatly in mass they often react quite differently to plasma effects. The different reactions also lead to charge separations in plasmas. If an electric field is applied to a magnetized plasma in a direction perpendicular to the magnetic field, the plasma will drift in a direction perpendicular to both the applied electric field and the magnetic field. This process is referred to as "**E** cross **B**" drift. Conversely, when a magnetized plasma is observed to be drifting, there must be an electric field perpendicular to the magnetic field in the frame of reference of the observer. If an observer is moving with the plasma, she or he sees no drift and hence in this frame of reference there is no electric field. Thus the electric field in a plasma is frame dependent; it depends on the velocity of the observer.

Electric fields have both magnitude and direction. They accelerate charged particles in the direction of the electric field. The energy gained by a particle moving in an electric field is the particle's charge times the electric field times the distance traveled. The electric field times the distance is called the potential drop or potential difference between two points. An electric field is most often detected by measuring the electrical potential difference between two points and dividing by the distance between them. Spacecraft designed to measure electric fields in space often carry long (up to 200 m) antennas. The common units of electric fields are volts/meter and statvolts per centimeter. One statvolt/cm equals 3×10^5 V/m. A typical electric field in the equatorial magnetosphere of the Earth is 0.2 mV/m. A typical electric field in the solar wind outside of the Earth's magnetosphere is 2 mV/m.

I. THE PHYSICS OF SOLAR SYSTEM ELECTRIC AND MAGNETIC FIELDS

A. Sources of Magnetic Fields

There are four basic equations governing the interrelationship of charges, currents, and magnetic and electric fields. These four equations are called Maxwell's laws. In an electrical resistor, current is usually directly proportional to the applied electric field. In a magnetized flowing electrical conductor, an additional term arises because of the electrical field associated with the motion of the conductor. The equation describing the relationship between the current, the electric and magnetic fields, and the flow velocity is called the Ohm's law. Maxwell's laws and Ohm's law can be combined to give what is known as the dynamo equation. This equation shows that the change in a magnetic field is the difference between the resistive decay of currents and the regeneration of the field due to fluid motion. If there is no motion in the conducting fluid, the magnetic field will decay with time. If the fluid core of the Earth froze, the terrestrial magnetic field would decay in a few thousand years. The Jovian field would decay over a few hundred million years, but the solar magnetic field decay time would be comparable to the age of the solar system. Hence, the solar magnetic field could conceivably be in part primordial, present from the time of formation of the Sun due to the compression of preexisting interstellar magnetic fields. Such a primordial field could explain only a steady component.

Cool dark regions on the sun called sunspots are associated with strong magnetic fields. The number and area of sunspots varies with an approximate 11-year cycle. The principal part of the solar field reverses approximately every 11 years producing a solar magnetic cycle with a length of two sunspot cycles. Because of its effect on geomagnetic records it can be shown to have continued for at least 150 years. Evidence for the solar cycle exists for over 2000 years in the available, albeit irregular, optical observations, and various proxy data, such as the fraction of the radioactive isotope of carbon, C^{14}, measured in tree rings. This varying solar magnetic field, the magnetic field of the outer planets, and that of the Earth must be actively

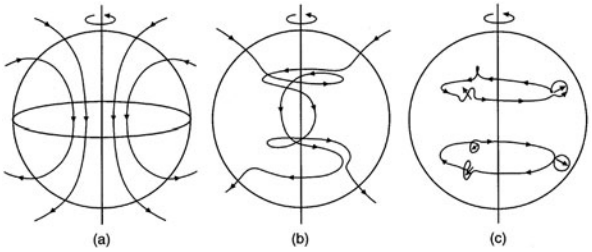


FIGURE 1 Schematic illustration of a self-regenerative magnetic dynamo. (a) The inward extrapolation of the field observed on the surface of the Earth. (b) How the field lines sketched in panel (a) become twisted in azimuth by the rotation of the core, which varies with depth. (c) The effect of rising convective cells on the azimuthal field of panel (b).

maintained by a generator of magnetic fields (i.e., a dynamo).

One approach to studying dynamos is to seek patterns of motion of the conducting fluid that can generate magnetic fields. This kinematic approach is not self-consistent because it does not generate the requisite velocity field from first principles. Furthermore, the magnetic field so generated usually acts on the conducting fluid, thus altering the motion. Figure 1 shows how a planetary field might be self-regenerative if it has a conducting fluid core. The left-hand panel shows a typical planetary magnetic field with field lines confined to planes that contain the axis of rotation of the planet. Deep in the fluid core, as shown in the middle panel, this field is twisted out of these planes by differential rotation of the core. At different depths the core rotates more rapidly. This causes an azimuthal component around the rotation axis, which may become much larger than the original field sketched in Fig. 1a. We know such a differential motion, or shear, in the fluid motions exists in the Earth because features in the terrestrial magnetic field drift slowly westward. We also see differential motions in the solar photosphere. Sunspots move faster at the equator than at higher latitudes. This differential motion, thought to extend well down into the interior, is called the omega effect. Heating, such as produced by radioactive sources or by the freezing out of a solid inner core in the Earth or nuclear fusion in the Sun, produces rising convective cells. These rising convective cells carry the azimuthal field upward as shown in the rightmost panel in Fig. 1. The rotation of the planet causes these convective cells to twist and thus creates a component of magnetic field parallel to the initial field. This new component replaces any field lost due to resistive decay. In this way differential rotation and upward convection combine to regenerate the planetary magnetic field. This process is thought to act in the Earth, Mercury, Jupiter, Saturn, the Sun, and many stars. However, the theory of this process is not yet sufficiently developed to provide a prediction of the strength of a planetary dynamo, even if the physical and chemical structure of the interior of the planet were well known. Table I shows the properties of the dipole moments of the eight innermost planets and the Earth's moon.

TABLE I Magnetic Dipole Moments of Planetary Bodies

Planetary body	Dipole moment ($T\text{-}m^3$)	Tilt (deg)
Mercury	$2–6 \times 10^{12}$	Unknown
Venus	$<8 \times 10^{10}$	—
Earth	8×10^{15}	10.5
Moon	$<1 \times 10^{9}$	—
Mars	$<1 \times 10^{11}$	—
Jupiter	1.5×10^{20}	9.5
Ganymede	1.4×10^{13}	5 ± 5
Saturn	4.6×10^{18}	0
Uranus	3.8×10^{17}	59
Neptune	1.8×10^{17}	47

One final source of planetary magnetic fields is natural remanent magnetization of solid materials. The most common means of acquiring such natural remanence is the cooling in an external magnetic field of magnetic material through one or more "blocking" temperatures below which the material can retain its acquired magnetization. Typical blocking temperatures are several hundred degrees Celsius and typical carriers of remanence are small particles of free iron metal and nickel and the iron oxide magnetite, Fe_3O_4. It is often possible to determine the direction and magnitude of the ancient magnetizing field from rocks containing natural remanence. Such studies have been crucial on the Earth for demonstrating that continents drift and the ocean floor spreads and that the terrestrial field periodically but irregularly reverses its direction. Lunar rocks also possess natural remanent magnetism, leading to the conjecture that the Moon once possessed its own dynamo-generated internal magnetic field that ceased operating over 3 billion years ago.

B. Charged Particle Motion in Electric and Magnetic Fields

In order to understand the structure and behavior of solar system magnetic and electric fields, it is necessary to understand how charged particles move in these fields, for, to a large extent, these electric and magnetic fields arise self-consistently from these same particles. As shown in the leftmost panel of Fig. 2, charged particles gyrate around magnetic field lines. The frequency of this motion is proportional to the strength of the magnetic field and the

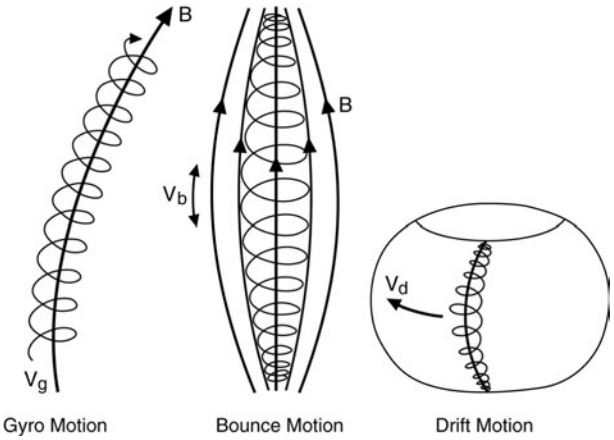


FIGURE 2 The three periodic motions of charged particles: gyro motion, bounce motion, and drift motion.

charge on the particle and inversely proportional to the mass. A proton in a 100-nT field gyrates around the field 1.5 times per second. An electron in a 100-nT field gyrates 2800 times per second. The direction of rotation of the charged particle depends on the sign of the charge. Positively charged particles gyrate in a left-handed sense, clockwise, viewed with the magnetic field pointing toward the observer. Electrons rotate in a right-handed sense. These circulating charged particles carry a current around the magnetic field in such a direction to reduce the magnetic field strength. The radius of the orbit of a charged particle, its gyroradius, is proportional to the velocity of the particle perpendicular to the field and the mass of the particle and inversely proportional to its charge and the magnetic field strength. The area enclosed by this orbit times the current due to the gyration of a particle (i.e., its charge times its gyrofrequency) is called the magnetic moment of the particle. It is equal to the energy associated with the motion of the particle perpendicular to the magnetic field divided by the magnetic field strength. If the variation in magnetic field strength is sufficiently slow in space or time, the magnetic moment of a particle is conserved. The magnetic moment is also called the first adiabatic invariant.

If magnetic field lines converge as they are shown to do in the middle panel of Fig. 2, the field strength as seen by the particle increases as the particle goes from the middle to either end of the field line. Because the magnetic moment of the particle is conserved, the perpendicular energy of the charged particle will increase until it is equal to the total energy of the particle. Because at the point where this occurs the particle has no forward velocity, and because the same forces are still acting on the particle that decelerated its forward motion, it reflects and starts moving away from this "mirror point." A planetary magnetic field generally has two points of convergence, one in the northern hemisphere and one in the south. Thus, particles will bounce back and forth trapped between mirror points. The bounce frequency is determined by how fast the particles are moving along the magnetic field line and by the length of the magnetic field. The parallel momentum of a charged particle is its mass times its velocity measured along the magnetic field. The integral of the parallel momentum over a bounce period (i.e., the momentum summed over each portion of the path) is conserved. This quantity is called the second adiabatic invariant. If mirror points of a particle move closer together, the particle must gain energy along the field due to this conservation principle. This process is called Fermi acceleration.

In nonuniform magnetic fields, charged particles drift perpendicular to the magnetic field. If the magnetic field strength decreases with altitude as it does in the terrestrial magnetic field, the gyroradius of a particle will be larger when it is farther from the Earth in its gyromotion than when it is closer. This leads to a net drift of the particle whose direction depends on the charge of the particle. In the Earth's magnetic field, which in the equatorial regions points northward, positively charged ions such as protons drift westward opposite to the Earth's rotation and electrons drift eastward. This process is referred to as gradient drift. If a particle is bouncing back and forth along a curved field line as shown in the rightmost panel of Fig. 2, it will experience a centrifugal force outward as its motion is deflected by the curved field. This has exactly the same effect as a field gradient, and electrons and protons drift in opposite directions. The drift path around a planetary magnetosphere is generally closed for most nonrelativistic particles in most planetary fields if the planetary field approximates that of a dipole. This drift requires from minutes to hours, compared to seconds to minutes for the bounce motion, and milliseconds to seconds for the gyromotion. When a charged particle completes its motion around a planetary magnetosphere the path traced out by the drifting bouncing particle is a roughly spheroidal shape with open ends. This so-called drift path or shell encloses a certain amount of magnetic flux, or equivalently a certain number of magnetic field lines. The magnetic flux enclosed by the drift path, or equivalently the magnetic flux through the open ends, is conserved when the magnetic field changes slowly on the time scale of the drift motion. This conserved amount of magnetic flux is referred to as the third adiabatic invariant. If the magnetic field of the Earth, for example, increased, then the radiation belts would move outward to conserve this invariant. The gradient and curvature drift of charged particles around the Earth's equator cause a net current to flow there, called the ring current. The ring current causes a depression in the strength of the

magnetic field observed on the surface of the Earth. The size of this depression is linearly related to the energy of the ring current particles. If the particles in the terrestrial trapped radiation belts possessed 2.8×10^{15} J of kinetic energy, there would be a 100-nT depression with the horizontal component of the Earth's surface field. A geomagnetic storm is such a period of enhanced ring current.

A charged particle is accelerated by an electric field. In the case of an electric field at right angles to a magnetic field, a gyrating particle will be accelerated for one half of its gyration and decelerated in the other half. When it is moving fastest, its gyroradius is largest and when it is moving slowest, its gyroradius is smallest. Thus, it moves farther in one half of its gyrorotation than the other. Protons and electrons gyrate in opposite directions about the field and also are accelerated on opposite halves of their gyration. The net effect is that electrons and protons drift in the same direction and at the same velocity perpendicular to both the magnetic and electric fields. In the Earth's magnetic field, which is northward in the equatorial regions, an electric field from dawn to dusk produces drift toward the Sun. Since electrons and ions drift together, there is no current associated with this drift.

C. The Physics of Plasmas

Thus far we have considered only single-particle motion. However, a gas of charged particles can exhibit collective behavior. The term plasma is usually restricted to a gas of charged particles in which the potential energy of a particle due to its nearest neighbor is much smaller than its kinetic energy. If we put a test charge in a plasma, it gathers a screening cloud of oppositely charged particles around it that tends to cancel the charge. In effect, the membership of a given particle in many screening clouds produces the collective behavior. Beyond some distance, called the Debye length, there is no observable effect of an individual charge. The Debye length is proportional to the square root of the ratio of the plasma temperature to its density. A plasma such as the solar wind with a temperature of 10 eV and a density of 10 per cubic centimeter has a Debye length of 740 cm. The number of particles in a cube with the dimensions of the Debye length is called the plasma parameter. The value of this parameter determines whether the ions and electrons can be treated as a plasma exhibiting collective behavior or as an ensemble of particles each exhibiting single-particle behavior. We can consider an electron–ion gas to be a plasma when the plasma parameter is much greater than 1. As illustrated in Fig. 3, in the Earth's ionosphere and the Sun's outer atmosphere (the corona) this number is about 10^5. In the Earth's magnetosphere and in the solar wind it is about 10^{10}.

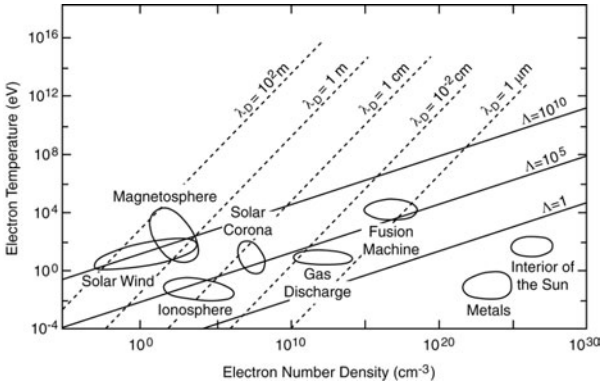


FIGURE 3 The Debye length λ_D and the plasma parameter Λ for plasmas of geophysical interest.

The characteristic frequency of a plasma at which it would oscillate if the ions and electrons were pulled apart and allowed to move back together (a collective plasma effect) is called the plasma frequency. It is equal to 9 kHz times the square root of the number of electrons per cubic centimeter. The maximum plasma frequency in the Earth's ionosphere is somewhat greater than about 10 MHz. The number of collisions per second in a fully ionized plasma is very roughly the plasma frequency divided by the plasma parameter. Thus, particles in plasmas can oscillate many, many times between collisions, and hence plasma processes are often referred to as collisionless processes. When the plasma parameter is large, charged particles move in almost straight lines. As the plasma parameter decreases, the individual interactions between charged particles become more important and large-angle deflections become more frequent. Eventually, for small plasma parameters, electrons become trapped in the potential wells of individual ions. This same effect occurs in metals and the interior of the Sun.

The key to understanding the behavior of the electric and magnetic fields in the solar system lies in understanding the behavior of plasmas and the various instabilities that transfer energy from one form to another in a plasma. The collective interactions allow us often to ignore the individual particle nature of a plasma and consider it to be an electrically conducting fluid. The laws governing the behavior of this fluid are Maxwell's equations, Ohm's law, and the conservation of mass and momentum. Use of these equations is called the magnetohydrodynamic, or MHD, approximation.

Plasmas in the solar system often find themselves in unstable situations in which the MHD equations would predict a rapid change of configuration. For example, there is the interchange instability in which entire magnetic flux tubes interchange position because by doing so they acquire a lower energy state. This might occur if a heavily

loaded magnetic flux tube lay on top of a lightly loaded magnetic flux tube in a gravitational field. This is analogous to the situation in which a heavy fluid sits on top of a lighter one. Another MHD instability of some importance is the Kelvin–Helmholtz or wind-over-water instability, in which surface waves on a boundary are induced by a shear in the flow velocity at the boundary. This instability is often invoked to explain magnetic pulsations in the Earth's magnetosphere. It is analogous to the mechanism for the formation of ocean waves.

Plasmas, because they are collisionless, can have different temperatures along and across magnetic field lines. However, too large a difference can be unstable. The fire hose instability arises when the thermal pressure along a field line exceeds the sum of the pressure across a field line and the magnetic pressure. When this occurs the magnetic field lines wiggle back and forth like a fire hose whose end is not being held. An instability also occurs when the thermal pressure across a field line greatly exceeds the pressure along it. This is known as the mirror instability and it creates pockets of denser plasma along the field lines.

Many of the various plasma instabilities cause fluctuations in the plasma: oscillations in number density, in electric and magnetic field strength and direction, in the velocity of the plasma, etc. Such oscillations can also be stimulated by processes external to the plasma being studied. There are three types of magnetohydrodynamic waves: fast, intermediate, and slow. Fast waves compress both the plasma and magnetic field. Intermediate waves bend the magnetic field lines but do not alter the density or magnetic field strength. Slow waves compress the field when they are decompressing the plasma and vice versa so that the total pressure, thermal plus magnetic, is nearly but not quite constant. Such waves are found throughout the solar system plasma. The velocity of an intermediate wave, also called the Alfven velocity, is proportional to the magnetic field strength and inversely proportional to the square root of the mass density. In a plasma of nine protons per cubic centimeter and a field strength of 6 nT, which is typical of the solar wind near Earth, the Alfven velocity is 44 km/sec. Fast and slow waves travel somewhat faster and slower than this velocity by a factor that depends on the temperature of the plasma and the direction of propagation relative to the magnetic field.

Fast and slow waves can steepen into shock waves if their amplitudes are sufficient. A shock wave is a thin discontinuity that travels faster than the normal wave velocity and causes irreversible changes in the plasma. Such a collisionless shock is analogous to the shock produced by a supersonic aircraft flying through a collisional gas, air. The process of shock formation is described as steepening because the passage of a fast or slow wave alters the plasma so that a following wave will travel faster and thus catch up with the first wave. Thus the trailing parts of a wave catch up with the leading part and a steep wavefront is formed. This process is similar to the steepening of ocean waves as they approach the shore. The steepening of shock waves in a plasma is limited by collective processes occurring in the plasma over sometimes rather short scale lengths, such as the ion gyroradius. Many of the processes occurring in collisionless shocks generate plasma waves, both electromagnetic waves that transmit energy and electrostatic waves that do not. These waves, oscillating at and above frequencies in the neighborhood of the gyrofrequencies of the ions and electrons, heat the plasma and equalize the pressures along and across the field. The thickness of collisionless shocks is often a fraction of the ion inertial length, which is the velocity of light divided by the ion plasma frequency (a factor of 43 less than the electron plasma frequency mentioned earlier if the ions are protons). Numerically, this is equal to 228 km divided by the square root of the number of protons per cm^3. In a typical solar wind plasma near the Earth a shock will be about 40 km thick.

Collisionless shocks are found in front of all the planets, forming standing bow waves much like the bow wave in front of a boat. Such a wave occurs because the solar wind must be deflected around each of the planetary obstacles. However, the velocity of the solar wind far exceeds the velocity at which the pressure needed to deflect the flow can propagate in the solar wind. The only means by which a planet can deflect the supersonic solar wind is to form a shock wave that slows down, heats, and deflects the flow.

Solar flare-initiated blast waves also cause shocks in the solar wind that are convected out past the planets. When they reach the Earth, such shock waves cause sudden compressions in the Earth's magnetic field that are observed by surface magnetometers. A sudden compression followed by an injection of energy into the Earth's ring current is called a sudden storm commencement, or SSC. Otherwise a sudden compression of the Earth's magnetic field is called a sudden impulse, or SI. Sometimes shock waves propagating toward the Sun, or reverse shocks, are carried outward by the very supersonic solar wind. These can cause negative sudden impulses in which the Earth's surface magnetic field suddenly decreases. It was the occurrence of the negative and positive sudden impulses in the Earth's magnetic field that originally led to the postulate that collisionless shocks could exist in the solar wind plasma. In ordinary gases, shocks require interparticle collisions to heat the gas across the shock front. In a collisionless plasma, this heating occurs both through oscillating magnetic and electric fields and through a steady-state

process by which a small fraction of the ions get reflected by the shock and thus attain a high thermal energy relative to the flowing solar wind.

D. Conductivity and Electric Field Sources

As noted above, in a plasma without collisions, an electric field perpendicular to a magnetic field causes a drift of both the electrons and the ions in the same direction and hence no electrical current. In the dense plasmas of planetary ionospheres and the solar photosphere, collisions either with other charged particles or with neutral particles occur frequently enough that the collisions modify the response of the plasma to an applied electric field. When collisions occur much more rapidly than either the electron or ion gyrofrequencies, the charged particles are no longer controlled by the presence of the magnetic field. The charged particles are then said to be unmagnetized and the electric field drives a current parallel to the electric field as in an ordinary conductor. The ions are unmagnetized at intermediate collision frequencies because of their lower gyrofrequency. They drift parallel to the electric field, while electrons, because their gyrofrequency is much higher, are magnetized and drift perpendicular to the magnetic field. Thus, there is a component of current carried by electrons perpendicular to the applied electric field and a component of current carried by ions parallel to the electric field. The ratio of proportionality between the current and the applied electric field is called the conductivity of the plasma. The conductivity perpendicular to the magnetic field and parallel to the electric field is called the Pederson conductivity. The conductivity perpendicular to both the magnetic and electric fields is called the Hall conductivity. The conductivity parallel to the magnetic field is called the direct or longitudinal conductivity. At some altitudes in the terrestrial ionosphere the Hall conductivity greatly exceeds the Pederson conductivity and electric current flows mainly perpendicular to the applied electric field.

Collisions couple the neutral and ionized atmospheres of planets, including the Earth. Motions of either one couple into the other. Neutral winds are driven by solar heating. At low altitudes in the terrestrial ionosphere where both ions and electrons are unmagnetized these collisions simply carry both the ions and electrons along with the neutral wind and cause a drift of the plasma. At high altitudes, the collisions cause the electrons and ions to drift in opposite directions perpendicular to the direction of the wind resulting in a current perpendicular to the wind associated with the differential flow of the ions and electrons. At intermediate altitudes the electrons are magnetized and the ions unmagnetized so that there is an electron current perpendicular to the wind and the magnetic field, and an ion current parallel to the wind. Thus, neutral winds as well as applied electric fields can drive currents. These currents cause variations in magnetic records taken on the surface of the Earth. Variations associated with the normal solar heating-driven convection of the ionosphere have been called S_Q variations because they are in phase with the Sun and are present on quiet days. Daily variations at geomagnetically active times are called S_D variations. Variations caused by the effects of lunar gravitation on atmospheric circulation at quiet times are called L_Q variations. The S_Q variations are caused by both solar gravitational and heating effects, whereas L_Q variations are due to gravitational tides only.

The electrons and ions that surround a planet and comprise its ionosphere can be set in motion from above as well as below. The expanding solar atmosphere, or solar wind, flows by the planets at an exceedingly high velocity, about 400 km/sec on average. Despite the fact that this solar wind is very tenuous, it can couple into a planetary ionosphere or magnetosphere by viscously dragging on the magnetopause which couples to the ionopause, the outer boundary of the ionosphere. In a magnetized ionosphere like the Earth's ionosphere, such motion is equivalent to an electric field as discussed above. Because the parallel conductivity along magnetic field lines is quite high, there is little potential drop along the field lines and the potential drop appears across field lines in the lower ionosphere, driving currents there in the complex manner discussed in the previous paragraph. In short, drag at high altitudes in a planetary magnetosphere can cause circulation in the ionosphere. This, in turn, leads to complex current patterns at low altitudes.

Because the solar wind is magnetized, there is an additional component to the drag on a planetary magnetosphere in addition to the normal particle and wave transfer of momentum across the boundary. When the interplanetary magnetic field has a component antiparallel to a planetary magnetic field where the solar wind and planetary magnetosphere first come into contact, the two fields can become linked in a process called reconnection. This process increases the drag on the magnetized planetary plasma, leads to a long magnetized tail behind the planet, and causes ionospheric flow across the polar cap, which returns at lower latitudes. This process of reconnection is believed to be the primary controlling mechanism for almost all of geomagnetic activity, except for sudden impulses in ground magnetograms that are associated with shocks in the solar wind passing the Earth.

Because the Earth rotates and the terrestrial magnetic field lines are good electrical conductors, the plasma on the field lines tends to rotate with the Earth. The electric field associated with this rotation is called the corotational electric field. The motion of plasma due to the drag of the

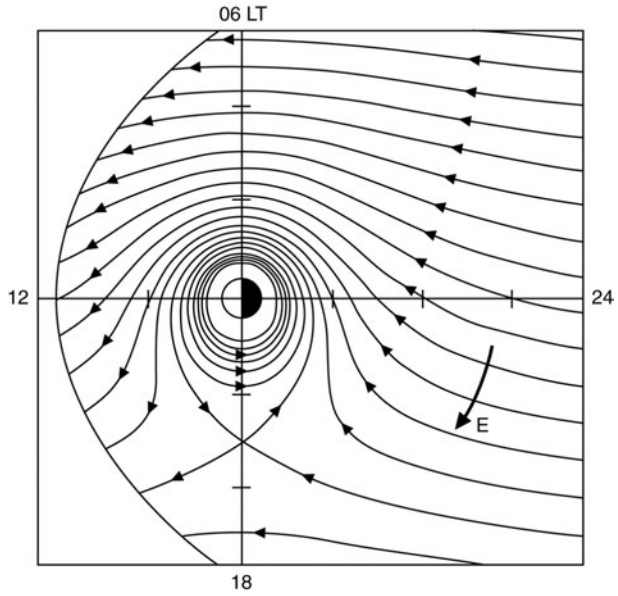


FIGURE 4 Drift paths of cold plasma in the Earth's magnetic field as a result of the superposition of a uniform electric field due to the interaction of the solar wind with the Earth and the corotation electric field. Because of the strong coupling of the magnetospheric and solar wind electric fields in this model, it is called the open magnetospheric model. The drift path flows from midnight (24 hr local time, LT), principally past the dawn meridian at 6 LT, and out through the magnetopause near noon, 12 LT. The electric field direction *E* is perpendicular to the streamlines.

solar wind, which at high latitudes is directed away from the Sun over the polar caps, is toward the Sun at lower latitudes. The electric field associated with this motion is called the convection electric field. The combination of the two fields produces a plasma circulation pattern in the Earth's equatorial magnetosphere as shown in Fig. 4. In this figure the large arrow labeled *E* represents the direction of the convection electric field. It is perpendicular to the magnetic field, which is out of the page and to the direction of flow, or convection, of the plasma, which is indicated by the streamlines labeled with smaller arrows. Most of these streamlines are open and carry low-energy plasma from the nightside of the magnetosphere to the dayside and out through its boundary. A subset of the streamlines in the inner magnetosphere is closed. In this region the plasma rotates with the Earth, both being supplied by and losing particles to the Earth's ionosphere along magnetic field lines.

II. THE SOLAR WIND

When during a solar eclipse the moon blocks all the light from the photosphere, light scattered from electrons in the solar corona can be seen. The density of the solar corona is very structured, and this structure is controlled by the solar magnetic field that changes significantly from month to month, passes through quiet phases every 11 years, and reverses polarity just after every activity maximum producing a 22-year magnetic cycle. The corona is also hot, so hot that it expands supersonically into interplanetary space. It reaches a velocity of typically 440 km/sec well before reaching Mercury and continues at this velocity well beyond Pluto. The density of the solar wind falls off as the inverse square of the heliospheric distance, so that at the orbit of the Earth, one astronomical unit (AU) from the Sun, there are about seven electrons and seven protons in every cubic centimeter with a temperature of about 100,000 K. Electrons and ions do not collide under these conditions and therefore are frozen to the magnetic field line on which they began their journey from the Sun. The energy density in the flowing solar wind is much greater than that in the magnetic field so the magnetic field does not act as a brake on the flow, rather the plasma drags out the magnetic field. The thermal velocity of the ions is much less than the radial bulk flow of the plasma, but not so for the much lighter electrons, which can communicate with the Sun along the stretched-out spiral magnetic field. Thus, one should visualize the electrons as streaming (back and forth) along the magnetic field line while the protons move radially outward with no contact with the Sun, dragging the field line with them. Time variations on the Sun may change the density of electrons on the field line at the surface of the Sun, but the electric potential due to the existence of an excess charge at any point adjusts the velocities of the electrons just the right amount to maintain charge neutrality. The thermal velocities of the electrons are not distributed according to a Maxwellian distribution, nor are they the same along and across the field. This complicates the radial variation of the effective temperature of the electrons and the angular distributions.

Much like the Earth, the Sun's magnetic field has a north pole and a south pole whose locations switch periodically, about every 11 years. There is also a magnetic equator. Unlike the Earth, the magnetic equator tilts significantly over the course of the 11-year sunspot cycle and unlike the Earth, the strength of the higher order moments of the magnetic field can be very strong. At the orbit of the Earth much of this complexity is seldom seen. What is important is the polarity of the dipole moment and the tilt of the magnetic equator or, as it is more commonly called, the heliospheric current sheet. As the sun rotates relative to the Earth it carries this current sheet past the Earth in a 27-day cycle producing a characteristic in/out pattern of the magnetic field as the Earth moves above and below the current sheet. This pattern is referred to as the sector structure of the interplanetary magnetic field. Since the magnetic field is connected to the Sun on one end, it

forms a spiral pattern as it expands. The strength of the radial component of the magnetic field decreases as r^{-2} and of the tangential components as r^{-1}, resulting in a magnetic field spiral that becomes tighter and tighter as it convects outward. At the orbit of the Earth the angle of the magnetic field to the radial direction from the Sun is about 45°.

The solar wind velocity varies with heliomagnetic latitude, being about 800 km/sec above about 30°. When the magnetic equator is tilted with respect to the rotational equator, the fast solar wind regions will catch up with the preceding slow solar wind. This collision causes a pileup in the collisionless solar wind because one magnetized plasma region cannot penetrate another except under special circumstances. This collision is called a stream–stream interaction and it compresses the field strength and the plasma density and the causes deflections to the east and west. These regions in turn have greater influence on the magnetosphere when they arrive at 1 AU due to these larger densities and magnetic fields.

The most effective disturbances from the Sun are those associated with coronal mass ejections (CMEs), in which large segments of the corona come flying off accompanied by a strong magnetic field. This fast palsma flow with its strong steady fields is very effective at producing a disturbed terrestrial magnetosphere. These disturbances in the solar wind have been called interplanetary coronal mass ejections (ICMEs) to distinguish them from the causative disturbance whose properties we do not fully understand. ICMEs often drive a leading bow shock in the solar wind. They also often contain multiple magnetic flux ropes or magnetic clouds that are twisted tubes of strong magnetic flux, containing perhaps as much as 50 terawebers (TWb) of magnetic flux. It is the interaction of these twisted tubes of magnetic flux that cause the largest geomagnetic disturbances or storms.

III. PLASMA INTERACTIONS WITH UNMAGNETIZED ATMOSPHERELESS BODIES

The simplest interaction of a plasma with a solar system body is that with an unmagnetized atmosphereless body, as in the interaction of the solar wind with the Earth's moon. The electrons and ions in the solar wind strike the lunar surface and are absorbed, leaving information about the solar wind energy implanted in the lunar soil but undergoing no esoteric plasma processes. This process leaves a cavity behind the Moon. The solar wind attempts to close behind the Moon to fill in this cavity. As illustrated by Fig. 5, the closure depends on the direction of the interplanetary magnetic field relative to the direction of the

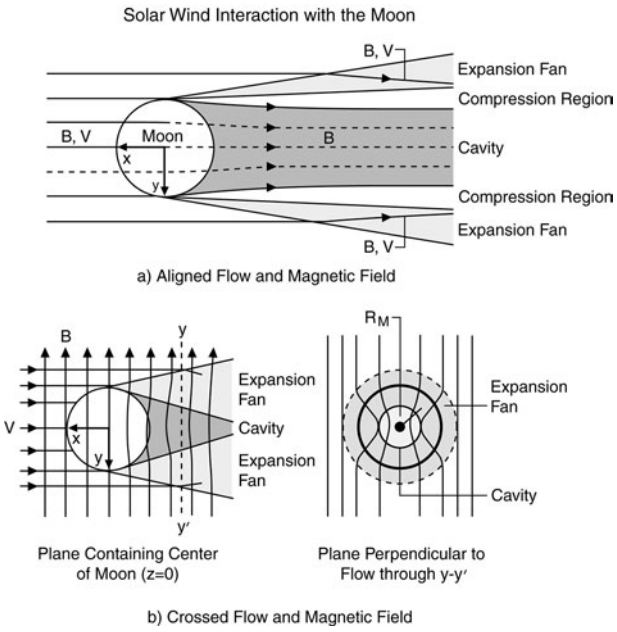


FIGURE 5 Interaction of the solar wind with the Moon for (a) flow *V* aligned with magnetic field *B* and (b) crossed flow and magnetic field. Left: Plane containing the center of the Moon ($z = 0$). Right: Plane perpendicular to flow through y–y'.

solar wind flow. If the magnetic field is aligned with the flow, the cavity behind the Moon is filled with magnetic flux and the cavity closes only slightly, so that pressure balance with the solar wind pressure is maintained. If the magnetic field is perpendicular to the flow, then the solar wind closes slowly behind the Moon as plasma flows along field lines at the thermal speed. The thermal speed is the average random velocity of the particles relative to the drift or bulk velocity of the flow.

The Moon is not completely unmagnetized. It possesses remanent magnetization from an earlier epoch. Some regions are coherently magnetized over hundreds of kilometers and sufficiently so that they have enough magnetic pressure to deflect the solar wind when the solar wind flow is tangential to the lunar surface in the region known as the solar wind limbs, or terminator. This deflection causes a disturbance to be launched into the solar wind, but it is not strong enough to form a shock. These disturbances have been called limb compressions.

The Moon has an insulating outer shell, but its interior, being much hotter, is quite electrically conducting. The size and conductivity of this highly conducting region are such that it would take at least many days and perhaps hundreds of years for a magnetic field to diffuse from outside the core to the interior of the core. Thus magnetic observations can be used to probe the interior lunar electrical conductivity. One way to do this is to use measurements obtained when the Moon is in the near-vacuum conditions

of the geomagnetic tail lobes. A satellite, such as one of the Apollo 15 and 16 subsatellites, flying in low-altitude lunar orbit can measure the distortion of the field caused by its exclusion from the lunar core. A second technique is to measure the frequency spectrum of magnetic fluctuations in the solar wind on a spacecraft and then measure the frequency spectrum of magnetic fluctuations seen on the lunar surface as was done during the Apollo program. The alteration of the frequency spectrum can be used to determine the conductivity profile of the Moon. In particular, this technique can sound the conductivity profile in the cooler outer layers of the Moon. Its sensitivity is limited by the accuracy of the intercalibration of the two instruments used.

Two of the moons of Jupiter, Europa and Callisto, also appear to have an interior electrically conducting layer, but in the cases of these two bodies the layer is thought to consist of salty water rather than a metallic or molten core. At both bodies the time-varying magnetic field associated with the rotation of its tilted dipolar magnetic moment is nearly completely excluded from the interior of the moon indicating that the conducting layer is close to the surface. The thickness of the subsurface oceans of Europa and Callisto is thought to be of the order of 100 km.

Plasmas interact with the moons of Jupiter and Saturn. Except for Titan and to a lesser extent Io, these moons have almost no atmospheres. Energetic charged particles from the radiation belts of these planets impact the surfaces of these Moons and sputter atoms from them. These atoms then become ionized and join the plasma surrounding the planets. Such sputtering also occurs in planetary rings. The collision of energetic particles with moons and rings is a significant loss process for the radiation belts of Saturn and to a lesser extent the radiation belts of Jupiter.

Dust particles, both those in rings and those not in the rings, can become electrically charged. For small dust particles, the charge can significantly alter their motions because the particles will feel the forces of the planetary electric and magnetic fields as well as of gravity. For instance, a planetary electric field is typically such as to enforce corotation of the plasma in the planetary magnetosphere with the rotation of the interior of the planet, whereas the orbital velocity of uncharged particles varies with radial distance. So, too, will interplanetary and cometary dust be affected by the motional electric field of the solar wind.

Finally, the interaction of the solar wind with an asteroid should resemble the interaction with the Moon except that perhaps the cavity will not be as well defined because the asteroids are all smaller than the Moon and hence gyroradius effects become more important. If the asteroid outgasses, as might an asteroid that is a nearly extinct comet, then its interaction region would be quite large and its interaction processes quite different from those described above. We defer such discussions to Section V.

IV. PLASMA INTERACTIONS WITH MAGNETIZED BODIES

There are at least seven strongly magnetized bodies in the solar system, Mercury, the Earth, Jupiter, Saturn, Uranus, Neptune, and the moon Ganymede. Each of these bodies provides us with a different aspect of the interaction of a flowing plasma with a magnetic field. Mercury gives a small magnetosphere, or magnetic cavity, both in absolute terms and relative to the size of the planet. Mercury also has no atmosphere or ionosphere. The Earth has a sizeable magnetosphere as well as a well-developed atmosphere and ionosphere. Jupiter is a rapidly rotating planet with an immense magnetosphere and a strong source of plasma deep in the magnetosphere. Saturn has a smaller but also rapidly rotating magnetosphere. However, its plasma sources are not as strong as Jupiter's sources. Uranus has a magnetic dipole axis that is at an angle of 60° to its spin axis, but, since its rotation axis is presently nearly aligned with the solar wind flow, it has an almost Earth-like interaction. Neptune's moment, tilted at 47° to its rotation axis, which is almost perpendicular to the solar wind flow, undergoes large changes in its angle of attack to the solar wind in the course of each day. Finally, Ganymede, the solar system's largest moon, has a magnetic moment large enough to create its own magnetosphere inside that of Jupiter.

A. Solar Wind Interaction with Mercury

Mercury is the smallest of the terrestrial planets, with a radius of 2439 km, intermediate between the Earth's moon and Mars in size. It rotates more slowly than the Moon, rotating with a period of 59 days compared to the Moon's 28 days. It is heavily cratered like the moon but differs from the Moon in its lack of synchronicity of rotational and orbital periods and in its density, which is 5.4 gm/cm^3 compared to the Moon's density of 3.3 g/cm^3. The high density indicates that Mercury has a significant iron core. This core apparently is sufficient to sustain an active dynamo despite the small size and slow rotation of the planet. Mariner 10 passed through the nightside magnetosphere of Mercury twice, in March 1974 and 1975, and detected a magnetic field arising from a planetary dipole magnetic moment of strength $4 \pm 2 \times 10^{22}$ G-cm^3. Figure 6 shows a sketch of the solar wind interaction with Mercury and the field lines in the Mercury magnetosphere. The nose of the magnetopause is only about 0.35 Mercury radii,

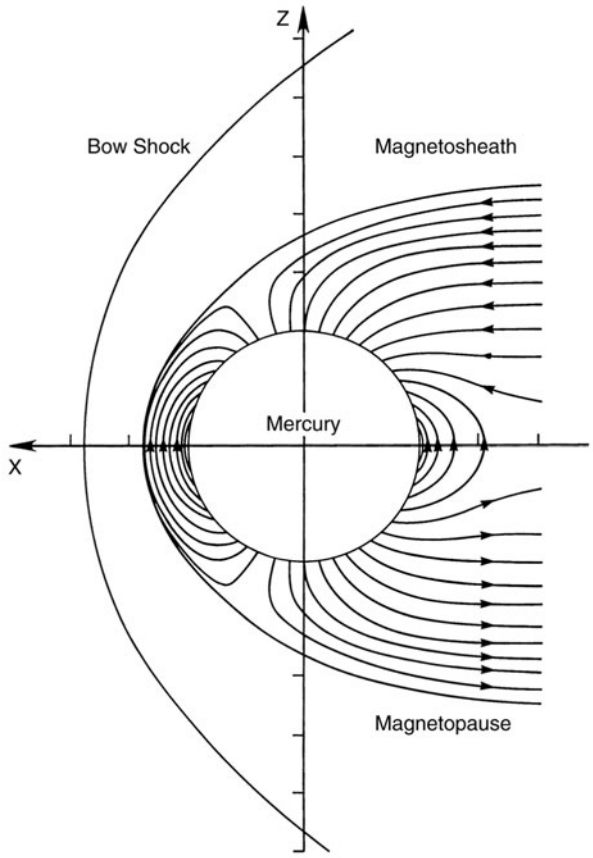


FIGURE 6 The solar wind interaction with Mercury illustrating the large fraction of Mercury's magnetosphere occupied by the planet.

R_m, above the surface of Mercury. The bow shock is at about $0.9R_m$ above the surface at the subsolar point. The magnetic field strength on the surface of Mercury is 300–500 nT. Its instantaneous value depends on variations in the solar wind pressure relatively more than the magnetic field at the surface of the Earth depends on these variations. Variations observed in the magnetic field during the passages of Mariner 10 past the planet have been interpreted in terms of a dynamic magnetosphere controlled to a large extent by internal processes, but there are few observations on which to judge. The magnetopause and bow shock seem similar in basic properties and structure to these same terrestrial boundaries.

B. Solar Wind Interaction with the Earth

The Earth, like Mercury, presents a magnetic obstacle to the solar wind. The Earth's magnetosphere, though, is about 20 times the size of Mercury's magnetosphere, being about 200,000 km wide at the dawn–dusk or terminator plane. Furthermore, at the feet of the field lines of the terrestrial magnetosphere there is a dynamically significant ionosphere. In front of this magnetosphere stands a bow shock wave much like that of Mercury, both in relative location and in microstructure. The closest approach of the bow shock to the center of the Earth under typical solar wind conditions is about 92,000 km or about 14.5 Earth radii. Under similar conditions the boundary of the magnetosphere (i.e., the magnetopause) lies at about 70,000 km or 10.9 Earth radii.

The Earth's dipole magnetic moment is 8×10^{15} T-m^3, which corresponds to a surface field of about 31,000 nT at the equator of the Earth and twice this at the poles. There are contributions from higher order terms such as the quadrupole and octupole terms. The relative importance of these higher order terms diminishes with altitude so that for most purposes we can consider the field in the inner magnetosphere to be dipolar. The Earth's internal dynamo is not steady. The most noticeable surface field change is a slow westward drift, but on much longer scales, millions of years, the field reverses. These reversals provide a useful clock for geophysical studies. Their existence also suggests that the Earth was not always so well shielded from the solar wind as it is today.

Figure 7 shows a noon–midnight meridian cross section of the terrestrial magnetosphere. The solar wind, after passing through the bow shock, flows through a region known as the magnetosheath and around the magnetopause. The electrons and protons in the magnetosheath near the magnetopause are reflected by the Earth's magnetic field. In this encounter with the Earth's magnetic field they turn 180° about the field line and escape from the magnetopause. The current associated with this gyroreflection self-consistently generates the proper electric current to bound the Earth's magnetic field under ideal conditions. Thus, plasma cannot directly enter the Earth's magnetic field in the subsolar region. The magnetic field inside the boundary has a pressure equal to the

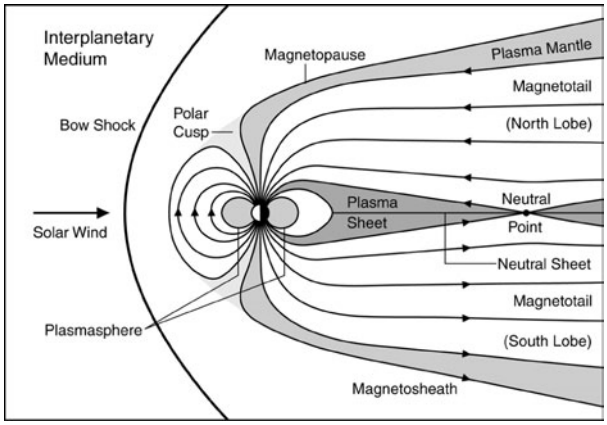


FIGURE 7 The interaction of the solar wind with the Earth.

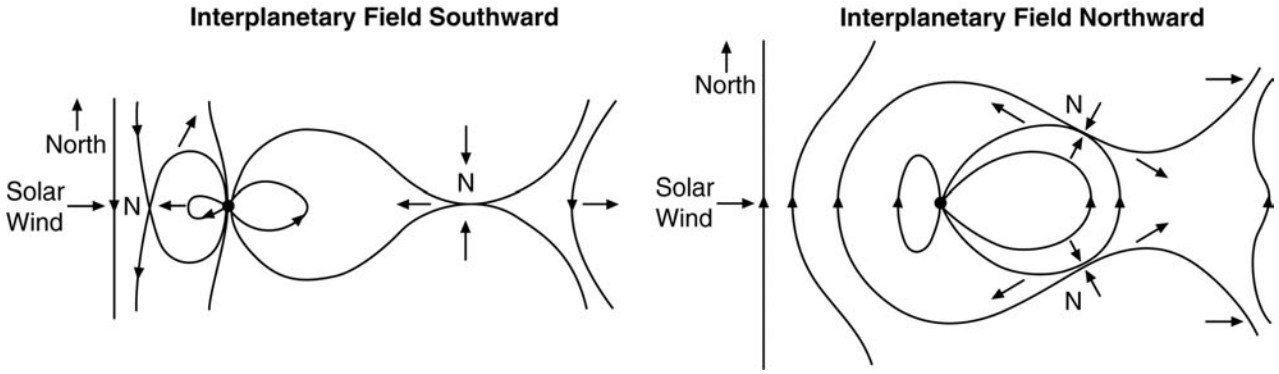


FIGURE 8 The reconnection process between interplanetary magnetic field lines and terrestrial magnetic field lines for situations in which the interplanetary magnetic field is southward or northward. The points labeled N are 'neutral' points in which the magnetic field strength is zero and across which the magnetic field threading the boundary reverses sign.

dynamic pressure of the solar wind outside the magnetosphere.

Behind the Earth, a long magnetic tail stretches for perhaps 1000 Earth radii or more. If there were no viscosity in the solar wind interaction, the magnetopause shape would be determined only by normal stresses perpendicular to the magnetopause, and the resulting shape would be a teardrop. Viscosity is presently thought to be supplied in part by the Kelvin–Helmholtz instability discussed above and in part by the process of reconnection discussed below. As the flow in the magnetosheath passes above and below the polar regions, it passes a weak point in the magnetospheric field that marks the beginning of the tail. The two regions, one in the north and one in the south, are called the polar cusps. In these regions, solar wind plasma can reach all the way down to the ionosphere. Plasma can enter the magnetosphere here and form a boundary layer that has been called the plasma mantle. In a reconnecting magnetosphere the plasma can drift across the tail to a null field point, called the X-point, where it is accelerated. There, some of it enters the closed field region of the magnetosphere. The process of reconnection effectively couples the electric field of the solar wind into the electric field of the magnetosphere. Since motion of magnetized conductors is equivalent to electric fields, we may also view this as the motion of the solar wind causing motion of the terrestrial plasma. However, this coupling of the two magnetized fluids cannot occur without breakdown of the very high conductivity of the fluids at least in some limited region.

The net result of this process is flow in the outer magnetosphere, but the deep interior of the magnetosphere is relatively unaffected by these flows and continues to corotate with the Earth. In this interior region, the ionospheric plasma, which is continually upwelling from the dayside ionosphere, can build up in density to a level at which the upwelling is matched by an equal downward flow. This high-density region of cold ionospheric plasma is called the plasmasphere. In this region the magnetic field is quite dipolar in character and the electric field is simply that necessary to make the plasma rotate uniformly.

The strongest coupling between the solar wind and the magnetosphere is through the process known as reconnection. This process can be steady state, slowly varying, or highly temporally varying. First we examine the case of steady-state reconnection. Then, we examine how variations in the reconnection rate lead to a dynamic magnetosphere and finally we examine the effects of temporally and spatially patchy reconnection. Figure 8 shows the field lines in an idealized solar wind–magnetosphere interaction. In the top panel, the interplanetary, or solar wind, field lines are in the direction opposite those in the Earth's magnetosphere. Due to a breakdown in electrical conductivity at the "nose" of the magnetosphere, the interplanetary and planetary magnetic fields join at this point. The flow of the solar wind carries the ends of these joined fields over and under the magnetosphere, pulling the magnetospheric field lines over the poles. These field lines sink in the tail until they meet in the center of the tail at the X-point where they reconnect once more. Here they form a closed field line, which touches the Earth at both ends, and also an interplanetary field line that does not touch the Earth at all. This interplanetary field line flows away from the Earth. The newly closed field line flows toward the Earth, around the Earth out of the plane of the figure, and then joins up with a new interplanetary magnetic field line.

When the interplanetary magnetic field is northward, parallel to the Earth's magnetic field, as in the bottom panel, reconnection apparently can still occur at high latitudes behind the polar cusps. This reconnection removes flux from the magnetotail and adds it to the dayside

magnetosphere if the same field line connects to both the North and South of the magnetotail. Otherwise, the magnetosphere is simply stirred by this process but no net flux transfer occurs.

The processes sketched here are steady state. If, instead, as occurs in practice, the rates of reconnection and hence flux transfer from one region to another vary and vary differently, flux can build up in one region or another. The sequence that frequently occurs is a sudden increase in the dayside reconnection rate with a southward field in the solar wind. The magnetic flux in the magnetotail increases until the reconnection rate in the tail increases to remove the flux from the tail. This latter rate of reconnection can exceed the rate that built up the flux enhancement in the tail. This generates rapid flows and high electric fields as plasma is accelerated both toward and away from the Earth.

The accelerated plasma, in addition to filling the radiation belts and adding energy to the ring current flowing in the magnetospheric equator, also powers the auroral displays. Aurora occur when energetic particles from the magnetosphere collide with neutral atoms and molecules, putting their electronic shells into an excited state from which they later decay by giving off light. During these times, flows in the outer magnetosphere can well exceed the flows possible in the ionosphere at low altitudes. This can occur only if the magnetic field line has less than infinite conductivity, and therefore electric fields can appear parallel to the magnetic field lines. In such situations, particles bouncing on flux tubes become accelerated in the "parallel" electric field and cause further auroral displays as they excite atmospheric atoms by collision. The overall sequence of events in the Earth's magnetosphere and ionosphere at such times is usually referred to as a magnetospheric substorm.

Even this complex process is an oversimplification. In practice, the reconnection process appears to be very unsteady. This unsteadiness appears to lead to the formation of magnetic islands on both the dayside and the nightside. This unsteadiness also leads to oscillations in the magnetospheric plasma inside the magnetosphere, adding to those oscillations that may be associated with the Kelvin–Helmholtz instability. Yet another source of waves is the region of space upstream of the Earth's bow shock yet connected to it by field lines. Beams of particles escaping from the bow shock generate waves. These waves are blown back toward the Earth by the supersonic solar wind. If the waves are generated in just the right region, they can be carried to the magnetopause and also add to the oscillations in the magnetosphere. Thus the magnetospheric field lines are almost always vibrating to some degree. Some of these vibrations can occur at resonances of the field line and large standing waves build up. Often the oscillating electric field associated with these waves far exceeds the background or steady electric field value. However, typically the magnetic amplitude is at most a few percent of the background field.

The presence of these oscillating electric and magnetic fields has very significant consequences for the trapped energetic charged particles in the Earth's radiation belts. In a dipolar field such as the Earth's, charged particles have three separable periodic components of motion: gyration around the magnetic field, bounce motion back and forth along the field, and drift around the Earth. As discussed earlier, in Section I.B, these three components of motion are associated with three conserved quantities or adiabatic invariants. These quantities remain unchanged unless the assumptions about the constancy of the fields in which the particles move are violated. The lowest frequency oscillations in the magnetosphere, with periods of many minutes to hours, can violate the third adiabatic invariant, that associated with the drift of particles around the Earth causing them to diffuse inward or outward. Fluctuations with periods of seconds to minutes can resonate with the bounce motion causing particles to mirror at different points. Fluctuations at periods of milliseconds to seconds can resonate with the gyro motion of particles causing particles to have different helical paths along the field and to enter the atmosphere rather than bounce back along the magnetic field line.

A very important set of these resonant wave–particle interactions are known as wave–particle instabilities. If waves at the resonant frequency do not exist, they may be spontaneously generated by the plasma if the net effect of the waves on the particles includes a transfer of energy to the waves. For example, if an electron traveling along a magnetic field with some energy across the field and some energy along the field met a right-handed electromagnetic wave head on, it would resonate with the wave if the particle velocity along the magnetic field caused the wave to appear to the particle to be oscillating at its gyrofrequency. This resonance would cause energy to be transferred from the parallel motion to the perpendicular motion of the electron or vice versa depending on the phase of the interaction. This resonant oscillation also would transfer energy from the wave to the particle in the former case and from the particle to the wave in the latter. If, for a given resonant parallel electron velocity, there are more electrons with energy predominantly perpendicular to the field than along the magnetic field, this random resonance will cause a net diffusion from the perpendicular energy state to the parallel one. Since this liberates energy from the electron, the wave gains energy. This process is but one example of a wave–particle instability. There are many such instabilities in the magnetospheric plasma causing both electromagnetic waves and

electrostatic waves. These instabilities help the magnetosphere regulate itself and return to an equilibrium configuration when disturbed.

C. Solar Wind Interaction with Jupiter

The magnetosphere of Jupiter is immense. It could easily contain the Sun and its corona. If one could see it in the night sky, it would appear to be much larger than the full Moon. It also rotates very rapidly, because the planet rotates rapidly and there is good electrical connection between the planet and the magnetosphere. Furthermore, the magnetosphere is filled with heavy ions that appear to come mainly from the satellite Io. These characteristics combine to provide Jupiter with a very interesting and different magnetosphere than that of the Earth.

The reason for the enormity of the Jovian magnetosphere is that the planet's intrinsic magnetic moment of 1.5×10^{20} T-m^3 is over 16,000 times greater than that of Earth, and that the solar wind that confines the magnetospheric cavity is over 25 times weaker. The resulting average distance to the magnetopause at the subsolar point is 70 planetary radii. The bow shock at the subsolar point is on average 85 Jovian radii. The surface magnetic field of Jupiter has a greater contribution from higher order terms (quadrupole, octupole, etc.) than the surface field of the Earth. This is at least in part due to the fact that the conducting core of Jupiter, in which the dynamo currents flow, is relatively closer to the surface of Jupiter than is the Earth's core to the surface of the Earth.

Well inside the Jovian magnetosphere, at 5.9 Jovian radii, orbits the moon Io with its sulfurous volcanoes. These volcanoes release into the Jovian magnetosphere sulfur monoxide, sulfur dioxide, sodium, and other gases, which in turn are ionized and become part of the trapped plasma of the Jovian magnetosphere. Sputtering from Io, the other satellites, and the Jovian rings also adds atoms and ultimately plasma to the Jovian magnetosphere. Because the Jovian field lines are good electrical conductors, the plasma is forced to corotate with the solid body of the planet over much of the magnetosphere. This acceleration is accomplished through an electrical current system closing through the Jovian ionosphere and the magnetospheric plasma joined along field lines. This current system also flows most strongly near Io as Jupiter attempts to force the newly created ions into corotation. These processes are thought to generate large potential drops along the field lines in the vicinity of Io. These electrical potential drops can accelerate charged particles to high energies and may be responsible for some of the radio emissions from Jupiter at decametric (10 m) wavelengths. At decimetric wavelengths (10 cm) radio emissions are due to synchrotron radiation from the intense fluxes of relativistic electrons near the planet. Auroral emissions are also observed at Jupiter and are thought to be caused in much the same way as terrestrial aurora.

The centrifugal force on the plasma in the Jovian magnetosphere far exceeds the gravitational force over much of the magnetosphere. On the front side of the magnetosphere the solar wind opposes this centrifugal force and near-static equilibrium is reached. If too much mass is added to flux tubes, then they can interchange their positions (via the interchange instability) with lighter flux tubes further out in the magnetosphere, but generally the flow corotates azimuthally with the planet.

As illustrated in Fig. 9, one major effect of these rapidly rotating mass-loaded magnetic field lines is that of stretching the magnetospheric cavity from the more common spherical shape to a more disklike shape. This effect can be seen in observations of the magnetic field and of the location of the bow shock. The streamlined shape of the disk-shaped magnetosphere allows the shock to stand closer to the magnetopause.

The plasma added to the magnetosphere at Io at an average rate of perhaps 500 kg/sec moves slowly outward at a rate of only meters per second at first but reaches a velocity of over 40 km/sec at a distance of about 50 Jovian radii. Thus many months are required for the transport of plasma to the distant magnetosphere where it can be lost down the tail, but most of this time is spent traversing the inner magnetosphere. Since in a collisionless plasma the

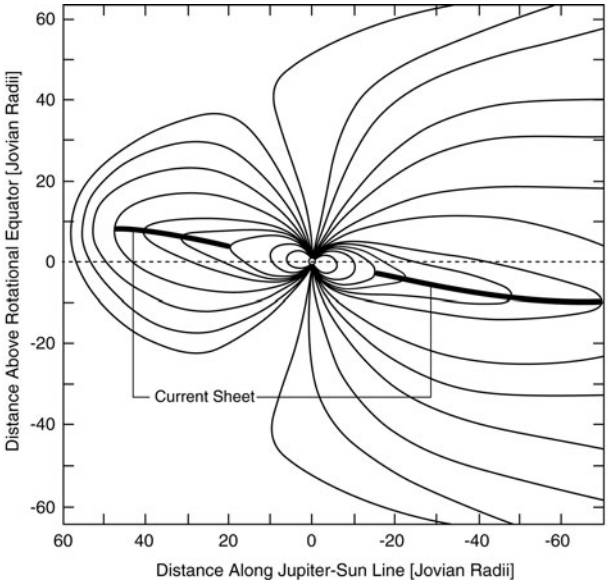


FIGURE 9 The magnetic field lines in the noon–midnight meridian of the Jovian magnetosphere showing the current sheet associated with centrifugal force exerted by the plasma added to the magnetosphere by the moon Io.

charged particles are constrained to stay on the same magnetic field lines in the absence of parallel electric fields, the outward motion of the plasma that is needed to maintain a steady-state density profile also transports outward the magnetic flux. Since Jupiter's internal dynamo is producing a rather constant magnetic flux, the outward plasma transport would soon deplete the magnetic flux of the planet if there were no way to separate the plasma from the magnetic field. Two mechanisms can do this: particle scattering that causes some particles to move parallel to the magnetic field and enter the atmosphere, and reconnection of oppositely directed magnetic fields across the magnetodisk that forms magnetized islands of ions with no net magnetic flux. Substorms, similar to those in the Earth's magnetotail have been reported. These liberate a similar amount of magnetic flux from the ions as was loaded with plasma at Io, about 80,000 Wb/sec on average. Once emptied of their plasma these magnetic flux tubes move buoyantly back to the interior of the Jovian magnetosphere much like a stream of air bubbles rising from the bottom of a sealed jar with a small hole in the bottom.

The size of the Jovian tail is immense. It is about 200 R_J (14 million km) across or more than 40 times the width of the Earth's tail. It is at least 4 AU long, extending all the way to Saturn.

D. Solar Wind Interaction with Saturn

The magnetic moment of Saturn is a factor of 32 less than that of Jupiter but it is immersed in a solar wind whose pressure is a factor of 4 less than that at Jupiter. The net result is a Saturn magnetosphere that, while it is only one fourth of the size of that of Jupiter, still dwarfs the magnetosphere of the Earth. Saturn's magnetosphere seems to be less inflated than that of Jupiter. The intrinsic magnetic field of Saturn is highly unusual. The surface magnetic field strength at the equator is 0.21 μT, not much different than that of the Earth. However, the magnetic moment of Saturn is almost perfectly aligned with the rotation axis, whereas the magnetic dipole moments of the Earth and Jupiter are tilted by about 10° to the rotational axes of the planets. The sizeable tilt of planetary dipole magnetic fields is thought to be essential to the dynamo process. Thus, the alignment of the Saturn magnetic moment is quite puzzling. The contribution to the surface field by the higher order moments is much less than that at Jupiter indicating that the depth at which the dynamo is acting at Saturn is greater than at Jupiter.

Satrun has many small and intermediate-sized moons and a well-developed ring system. These bodies are bombarded by the radiation belt particles, mass is sputtered from their surfaces, and mass is added to the magnetic field lines. The Saturn ionosphere, like the Jovian ionosphere, attempts to accelerate this plasma to corotational velocities. However, the mass-loading rates in the inner magnetosphere are not as large as at Jupiter and little distortion of the magnetosphere results. The one large moon of Saturn, Titan, orbits at the outer edge of the magnetosphere at 20 Saturn radii. Its dense atmosphere does strongly interact with the corotating magnetospheric plasma. However, the resulting distortion of the overall shape of the magnetosphere is much less than Io's effect on the Jovian magnetosphere. Again, this is evident from the direction of the magnetic field in the magnetosphere of Saturn and the location of the bow shock relative to the magnetopause.

The radiation belts of Saturn are much more benign than those of Jupiter and the radio emissions are much less intense. The rings absorb the charged particles as they diffuse radially in toward the planet. Because it is the particles in the innermost part of a planetary magnetosphere that are most energetic, the rings play a significant role in reducing the particle and radio flux. Ring particles can become electrically charged. It is thought that some of the exotic behavior of the rings such as the appearance of radial spokes in the rings may be the result of the effects of Saturn's magnetic and electric fields on these charged dust particles. The Cassini spacecraft launched in 1997 is scheduled to orbit Saturn in 2004 and extend our understanding presently based on three Pioneer and Voyager flybys.

E. Solar Wind Interaction with Uranus and Neptune

Voyager 2 flew through the magnetospheres of Uranus and Neptune in January 1986 and August 1989, respectively. The magnetic moment of Uranus is a factor of 12 smaller than that of Saturn, and the best-fit offset dipole field is tilted a surprising 59° to the rotation axis. The magnetic moment of Neptune is a factor of two smaller than that of Uranus and tilted at 47° to the rotation axis. The radiation belts of both planets are quite benign. Otherwise, the magnetospheres are quite terrestrial in configuration, with a standing, very high Mach number, bow shock in front and a long magnetic tail in the antisolar direction.

F. Interacting Magnetospheres of Jupiter and Ganymede

The Ganymede magnetosphere shown in Fig. 10 differs greatly from that of the Earth shown in Fig. 7. First, the velocity of the Jovian magnetospheric plasma past Ganymede is slower than that of the compressional wave that is required to deflect the flow around Ganymede. Thus the compressional wave can run far ahead of Ganymede so that the flow is deflected gradually around Ganymede and

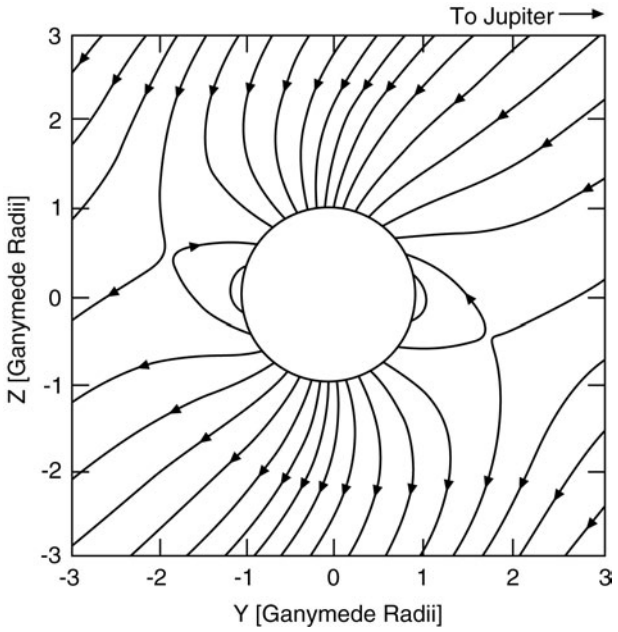


FIGURE 10 Ganymede magnetosphere shown in the magnetic meridian plane of Jupiter's magnetosphere.

no bow shock is formed. Second, the external magnetic conditions are relatively constant so that the Ganymede magnetosphere is quite steady with no substorm-type processes. Third, the strong external field limits the magnetosphere to a nearly cylindrically symmetric tube that wobbles with the nodding of the external Jovian field as the Jovian tilted dipole rotates. Finally, there is no cold plasmasphere in the inner part of the Ganymede magnetosphere. The ionosphere and slow rotation compared to the transport velocity induced by the Jovian magnetosphere are just too weak to produce such a feature. Nevertheless, it has some of the character of the terrestrial magnetosphere: a polar cap whose magnetic field lines connect to the external, flowing plasma and a region of closed field lines that intersect the surface of the body on both ends.

V. PLASMA INTERACTIONS WITH IONOSPHERES

A planetary ionosphere is the ionized upper atmosphere of a planet or moon, usually caused by solar ultraviolet radiation, but often impact ionization caused by charged particles plays an important role. If a body has an ionosphere, it also has a neutral atmosphere. The plasma also can interact directly with the neutral atmosphere. Thus, it is often difficult to separate those effects due to the interaction of a magnetized plasma with a conductor (the ionosphere) and those due to charged particle–neutral particle interactions. This is particularly true for the interaction of the moon Io and the Jovian magnetosphere. Nevertheless, in the next two sections we attempt just that. In this section, we discuss the interaction of the solar wind with Venus and then with Mars, both of which are probably dominantly, but not exclusively, controlled by ionospheric behavior.

A. Solar Wind Interaction with Venus

The planet Venus has been visited by many spacecraft including the long-lived Pioneer Venus orbiter, one of whose objectives was to study how the solar wind interacted with the Venus ionosphere. Figure 11 shows schematically the interaction of the solar wind with Venus. Despite the fact that Venus has no detectable intrinsic magnetic field, the solar wind is deflected about the planet and a bow shock formed. The ionospheric pressure balances the solar wind pressure to stand off the solar wind. The pressure of the solar wind is applied to the planetary ionosphere through a compression and pileup of magnetic field lines over the forward hemisphere of the planet. Magnetic field does penetrate the ionospheric barrier in two ways. First, bundles of magnetic field about 20 km across break through the ionopause and slip into the planetary ionosphere. The ionosphere sweeps these flux ropes to the nightside. However, as they move through the ionosphere, these tubes can become highly twisted and form links like a highly twisted rubber band.

The magnetic field also can diffuse and convect into the ionosphere. Usually the rate of diffusion is small enough that any field that enters the planetary ionosphere is swept away by ionospheric flows although, at the subsolar point, downward motion of the ionospheric plasma also can transport magnetic flux from the ionopause to low altitudes. When the solar wind pressure is high, especially when it exceeds the peak ionospheric pressure, diffusion

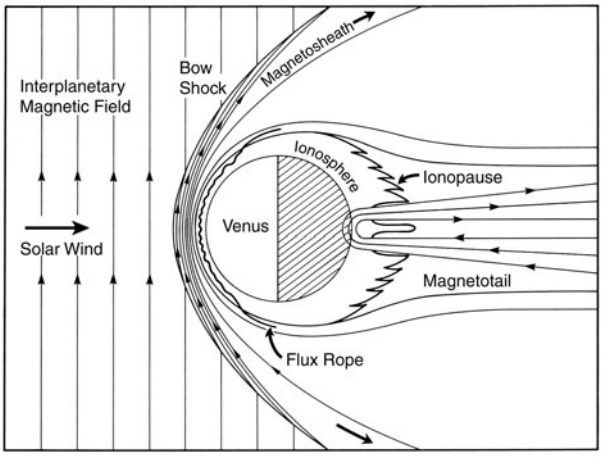


FIGURE 11 The solar wind interaction with Venus.

and downward transport of magnetic flux can become fast enough to magnetize the Venus ionosphere with a horizontal field of up to ~150 nT strength. At such times ion-neutral collisions in the atmosphere help support the ionosphere against the solar wind pressure. The magnetic field that enters the Venus ionosphere from the solar wind still has its ends in the solar wind and is dragged antisunward by the solar wind flow. This process contributes to the formation of a magnetic tail behind Venus.

The solar wind interaction with Venus is solar-cycle dependent with the bow shock farthest from the planet at the peak of the solar cycle when the solar ultraviolet radiation is the strongest. At solar minimum when the solar ultraviolet radiation and the Venus ionsphere are weaker the solar wind penetrates closer to the planet and more directly interacts with the neutral atmosphere. Thus, at solar minimum the ionosphere is more strongly magnetized and fewer electrons and ions are transported toward midnight to support the night-time ionosphere.

B. Solar Wind Interaction with Mars

The exploration of the planet Mars has been plagued with many problems and only recently has our understanding of the solar wind interaction with Mars approached our understanding of the interaction with Venus. This new understanding derives first from the 1989 Russian Phobos mission that provided 3 months of data from an initial elliptic orbit and then a circular orbit close to the moon Phobos. It also derives from the 1997 U.S. Mars Global Surveyor (MGS) mission that used aerobraking at altitudes below 140 km. These missions both reveal a very Venus-like interaction in which the solar wind interacts with the planetary ionosphere and any planetary magnetic field plays at most a very minor role. MGS additionally provided evidence for patches of strong surface remanent magnetization that show that, although Mars does not presently have an active dynamo, it once did.

VI. PLASMA INTERACTIONS WITH NEUTRAL GAS

The epitome of the interaction of a plasma with a neutral gas is the formation of a cometary tail in the solar wind. A cometary nucleus evaporates when it is close to the Sun. The expanding cloud of neutral gas is ionized by the solar ultraviolet radiation as well as by charge exchange with the solar wind and by impact ionization. This ionization makes the solar wind heavier and it slows down. However, the ends of the magnetic field lines are not affected and they continue to move at a rapid rate antisunward. This stretches the field lines out in a long magnetic tail behind a comet. In this section we will discuss our present understanding of comets. Then we examine the same processes as they occur at Venus, Io, and Titan. The major difference in the interactions with these three bodies is that their neutral atmospheres are gravitationally bound to them. This restricts the region of mass addition.

A. Solar Wind Interaction with Comets

For purposes of understanding the solar wind interaction with comets, the cometary nucleus can be thought of as a reservoir of frozen gases and dust. The gas may be locked up in a lattice of other material. Such an assemblage is called a clathrate. When the gas is warmed up by the approach of the comet to the Sun, it evaporates, expanding supersonically and carrying dust particles with it. The evaporation can occur from a limited region of the nucleus and form a jet, or it can occur rather uniformly. The resulting cloud of dust orbits with the comet, although light pressure changes the effective force of gravity on the dust so that the dust follows a slightly different path trailing out behind the comet. This dust cloud forms what is known as a type II cometary tail. The neutral gas expands to great distances up to several million kilometers or more in an active comet, with an emission rate that may exceed 10^{30} molecules per second. The neutral hydrogen in this cloud extends to millions of kilometers as can be determined from observations of the solar Lyman *a* scattered from the neutral hydrogen and forming a bright ultraviolet halo about the comet. The neutral gas becomes ionized by solar ultraviolet radiation, by charge exchange with solar wind ions, and by impact with solar wind electrons. Photoionization simply adds mass to the solar wind, as does impact ionization. Charge exchange does not add mass to the flow if the charge exchange is symmetric (i.e., between nuclei of the same species). However, a fast ion turns into a fast (400 km/sec) neutral in this process, so momentum is removed from the solar wind. If the charge exchange is between a solar wind proton and a heavy neutral atom such as oxygen, the solar wind plasma gains mass. All these processes lead to a slowing down of the solar wind flow because the electric field of the solar wind immediately accelerates the newly created ion up to the speed of the flow across the magnetic field. Since momentum is conserved, this acceleration of the ion must be accompanied by a slight deceleration of all the neighboring ions.

The energy of the solar wind flow is converted to energy of gyromotion of the decelerated heavy ions. Ions with energies of the order of 100 keV may be detected millions of kilometers from a comet. Near the comet the energy density in the picked-up ions may significantly exceed the energy density in other constituents in the plasma. The

large amount of energy in these heavy ions in turn leads to plasma instabilities and much turbulence downstream of the nucleus.

The interaction with a comet creates a slow spot in the solar wind. Well away from this slow spot, the solar wind is moving at an undiminished rate. Since the fast regions and slow regions are linked by the magnetic field lines, the magnetic field gets stretched out, forming a long tail. This long tail is filled with ions, at least in its central region, and can often be readily seen in bright comets. This is a type I or ion tail. Figure 12 shows the magnetic field lines obtained in computer simulations of Comet Halley for four different conditions, slow, normal, and fast solar wind velocity, and, on the bottom, normal velocity but enhanced mass addition rate.

If a comet is weak, there is no formation of a bow shock wave. However, if the comet is strong, the solar wind is unable to accommodate the added mass, a shock wave is formed, and the incoming flow is deflected around the heaviest mass-loading region. The newly added mass may, in fact, have enough pressure to create an ionopause and a magnetic field void in the interior of the interaction region. Such a void was observed at closest approach to Halley by the Giotto spacecraft.

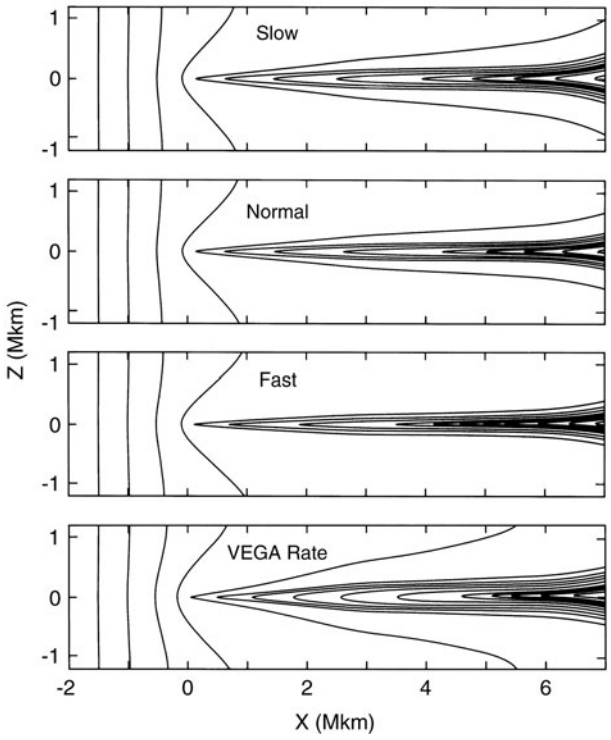


FIGURE 12 Magnetic field lines for four different computer simulations of Comet Halley (top three panels) using a cometary outgassing rate appropriate to the time of the Giotto encounter and solar wind conditions appropriate for slow, normal, and fast streams. Bottom: The cometary magnetic field for normal solar wind and VEGA encounter mass-loading rates. The solar wind moves from left to right. The visible ion tail corresponds to the region of greatest bending of the magnetic field.

Missions to comets—Giotto at both Halley and Grigg-Skjellerup; VEGA 1 and 2 at Halley; and ISEE-3 at Giacobini-Zinner—encountered strong ultralow-frequency wave turbulence. The VEGA 1 and 2 spacecraft at closest approach to Halley also saw ion condensations, indicative of the mirror instability, in which ions form pockets of high density and low magnetic field strength surrounded by regions of strong magnetic field and low ion density.

The tails of comets undergo many dynamic events as they encounter changing solar wind magnetic field orientations and changing solar wind conditions. The most dramatic of these is the so-called *tail disconnection event*, where the type I ion tail appears to be torn away from the comet. Undoubtedly, the behavior of the ion tail at these times is strongly affected by the tension in the bent magnetic field lines. This tension acts to accelerate the ion tail plasma to the ambient solar wind velocity.

B. Formation of the Venus Magnetotail

The neutral atmosphere of Venus is gravitationally bound to the planet. However, both the hydrogen and oxygen in the Venus atmosphere have hot components that can reach many thousands of kilometers in altitude before returning to the planet. As the solar wind passes through this so-called exosphere, it can pick up mass just as in the cometary interaction described above. This added mass can further slow down the solar wind and add magnetic flux to the Venus tail. Measurements behind Venus reveal a well-developed magnetic tail with two lobes as in the Earth's tail and a diameter of 4 Venus radii. However, at Venus the location of these two lobes depends on the orientation of the interplanetary magnetic field because they are formed by the draping of this field around the obstacle to the solar wind flow.

C. Plasma Interaction with Io

The interaction of the Jovian magnetospheric plasma with the moon Io is the engine that drives the dynamics of the Jovian magnetosphere, its aurora, and radio emissions. The tenuous, volcanically derived atmosphere of Io becomes ionized by photoionization, charge exchange, and impact ionization. As illustrated in Fig. 13, the strong Jovian magnetic field links this plasma to the rapidly rotating Jovian ionosphere. A current system is established that accelerates the newly added plasma to the angular speed of the ionosphere, which is much greater in linear

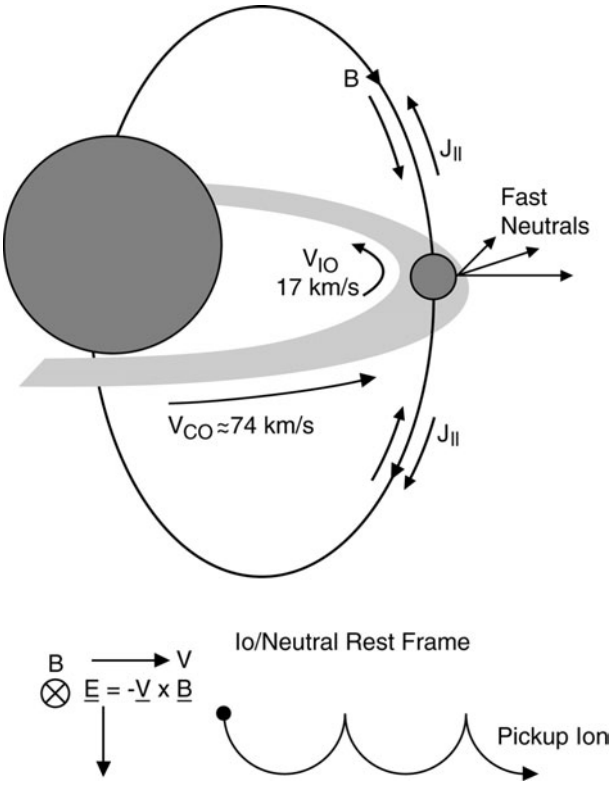


FIGURE 13 Torus interaction with Io. Electric currents $J_{||}$ flow along field lines from the ionosphere and close in the plasma produced at Io by sputtering and photoionization. The closure current through the plasma accelerates the plasma to corotational speeds. Individual ions follow a cycloidal path with an average velocity of 74 km/sec. Neutralization of ions executing cycloidal motion causes a spray of fast neutrals perpendicular to the magnetic field outward from Jupiter.

speed because Io is much farther from the rotational axis of Jupiter than is the ionosphere. Because the amount of mass added per second is large, perhaps 500 kg, the energy added to the plasma from the planetary rotation is about 500 GW, enough to light a bright light bulb for every person on Earth. Outward radial transport balances the addition of mass at Io. The transport process is very dynamic both because of the varying volcanic activity on Io and because of unsteadiness in the transport of plasma once injected into the magnetosphere. The effect of this added mass to the magnetosphere and its transport are described in detail in Section IV.C.

The region of newly added mass is surprisingly large but possibly quite thin because of the creation of a spray of neutrals caused when Io's atmosphere is ionized and then neutralized in Jupiter's rotating magnetosphere. The brief burst of acceleration while the particles are ionized causes them to leave Io rapidly perpendicular to the magnetic field. Where SO and SO_2 become reionized, strong waves at the ion gyrofrequencies are created. Mirror mode waves arise on the edge of the wake formed by Io in the corotating flow. These "waves" are condensations of plasma that produce large depressions in the magnetic field. In the wake itself the flow is slowed greatly and the plasma reaccelerates downstream from Io. A corresponding auroral spot at Io's footpoint and a long auroral wake downstream from that spot are seen in Jupiter's ionosphere. The formation of these auroras may be associated with the acceleration of the plasma in the wake and the coupling of the acceleration process to the ionosphere. Further, this magnetosphere–ionosphere coupling via field-aligned electric currents may also generate a number of radiofrequency phenomena that have been linked to Io. Overall most of the energetic phenomena in the aurora, the magnetosphere, and the radio emissions can be linked ultimately to Io. How much the solar wind adds to the energetics is still the subject of considerable debate.

D. Plasma Interaction with Titan

The interaction of the Saturnian magnetosphere with Titan is much less energetic than of the Io with the Jovian magnetosphere because the magnetic field at Titan is very weak compared to that at Io and the ionization rates much slower. Titan's atmosphere is dense, not tenuous, and is less sensitive to any present-day volcanic activity or outgassing of the planet. As a result any variability of the interaction of the Saturn on magnetosphere with Titan should depend more on the solar wind than on changes at Titan. The one passage of Voyager by Titan revealed strongly bent magnetic field lines and associated field-aligned currents. In many senses the interaction should resemble the solar wind interaction with Venus, rather than Jupiter's interaction with Io. Further study of the interaction of Titan with the Saturn magnetosphere will be obtained with the Cassini mission.

VII. CONCLUDING REMARKS

Plasmas in the solar system are complex entities in part because the currents, fields, and particle distributions are all interdependent physical parameters, linked through Maxwell's equations and the laws of classical mechanics and gravitation. Often intuition based on our observation of neutral gases and fluids fails us so that we find these plasmas behaving in quite unexpected ways. Much of the present understanding of space plasmas is based on a symbiotic relationship between theory and observation and assisted in recent years by large-scale computer simulations. This relationship will have to continue as we explore and attempt to understand the distant reaches of our solar system. The most distant reach of the solar system,

as far as electric and magnetic fields are concerned, is the heliopause, where the solar wind is stopped. Here the pressure of the galactic plasma, including its cosmic rays and magnetic field, is sufficient to balance the dynamic pressure of the solar wind. Our spacecraft have not reached the heliopause yet, but we expect that, if they keep operating, one of the Pioneer or Voyager spacecraft will one day penetrate this boundary, probably before the end of this century.

SEE ALSO THE FOLLOWING ARTICLES

AURORA • ELECTROMAGNETICS • GEOMAGNETISM • IONOSPHERE • MOON (ASTRONOMY) • PLANETARY SATELLITES, NATURAL • PLASMA SCIENCE AND ENGINEERING • SOLAR PHYSICS • SOLAR SYSTEM, GENERAL

BIBLIOGRAPHY

Battrick, B., and Rolfe, E. (Eds.). (1984). "Achievements of the International Magnetospheric Study (IMS)," European Space Agency, Noordwijk, The Netherlands.

Brandt, J. C., and Chapman, R. D. (1981). "Introduction to Comets," Cambridge University Press, New York.

Carovillano, R. L., and Forbes, J. M. (Eds.). (1983). "Solar Terrestrial Physics," Reidel, Dordrecht, The Netherlands.

Dessler, A. J. (Ed.). (1983). "Physics of the Jovian Magnetosphere," Cambridge University Press, New York.

Gehrels, T., and Matthews, M. S. (Eds.). (1984). "Saturn," University of Arizona Press, Tucson, AZ.

Grewing, M., Praderie, F., and Reinhard, R. (Eds.). (1988). "Exploration of Halley's Comet," Springer-Verlag, Berlin.

Hones, E. W., Jr. (Ed.). (1984). "Magnetic Reconnection in Space and Laboratory Plasmas," American Geophysical Union, Washington, D.C.

Hunten, D. M., Colin, L., Donahue, T. M., and Moroz, V. L. (Eds.). (1983). "Venus," University of Arizona Press, Tucson, AZ.

Kivelson, M. G., and Russell, C. T. (Eds.). (1995). "Introduction to Space Physics," Cambridge University Press, New York.

Kivelson, M. G., *et al.* (1998). "Ganymede's magnetosphere: Magnetometer overview," *J. Geophys. Res.* **103,** 19,963–19,972.

Lyons, L. R., and Williams, D. J. (1984). "Quantitative Aspects of Magnetospheric Physics," Reidel, Dordrecht, The Netherlands.

Merrill, R. T., and McElhinny, M. W. (1983). "The Earth's Magnetic Field: Its History, Origin, and Planetary Perspective," Academic Press, New York.

Potemra, T. A. (Ed.). (1984). "Magnetospheric Currents," American Geophysical Union, Washington, D.C.

Priest, E. R. (1982). "Solar Magnetohydrodynamics," Reidel, Dordrecht, The Netherlands.

Russell, C. T. (1980). "Planetary magnetism," *Rev. Geophys. Planetary Phys.* **18,** 77–106.

Russell, C. T. (Ed.). (1986). "Solar wind interactions," *Adv. Space Res.* **6**(1).

Southwood, D. J., and Russell, C. T. (Eds.). (1982). "Special issue on international magnetospheric study," *Rev. Geophys. Planetary Phys.* **20**.

Vilas, F., Chapman, C. R., and Matthews, M. S. (Eds.). (1988). "Mercury," University of Arizona Press, Tucson, AZ.