I. Thermal Evolution of Ganymede and Implications for Surface Features II. Magnetohydrodynamic Constraints on Deep Zonal Flow in the Giant Planets III. A Fast Finite-Element Algorithm for Two-Dimensional Photoclinometry

Thesis by

RANDOLPH LIVINGSTONE KIRK

In partial fulfillment of the requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1987

(submitted 9 January 1987)

Dedicated to my parents, and all who sail in them.

Acknowledgements

It is difficult to express the deep admiration and appreciation I feel for my thesis advisor, Dr. David Stevenson. Dave, you provide an example of scientific integrity and creativity I know I will always find hard to live up to. Thank you for suggesting the topics that led to parts I and II of this thesis, for your insight, your patience, and all those marvelous (though much too easy, of course) hikes you organized. May you always be around at crucial moments to remind me when I've "got to be careful..."

I am also indebted to my academic advisor, Dr. Peter Goldreich, for his support and helpful comments over the years. Thanks are due to Dr. Andrew Ingersoll and Dr. Bruce Murray for their interest in and suggestions about the various sections of my thesis. Dr. Duane Muhleman I thank for the pleasure of his forthright opinions, as well as for the stimulation his research has given me. I wish I'd gotten to try my hand at some radio science, Dewey... but I don't think any of us could have survived a *four* part thesis. To Dr. Yuk Yung, I am sorry I didn't diffuse down to your end of the hall more often. I thank the entire Planetary Science faculty for accepting the diversity of my interests and allowing me to write a thesis in three such isolated compartments.

Thanks to Dr. Joseph Kirschvink for piquing my interest in photoclinometry as a tool for looking at fossils, and for his good humor when I ran off and turned it into a remote sensing technique. Your friendship is appreciated and your enthusiasm is always infectious, Joe. Save the whales!

I am grateful to Professor James Westphal for making available and helping me use the PFUEI CCD camera and the computers that go with it. Without his assistance the photoclinometry algorithm would have been stillborn. Thanks also to Dr. Larry Soderblom for providing a tape of digital topographic data.

Thanks to Dr. Robert Sharp for many fine fieldtrips around the Southwest,

and for the opportunity to participate in Project Pahoehoe 1987. We planetary scientists have a tendency to go straight from the computer room to low earth orbit. We're lucky to have you, Dr. Sharp, to remind us what a real planet is like close up: messy but comprehensible.

To Dr. David Crisp, thanks for a number of insightful conversations, but particularly for your selfless willingness to straighten out computer problems of any magnitude. You truly win the "his brother's keeper" award.

Thanks to my fellow students, from my entering class out into the exponential tails on either side, for your friendship and fellowship in the past five and a half years. Thanks especially to Don and Carol, the two best office partners a boy could *ever* have, and to Jim and Jon, and Greg and Ken and Lottie, and to all the rest of you tads and geezers whose names I haven't forgotten but scarcely have room for.

To my long-time friends up north CJM and DHO, thank you for staying in touch over the last few years. Maybe I'll be able to hold up my end of the correspondence again now. I might also promise the same to my parents, but I'd rather take the opportunity to thank them for the morale support they have given me, year in and year out (not to mention the fruitcakes).

Finally, it is a pleasure to be able to thank the administrative staff for making the nonscience parts of life in South Mudd so easy. Particular thanks to Kay Campbell, our official "Mom," and to Donna Lathrop for yeoman service on the keyboards.

Abstract

The work is divided into three independent papers:

PAPER I:

Thermal evolution models are presented for Ganymede, assuming a mostly differentiated initial state of a water ocean overlying a rock layer. The only heat sources are assumed to be primordial heat (provided by accretion) and the long-lived radiogenic heat sources in the rock component. As Ganymede cools, the ocean thins, and two ice layers develop, one above composed of ice I, and the other below composed of high-pressure polymorphs of ice. Subsolidus convection proceeds separately in each ice layer, its transport of heat calculated using a simple parameterized convection scheme and the most recent data on ice rheology. The model requires that the average entropy of the deep ice layer exceed that of the ice I layer. If the residual ocean separating these layers becomes thin enough, then a Rayleigh-Taylor-like ("diapiric") instability may ensue, driven by the greater entropy of the deeper ice and merging the two ice mantles into a single convective layer. This instability is not predicted by linear analysis but occurs for plausible finite amplitude perturbations associated with large Rayleigh number convection. The resulting warm ice diapirs may lead to a dramatic "heat pulse" at the surface and to fracturing of the lithosphere, and may be directly or indirectly responsible for resurfacing and grooved terrain formation on Ganymede. The timing of this event depends rather sensitively on poorly known rheological parameters but could be consistent with chronologies deduced from estimated cratering rates. Irrespective of the occurrence or importance of the heat pulse, we find that lithospheric fracturing requires rapid stress loading (on a timescale $\lesssim 10^4$ years). Such a timescale can be realized by warm ice diapirism, but not directly by gradual global expansion. In the absence of any quantitative and self-consistent model for the resurfacing of Ganymede by liquid water, we favor resurfacing by warm ice flows,

which we demonstrate to be physically possible, a plausible consequence of our models, compatible with existing observations, and a hypothesis testable by Galileo. We discuss core formation as an alternative driver for resurfacing, and conclude that it is less attractive. We also consider anew the puzzle of why Callisto differs so greatly from Ganymede, offering several possible explanations. The models presented do not provide a compelling explanation for all aspects of Ganymedean geological evolution, since we have identified several potential problems, most notably the apparently extended period of grooved terrain formation (several hundred million years), which is difficult to reconcile with the heat pulse phenomenon.

PAPER II:

The observed zonal flows of the giant planets will, if they penetrate below the visible atmosphere, interact significantly with the planetary magnetic field outside the metalized core. The appropriate measure of this interaction is the Chandrasekhar number $Q = \frac{H^2}{4\pi\rho\nu\alpha^2\lambda}$ (where H = radial component of the magnetic field, ν = eddy viscosity, λ = magnetic diffusivity, α^{-1} = lengthscale on which λ varies); at depths where $Q \gtrsim 1$ the velocity will be forced to oscillate on a small lengthscale or decay to zero. We estimate the conductivity due to semiconduction in H_2 (Jupiter, Saturn) and ionization in H_2O (Uranus, Neptune) as a function of depth; the value $\lambda \simeq 10^{10}\,\mathrm{cm^2\,s^{-1}}$ needed for $Q = 1$ is readily obtained well outside the metallic core (where $\lambda \simeq 10^2\,\mathrm{cm^2\,s^{-1}}$).

These assertions are quantified by a simple model of the equatorial zonal jet in which the flow is assumed uniform on cylinders concentric with the spin axis, and the viscous and magnetic torques on each cylinder are balanced. We solve this "Taylor constraint" simultaneously with the dynamo equation to obtain the velocity and magnetic field in the equatorial plane. With this model we reproduce the widely differing jet widths of Jupiter and Saturn (though not the flow at very high or low

latitudes) using $\nu = 2500\,\mathrm{cm^2\,s^{-1}}$, consistent with the requirement that viscous dissipation not exceed the specific luminosity. A model Uranian jet consistent with the limited Voyager data can also be constructed, with appropriately smaller ν, but only if one assumes a two-layer interior. We tentatively predict a wide Neptunian jet.

For Saturn (but not Jupiter or Uranus) the model has a large magnetic Reynolds number where $Q = 1$ and hence exhibits substantial axisymmetrization of the field *in the equatorial plane.* This effect may or may not persist at higher latitudes. The one-dimensional model presented is only a first step. Variation of the velocity and magnetic field parallel to the spin axis must be modeled in order to answer several important questions, including: 1) What is the behavior of flows at high latitudes, whose Taylor cylinders are interrupted by the region with $Q \gtrsim 1$? 2) To what extent is differential rotation in the envelope responsible for the spin-axisymmetry of Saturn's magnetic field?

PAPER III:

It is shown that the problem of two-dimensional photoclinometry (PC) — the reconstruction of a surface $z(x, y)$ from a brightness image $B(x, y)$ — may be formulated in a natural way in terms of finite elements. The resulting system of equations is underdetermined as a consequence of the lack of boundary conditions for z, but a unique solution may be chosen by minimizing a function S expressing the "roughness" of the surface. An efficient PC algorithm based on this formulation is presented, requiring ~ 10.66 (four-byte) memory locations and $\sim 10^4$ floating multiplications/additions per pixel, and incorporating: 1) Minimization of the roughness by the penalty method, which yields the smallest set of equations. 2) Iterative solution of the nonlinear equations by Newton's method. 3) Solution of the linearized equations by an inner iterative cycle of successive over-relaxation, which takes advantage of the extreme sparseness of the system. 4) Multigridding, in which the solutions to the smaller problems ob-

tained by reducing the resolution are used recursively to greatly speed convergence at the higher resolutions, and 5) A rapid noniterative initial estimate of z obtained by exploiting the special symmetry of the equations obtained in the first linearization.

The algorithm is extensively demonstrated on 200 by 200 pixel synthetic "images" generated from digital topographic data for northern Utah over a range of phase angles. Rms error in the solution is $\sim 22\,\mathrm{m}$, out of $\sim 660\,\mathrm{m}$ total relief. The error is dominated by "stripes" with the same azimuth as the light source, resulting from use of the roughness criterion in lieu of boundary conditions; the rms error along profiles parallel to the stripes is only $\sim$ 2–8 m, depending on the phase angle. Satisfactory solutions are obtained even in the presence of quantization error, noise, and moderate blur in the image.

Applications of the PC algorithm to both remote sensing and photomacrography are sketched; a photoclinometric map of a low-relief Precambrian era fossil is presented as an example of the latter. Prospects for dealing with photometrically inhomogeneous surfaces, and an extension of the method to the analysis of side-looking radar data ("radarclinometry") are also discussed.

Table of Contents

List of Figures

PAPER I:

List of Tables

PAPER I:

PAPER II:

PAPER III:

PAPER I

Thermal Evolution of a Differentiated Ganymede and Implications for Surface Features

That is, hot ice and wondrous strange snow.

— William Shakespeare *A Midsummer Night'sDream*, V, I, 59

Thermal Evolution of a Differentiated Ganymede and Implications for Surface Features

R. L. KIRK AND D. J. STEVENSON

Division of Geological and Planetary Sciences
California Institute of Technology
Pasadena, California 91125

Published in modified form in *Icarus*
January, 1987

Contribution number 4255 from the Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125.

Abstract

Thermal evolution models are presented for Ganymede, assuming a mostly differentiated initial state of a water ocean overlying a rock layer. The only heat sources are assumed to be primordial heat (provided by accretion) and the long-lived radiogenic heat sources in the rock component. As Ganymede cools, the ocean thins, and two ice layers develop, one above composed of ice I, and the other below composed of high-pressure polymorphs of ice. Subsolidus convection proceeds separately in each ice layer, its transport of heat calculated using a simple parameterized convection scheme and the most recent data on ice rheology. The model requires that the average entropy of the deep ice layer exceed that of the ice I layer. If the residual ocean separating these layers becomes thin enough, then a Rayleigh-Taylor-like ("diapiric") instability may ensue, driven by the greater entropy of the deeper ice and merging the two ice mantles into a single convective layer. This instability is not predicted by linear analysis but occurs for plausible finite amplitude perturbations associated with large Rayleigh number convection. The resulting warm ice diapirs may lead to a dramatic "heat pulse" at the surface and to fracturing of the lithosphere, and may be directly or indirectly responsible for resurfacing and grooved terrain formation on Ganymede. The timing of this event depends rather sensitively on poorly known rheological parameters but could be consistent with chronologies deduced from estimated cratering rates. Irrespective of the occurrence or importance of the heat pulse, we find that lithospheric fracturing requires rapid stress loading (on a timescale $\lesssim 10^4$ years). Such a timescale can be realized by warm ice diapirism, but not directly by gradual global expansion. In the absence of any quantitative and self-consistent model for the resurfacing of Ganymede by liquid water, we favor resurfacing by warm ice flows, which we demonstrate to be physically possible, a plausible consequence of our models, compatible with existing observations, and a hypothesis testable by Galileo. We

discuss core formation as an alternative driver for resurfacing, and conclude that it is less attractive. We also consider anew the puzzle of why Callisto differs so greatly from Ganymede, offering several possible explanations. The models presented do not provide a compelling explanation for all aspects of Ganymedean geological evolution, since we have identified several potential problems, most notably the apparently extended period of grooved terrain formation (several hundred million years), which is difficult to reconcile with the heat pulse phenomenon.

1. Introduction

Thermal evolution modeling of planets and satellites is a frustrating game because there is usually little connection between what one can calculate and what one can observe. Large icy satellites have proved to be no exception. The startling diversity of geology on Ganymede revealed by the Voyagers and the puzzling dissimilarity of Callisto have prompted many efforts to understand these bodies, but little consensus has emerged. The field has progressed from an early elucidation of principles for their internal structure (Huaux 1951; Lewis 1971*a*, *b*; Consolmagno and Lewis 1976) and their solid-state convection histories (Reynolds and Cassen 1979; Parmentier and Head 1979; Cassen *et al.* 1980; Thurber *et al.* 1980) to specific, hence questionable, models for the resurfacing of Ganymede or the Ganymede-Callisto differences (Squyres 1980*a*; Schubert *et al.* 1981; McKinnon 1981; Shoemaker *et al.* 1983; Lunine and Stevenson 1982; Friedson and Stevenson 1983). For a recent review, see Schubert *et al.* (1986). From a geologic perspective, the existence of terrains with different crater densities and varying degrees of crater degradation, presumably the result of viscous relaxation, has led to some hope that the evolution can be reconstructed (*e.g.*, Passey 1982) but uncertainties in interpretation and rheology persist.

The significance of the work presented here lies not in the presentation of yet another parameterized thermal evolution calculation (these are all too easy to

perform) but in the identification of several processes and principles which, although embodied in the specific models presented here, may apply to a wide range of models for the evolution of large icy bodies, especially Ganymede. In particular, we have attempted (with admittedly limited success) to establish possible connections between interior evolution and the geologic evidence for Ganymede.

As in all modeling efforts, some assumptions are necessary. The most important of these is the assumption of a differentiated structure for Ganymede: a rock-rich core surrounded by a water ice mantle. This assumption is not compelled by data but strongly implied by calculations of satellite accretion (Schubert *et al.* 1981; Lunine and Stevenson 1982). These calculations also motivate our assumptions that Ganymede was initially "hot" (most of the H_2O in a liquid state) and retained only minor amounts of molecules more volatile than H_2O. (Minor quantities of water-soluble constituents, probably dominated by NH_3, nevertheless play an important role in our modeling.) We assume, implicitly, that parameterized subsolidus convection recipes, popularized for the icy satellites by Reynolds and Cassen (1979), provide an adequate quantification of the thermal history. The only heat sources assumed are radiogenic and primordial (accretional). We have also made a thorough and critical assessment of ice rheology and utilized the most recent laboratory data. With this background, our effort has identified and focused on the following features:

(i) *The Heat Pulse.* As a consequence of the H_2O phase diagram, all models of Ganymede that begin hot must develop an outer ice I layer, an underlying water ocean, and a deeper layer of high-pressure polymorphs of ice. The two ice layers undergo separate solid-state convection, with the average entropy of the deeper ice layer substantially exceeding that of the ice I layer. If cooling causes the intervening ocean to (almost) freeze, then the two layers may merge, to form a single, nearly isentropic, convecting layer. If this merging occurs,

excess heat from the deeper layer may be released by a Rayleigh-Taylor-like instability, causing warm ice diapirism, which helps transport a pulse of heat to the surface. Although this instability is not predicted by linear analysis (Bercovici *et al.* 1986), it is expected to happen for plausible perturbations of a background state of finite amplitude convection. We speculate that this overturn may be responsible, directly or indirectly, for resurfacing and grooved terrain formation.

(ii) *Rapid Fracturing.* Irrespective of whether the heat pulse occurred or was important, we find that fracturing of the Ganymedean lithosphere must have required the rapid imposition of stress ($\lesssim 10^4$ year timescale), since otherwise the ice responds primarily by creep. This argues against attributing resurfacing to a geologically slow process such as global expansion but is consistent with the stress history imposed by a warm ice diapiric upwelling.

(iii) *The Role of Impacts.* We see several important roles for large impacts. First, we argue that stirring of silicate fines from the ocean into the ice I layer by impact may be important in controlling the ice viscosity. Second, a very large impact may trigger the merging of the two ice layers and allow the heat pulse to occur when otherwise it would not, because of soluble impurities that limit freezing of the ocean at a thickness for which spontaneous overturn is impossible. Third, impacts may puncture the lithosphere, providing pathways for underlying warm ice to flow onto the surface. We consider it highly probable that the exogenic and endogenic histories of Ganymede are intimately coupled.

(iv) *Resurfacing by Ice Flows.* No quantitative and self-consistent model currently exists whereby liquid water is the resurfacing agent for Ganymede. We demonstrate that resurfacing by warm ice is physically possible, a plausible consequence of our models, compatible with existing observations, and a hypothesis

testable by Galileo.

The plan of the remainder of this paper is as follows. Section 2 contains a detailed description of our thermal evolution model, its first three subsections devoted to the terms of the energy balance equation: heat output from the core, sensible and latent heat of cooling, and convective heat transport to the surface. In Section 2.4 we outline the modifications that must be made to these terms when the liquid ocean is gone, and in 2.5 we summarize the range of results obtained.

Section 3 addresses at some length the "heat pulse" phenomenon, with subsection 3.1 devoted to the mechanism and timing of the pulse, and 3.2 to its important consequences for heat flow (and hence viscous degradation of impact craters) and lithospheric stresses due to warm ice diapirism.

Criteria for extension fracturing by both global expansion and regional lithospheric warping are developed in Section 4; although the latter is directly applicable to the results of Section 3.2, we present these results separately to emphasize their independence from any assumptions about the cause of the lithospheric stress.

Evolution of the core from its gravitationally unstable initial condition is discussed in Section 5. The short timescales obtained, even under conservative assumptions, for most of the core differentiation, suggest that this process did not play an important role in resurfacing.

Our discussion, Section 6, considers the plausibility of resurfacing Ganymede with solid, rather than liquid, H_2O and examines some possible mechanisms for the formation of grooves. Explanations for the lack of resurfacing on Callisto in light of our model are presented, and finally the possibilities for resolving some of these difficulties are assessed.

Following the paper is an Appendix in which we consider the available constraints, both observational and theoretical, on the viscosity of ice in the Ganymedean

mantle, which is the most important and ill-constrained input to our models.

2. The Thermal Model

The bulk of this paper concerns our modeling of the thermal evolution of a Ganymede-sized rock-ice body (physical properties summarized in Table I) whose outer regions have differentiated because of strong heating during formation (Schubert *et al.* 1981; Lunine and Stevenson 1982) or subsequently (Friedson and Stevenson 1983). Our model takes as its initial state a body with a liquid water ocean overlying a silicate core comprising the rock that accreted contemporaneously with the ocean water. An inner core of undifferentiated ice-rock mixture may also exist initially but does not enter into the thermal history model and is in any event rapidly eliminated by warming and differentiation of the core (see Section 5). Figure 2.1 illustrates the model structure at a time when part of the ocean has frozen but the core is not yet fully differentiated.

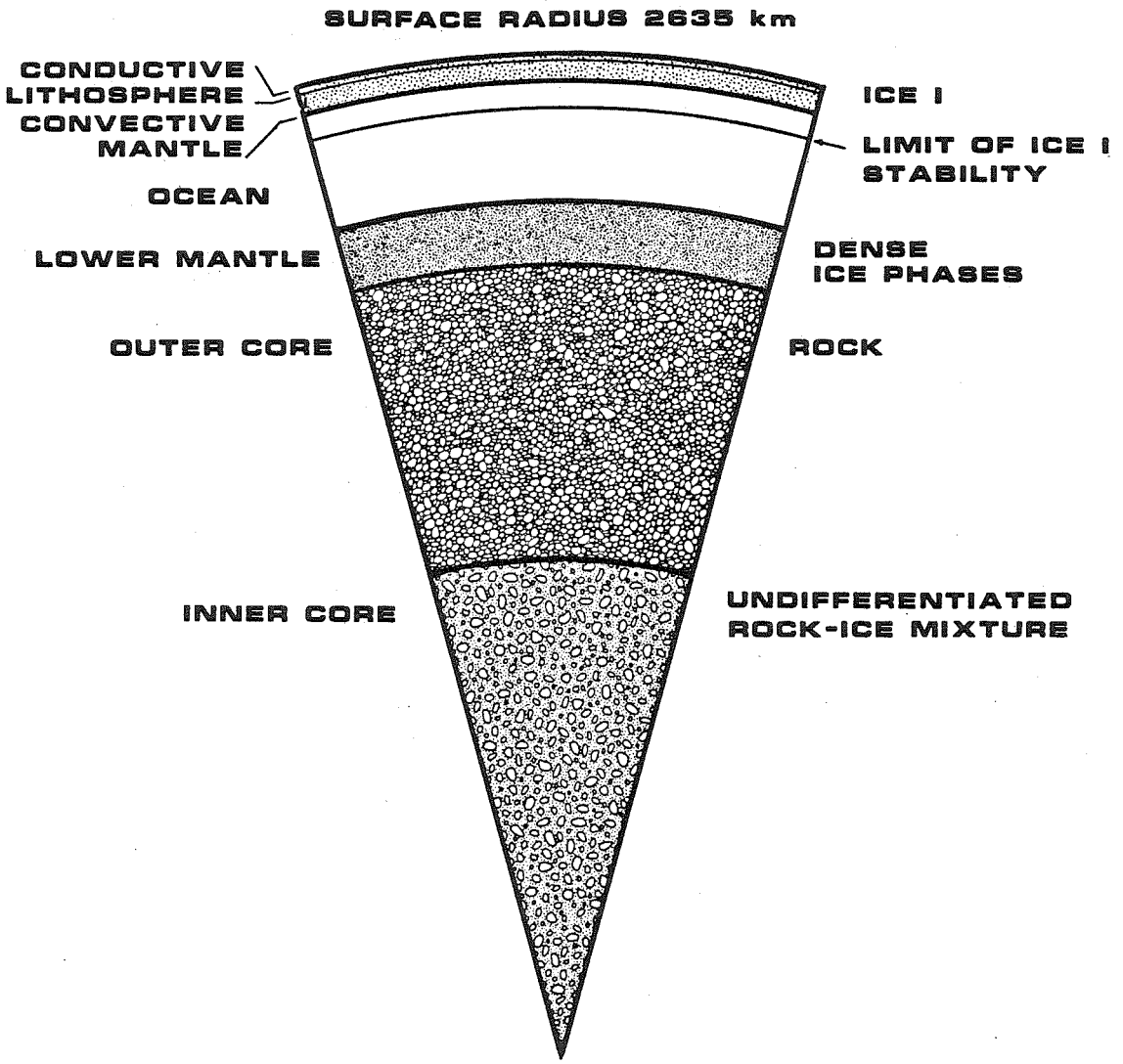
Table I. Physical Properties of Ganymede

Quantity	Units	Magnitude
R	m	2.635×10^6
M	kg	1.490×10^{23}
$\overline{\rho}$	$\mathrm{kg\,m^{-3}}$	1944
g	$\mathrm{m\,s^{-2}}$	1.43

The ocean will convect strongly (Lunine and Stevenson 1982) and will cool extremely rapidly until its surface reaches the zero-pressure melting point and an ice I crust begins to form. This first stage of cooling, lasting only $\lesssim 10^4$ years (Cassen *et al.* 1982) was not modeled by us. In the second period, the rate of cooling is controlled by conduction through the ice crust. To a first approximation, the temperature drop across the crust is constant, from about 130 K determined by thermal annealing of the regolith above (Passey and Shoemaker 1982) to the

Figure 2.1. Model interior structure of Ganymede during freezing of the ocean but before differentiation of the inner core.

Ganymede Pie



melting temperature (for a thin crust nearly the zero-pressure value) below. The crustal thickness thus increases roughly as $t^{1/2}$ and heat flow falls off as $t^{-1/2}$ until convection sets in or the radiogenic heating becomes an appreciable part of the total heat budget. Once the ice I layer reaches a critical thickness, which depends on its rheologic properties — typically after 5×10^6 to 5×10^7 years — it becomes unstable to convection in its lower portion (the "ice I mantle"). Thus begins the third major stage of cooling, in which heat loss is controlled by subsolidus convection, and the radiogenic output of the core becomes increasingly important. It is with the numerical modeling of this period, which extends to the present day, that we are primarily concerned. An exact description of Ganymede's thermal state in this period would involve solution of the fluid-mechanical equations for thermal convection, including phase change and conduction of heat, in three dimensions and as a function of time. By assuming a parameterized form (cf. Figure 2.2) for the globally averaged Ganymedotherm, however, we obtain a single ordinary differential equation expressing energy balance for the body:

$$4\pi R^2 F = 4\pi R_c^2 F_c + \left.\frac{dE}{dt}\right|_{sens} + \left.\frac{dE}{dt}\right|_{lat}. \qquad (2.1)$$

R and F are the radius and heat flux at the surface and R_c and F_c, those at the boundary of the core, and $\left.\frac{dE}{dt}\right|_{sens}$ and $\left.\frac{dE}{dt}\right|_{lat}$ represent, respectively, the sensible heat released by cooling and the latent heat of freezing. The energy deposited by infalling planetesimals is neglected in this energy budget; the cratering rate of Shoemaker and Wolfe (1982) corresponds to a miniscule energy flux (although the flux at the earliest times is unconstrained and could have been much larger). Similarly, tidal dissipation is small at present and was neglected, but a large initial free eccentricity, leading to significant dissipation in the first 10^9 years, cannot be ruled out (Cassen *et al.* 1982). The gravitational potential energy released by differentiation was ignored as well.

Our thermal evolution scenarios constitute time histories of the Ganymedo-

therm and the terms in the heat budget — in particular, F — that satisfy (2.1). In the following sections we describe the assumptions on which they are based.

2.1 Core Heat Flux

In the early stages of evolution, the heat flux from the core is determined by conduction. The core is initially isothermal with the ice above it, gradually warming as radionuclides in the rock decay. Relative to its initial temperature, at time t it is warmed by an amount

$$\Delta T_c = \frac{1}{C_c} \sum_i X(q_i) A(q_i) \frac{1 - e^{-\lambda_i t}}{\lambda_i}, \tag{2.2}$$

where C_c is the heat capacity of the core, and the sum is over radioisotopes q_i (in our model, the long-lived isotopes of K, Th and U) with decay constants λ_i, radiated power per mass of element A_i, and initial abundances X_i expressed as mass ratios.

The temperature of the core is elevated by ΔT_c at depth, but it drops to that of the overlying ice across a boundary region of width $\sim \sqrt{\kappa_c t}$, where κ_c is the thermal diffusivity of the core rock. The concomitant flux conducted across the boundary is

$$F_c = \frac{2 k_c \Delta T_c}{\sqrt{\pi \kappa_c t}}, \tag{2.3}$$

where k_c is the thermal conductivity of the rock. Combining (2.2) and (2.3), we obtain:

$$4\pi R_c^2 F_c = 8 \rho_c R_c^2 \sqrt{\pi \kappa_c t} \sum_i X(q_i) A(q_i) \frac{1 - e^{-\lambda_i t}}{\lambda_i t}. \tag{2.4a}$$

Table II gives our choices of nominal core density and inner and outer radii, along with rock parameters based on terrestrial ultramafic rocks (Birch 1942), while Table III gives the assumed "chondritic" abundances of ^{40}K, ^{232}Th, ^{235}U and ^{238}U (Kaula 1968) included in the model.

As the core warms, subsolidus convection will eventually become possible. To determine the time at which this occurs, the stability of the core against convection

Table II. Nominal Core Model

Quantity	Units	Magnitude
R_{ic}	m	1.305×10^{6a}
		0^{b}
R_c	m	2.06×10^{6a}
		1.96×10^{6b}
M_c	kg	9.4×10^{22}
ρ_c	$kg\,m^{-3}$	3000
g_c	$m\,s^{-2}$	1.7
k_c	$W\,m^{-1}\,K^{-1}$	3.0
C_c	$J\,kg^{-1}\,K^{-1}$	920
κ_c	$m^2\,s^{-1}$	1.1×10^{-6}
α_c		2.4×10^{-5}
η_{0c}	Pa s	1.7×10^{16}
A_c		29.4
T_{mc}	K	2300

[a] Before core differentiation.
[b] After core differentiation (used in thermal model).

Table III. Chondritic Radiosotope Abundances

Isotope	X (ppm)	A ($W\,kg^{-1}$ pure element)[a]	XA ($W\,kg^{-1}$ rock)	λ (Gy^{-1})
^{40}K	845	3.70×10^{-8}	3.13×10^{-11}	0.531
^{232}Th	0.04	3.33×10^{-5}	1.33×10^{-12}	0.0499
^{235}U	0.012	3.42×10^{-4}	4.10×10^{-12}	0.972
^{238}U	0.012	1.88×10^{-4}	2.26×10^{-12}	0.154
			Total: 3.89×10^{-11}	

[a] Natural isotopic abundances at 4.55 Gybp.

was calculated as part of the thermal model by evaluating the core Rayleigh number:

$$Ra_c(t) = \frac{\rho_c g_c \alpha_c \Delta T_c (\kappa_c t)^{\frac{3}{2}} \delta^4}{\kappa_c \eta_c (T(R_c) + (1 - \delta/2)\Delta T_c)}, \tag{2.5}$$

with the dimensionless boundary-layer thickness (expressed as a fraction of the thermal diffusion depth) δ chosen to satisfy:

$$\frac{\partial Ra_c}{\partial \delta} = 0. \tag{2.6}$$

A homologous-temperature formulation (equation A.1) was used for the silicate

viscosity, but with viscosity parameters appropriate to olivine at a pressure of 2 GPa (Table III). The criterion $Ra_c = 10^3$ was used for the onset of convection.

Proper treatment of the core heat flux after the onset of convection would require a parameterized convection model for the core operating in parallel to that for the icy shell. Such a model would, by virtue of the strong dependence of convective flux on temperature, self-regulate. We therefore made the much simpler assumption that the convective flux from the core is in instantaneous equilibrium with the rate of energy release by radioactive decay:

$$4\pi R_c^2 F_c = \frac{4\pi}{3}\rho_c R_c^3 \sum_i X(q_i)A(q_i)e^{-\lambda_i t}. \qquad (2.4b)$$

(One might object that a similar argument could be made for the conductive cooling of the ice mantles , thus rendering redundant our entire modeling effort, but, unlike the core, the mantles begin convection at a temperature well above their self-regulation point, and their cooling is strongly buffered by the latent heat of freezing.)

A possible source of departure of the radiogenic heating rate from the conductive-convective core model is the retention of radionuclides in the ocean through either of two mechanisms, both crucially dependent on the abundance of very small silicate grains. First, as we show in the Appendix, grains of radius $r \lesssim 100\,\mu$m may remain suspended in the convecting ocean. Second, John Lewis (personal communication, 1983) has pointed out that potassium is readily leached from chondritic fines by liquid water. Such oceanic radionuclides release their energy directly into the heat budget for the icy envelope, increasing the radiogenic heating at the earliest times when the core is conductive. Once the core is convective (which will occur later because its rate of warming is decreased) the total radiogenic contribution to the heat budget is equal to the equilibrium rate of energy release, independent of the distribution of the heat sources. Because of the uncertainty in particle size distribution which must be factored into these effects, we will present what are probably limiting cases: all

radionuclides in the core, and 30% of the radionuclides in the ocean.

For each assumption about the distribution of the radionuclides between core and ocean, we will present a range of thermal evolution scenarios in which their total abundance is a multiple of the nominal value. The "shape" of the core heat output is affected by the *relative* abundances of various nuclides only weakly because their half-lives are (with the exception of ^{235}U, which makes only a minor contribution to the total energy) long and similar. On the other hand, the range of plausible total radionuclide abundances is both large and difficult to quantify. Our nominal model is based on an analogy between Ganymede rock and chondritic meteorites, with a K/U ratio of 7×10^4 (Kaula 1968). Estimates of terrestrial radionuclide abundances, in comparison, involve substantial depletion of potassium, as well as uncertainties of a factor of two in uranium abundance (Ganapathy and Anders 1974; Wasserburg *et al.* 1964). Such estimates applied to Ganymede lead to core heat fluxes of 0.5–1.2 times the chondritic value. Although extreme potassium depletion in Ganymede is unlikely, uncertainties in radionuclide abundances probably lead to at least ±20% uncertainty in F_c.

A smaller uncertainty arises from the poorly known density of the core. Although F_c could range from 2500–3500 $\mathrm{kg\,m^{-3}}$, depending on the degree of hydration, the constraint implied by the known mean density of Ganymede (plus the assumption $R_{ic} = 0$, appropriate for times $\gtrsim 10^8$ y) leads to only ±4% variation of the factor $F_c R_c^2$ in (2.4a), and a ±15% variation in $F_c R_c^2$ in (2.4b).

2.2 Sensible and Latent Heat

Calculation of the heat released by cooling and by freezing is quite straightforward. Figure 2.3 shows typical Ganymedotherms superimposed on a simplified phase diagram of pure H_2O based on Hobbs (1974). Clearly, if one knows the rate of cooling $\frac{dT}{dt}$ as a function of depth (or equivalently pressure) one can find the rate of change

of sensible heat:

$$\left.\frac{dE}{dt}\right|_{sens} = \int_0^{P_{core}} C(P)\frac{\partial T(P)}{\partial t}\frac{dM}{dP}\,dP, \tag{2.7}$$

where $C(P)$ is the appropriate heat capacity for the phase occurring at pressure P and $M(P)$ is the mass of material lying above the pressure P, calculated assuming hydrostatic equilibrium. Note that the integral does not extend into the core, which was treated separately. Similarly, the rate of release of latent heat is given by:

$$\left.\frac{dE}{dt}\right|_{lat} = \sum_i L_i \frac{\partial T}{\partial t}(P_i)\left(\frac{dT_{mi}}{dP}\right)^{-1}\frac{dM}{dP}(P_i), \tag{2.8}$$

where the sum is over liquid-solid phase transitions occurring at temperatures $T_{mi}(P)$ and having latent heats L_i: at any given time, ice I and one of the dense ice phases III, V, or VI. Latent heats of solid-solid phase transitions do not contribute significantly to the heat budget, although they do cause small offsets of the convective adiabat where it crosses boundaries between solid phases. Thermodynamic data for ice taken from Hobbs (1974) are summarized in Tables IV and V. Heat capacities of the dense ice phases are not well known, so that of ice I was used throughout.

Table IV. Triple Point Data for Water

Triple Point	P (GPa)	T (K)	$L_{solid-solid}/C$ (K)[a]
Vap–L–I	6×10^{-7}	273.2	...
L–I–III	0.207	251.1	−10.88[b]
L–III–V	0.346	256.1	2.05
L–V–VI	0.626	273.3	0.48

[a] Temperature offset (in direction of increasing P) of an adiabat.
[b] Decreases to −4.59 K at the I–II–III triple point.

Further simplification of equations (2.7) and (2.8) results from the fact that in our parameterization of the Ganymedotherm, $T(P)$ is fully specified (for a given surface temperature) by a single tiepoint, conveniently taken as the pressure P_5 at the boundary between the liquid water ocean and the heavy ices beneath it. (Nomenclature for the segments of the parameterized Ganymedotherm, and the (P,T) points bounding them, is defined in Figure 2.2). The temperature T_5 at the pressure P_5 is

Figure 2.2. Parameterized forms of the Ganymedotherm . A rigid conductive cap extends from the surface to (P_1, T_1), and a thermal boundary layer from there to (P_2, T_2) in all cases. (a) Initial state of convection in three cells: ice I from (P_2, T_2) to (P_3, T_3) with lower boundary layer from there to (P_4, T_4); liquid water from (P_4, T_4) to (P_5, T_5); and dense ices from there to the core. (b) After bridging of the residual ocean. A single convective adiabat extends from (P_2, T_2) to the core. Offsets in the Ganymedotherm are due to the solid-solid latent heats. T_e is the temperature extrapolated to the surface on the adiabat.

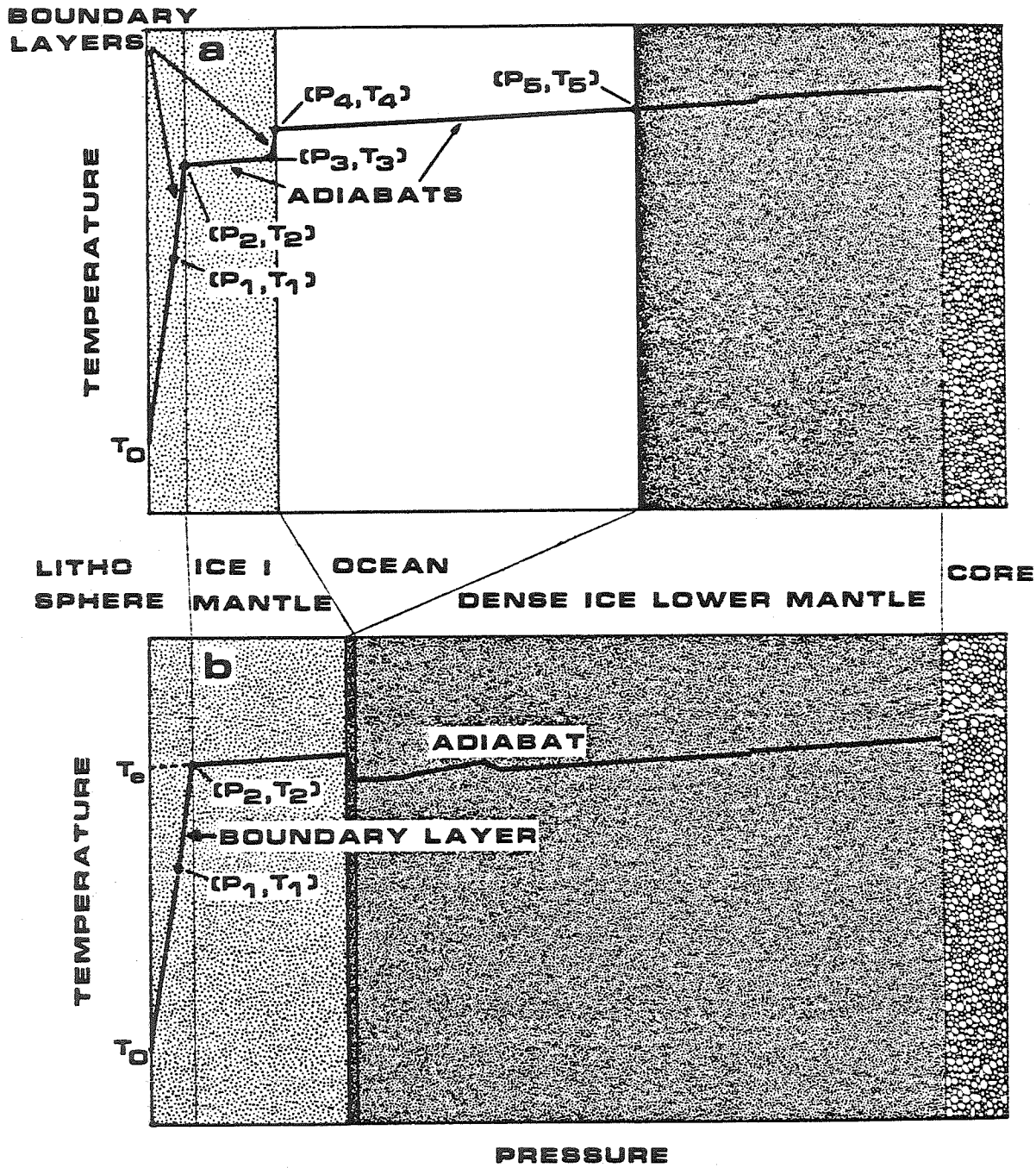

BOUNDARY LAYERS
a
(P4,T4)
(P5,T5)
(P3,T3)
ADIABATS
(P2,T2)
(P1,T1)
TEMPERATURE
TO
LITHO SPHERE
ICE I MANTLE
OCEAN
DENSE ICE LOWER MANTLE
CORE
b
ADIABAT
Te
(P2,T2)
BOUNDARY LAYER
(P1,T1)
TEMPERATURE
TO
PRESSURE

Figure 2.3. Approximate phase diagram of H_2O (neglecting solute effects on the melting point) with horizontally averaged Ganymedotherms. Depth scale is approximate, based on the density of liquid. Roman numerals indicate phases of ice. Ganymedotherms correspond to times indicated on the schematic heat flux history (inset): (a) Onset of freezing of ices I and VII. (b) Onset of convection in ice I. (c) Just before bridging of ocean. Ice I is colder than ice III because of boundary layer. (d) Just after bridging of ocean. Ice I is of equal entropy to and hotter than ice III. (e) Late stage of quasi-equilibrium with core heat output. Ice II region is avoided because of large latent heat.

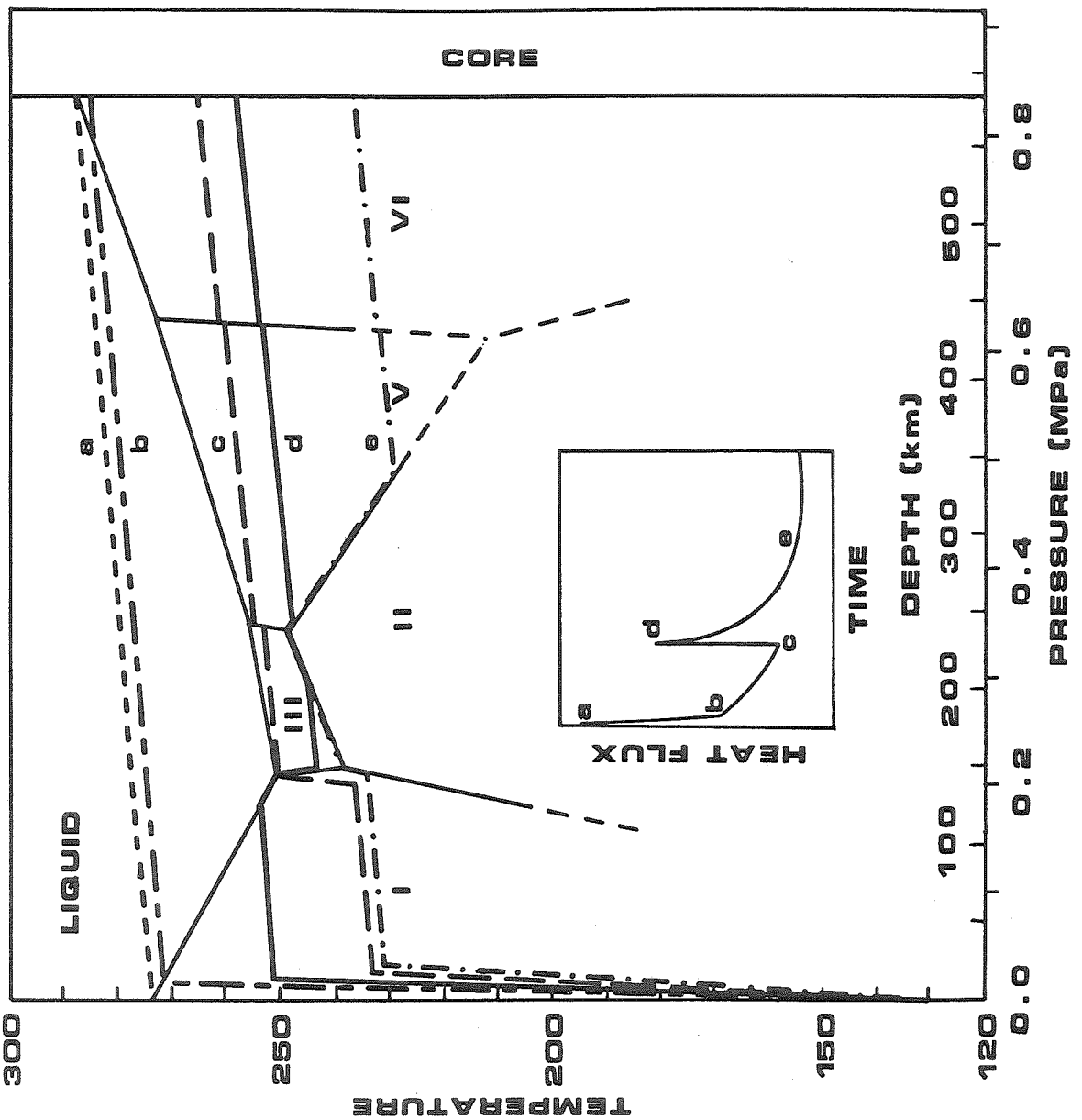

CORE
LIQUID
I
II
III
V
VI
a
b
c
d
e
HEAT FLUX
TIME
DEPTH (km)
100
200
300
400
500
PRESSURE (MPa)
0.0
0.2
0.4
0.6
0.8
TEMPERATURE
300
250
200
150
120

Table V. Thermodynamic Properties of the Phases of Water

Quantity	Units	Phase				
		Liquid	I	III	V	VI
ρ^a	$kg\,m^{-3}$	1000	917	1160	1270	1310
L^b	$kJ\,kg^{-1}$	...	284	235	277	294
C	$J\,kg^{-1}\,K^{-1}$	4180	$1925\,(T/250)^c$	[d]	[d]	[d]
$T_{m0}{}^e$	K	...	273.2	243.6	234.7	227.8
dT_m/dP^e	$K\,GPa^{-1}$	...	−106.3	35.97	61.65	72.7
$\alpha T/\rho C$	$K\,GPa^{-1}$	18.0	$21.19\,(T/250)^c$	[d]	[d]	[d]
α	K^{-1}		$1.56 \times 10^{-4}(T/250)^c$			
k	$W\,m^{-1}\,K^{-1}$		$2.60\,(250/T)^c$			
κ	$m^2\,s^{-1}$		$1.47 \times 10^{-6}(250/T)^{2\,c}$			

[a] At the lower-pressure triple point; cf Table IV.
[b] Latent heat of fusion, averaged over the pressure range of stability.
[c] One-parameter fit to variation from 130–273 K.
[d] Data unavailable; ice I value used.
[e] Coefficients of approximate liquidus: $T_m \simeq T_{m0} + dT_m/dP$.

given by the phase relations, $T(P)$ is assumed to be adiabatic in both the ocean and the lower mantle and in the upper mantle is determined by the solid-state convection equations discussed below. Thus, the rate of cooling everywhere, and all rates of phase conversion, depend on the rate at which P_5 changes:

$$\left.\frac{dE}{dt}\right|_{sens} + \left.\frac{dE}{dt}\right|_{lat} = f(P_5)\frac{dP_5}{dt}. \tag{2.9}$$

The function f depends on the heat capacities and latent heats, the equations of the adiabats and phase boundaries, and the mass distribution M(P). In a preliminary version of this work (Kirk and Stevenson 1983), the phase boundaries and adiabats of pure H_2O (Figure 2.3) and $\frac{dM}{dP}$ were approximated as linear functions of P, and L and C for each phase were taken as constant, allowing f to be written explicitly. Here, we retain these approximations but include the effects of solutes on the phase diagram in an approximate way. All solid phases were assumed subject to the same pressure-independent melting-point depression, proportional to the solute concentration:

$$\Delta T_{sol} = \frac{dT_m}{dx} x_0 \frac{(M(P_5) - M(P_4))_{t=0}}{(M(P_5) - M(P_4))_t}. \tag{2.10}$$

Here $M(P_5) - M(P_4)$ is the mass of liquid water remaining, and x_0 is the initial solute concentration expressed as a mole ratio. We adopt a cryoscopic constant $\frac{dT_m}{dx} = 116\,\mathrm{K}$, the value for ice I in the presence of NH_3 at zero pressure (Weast 1976, p. D221) and an initial mole fraction of dissolved nebular ammonia $x_0 = 10^{-3}$ taken from Lunine and Stevenson (1982). This value, based on equilibrium with a protojovian nebula, is a rough upper limit, since some ammonia could have been lost to space as the nebula dissipated, depending on the rapidity with which this occurred. We have also considered the influence of salts. Assuming the core rock contains chlorine in cosmic ratio to silicon (Anders and Ebihara 1982), all of which enters the ocean in the form of NaCl (sodium is not limiting, and the other halogens are much less abundant), we find that the melting point depression will be increased by about 12% over that for ammonia alone (Weast 1976, pp. D221, D252–D253). Again, this is a lower limit, since the core is unlikely to be completely leached. The effect of carbonates and sulfates should be even less, in view of their limited solubilities. Given the uncertainty in the ammonia abundance, we have not attempted to include the effects of salts or the variation of the effect of ammonia with pressure and concentration (Johnson *et al.* 1985).

2.3 Surface Heat Flux

Next we turn to a consideration of the heat lost to the exterior. A parameterized convection model was used to estimate the heat flow passing through the ice I mantle to the surface; this region, being the coldest and most viscous, controls the rate of heat loss from the deep interior, unless the viscosity of a high-pressure phase of ice is enormously larger at the same homologous temperature.

The convection model used yields the approximate temperature distribution and heat flow in the upper mantle as a function of its thickness and the temperatures at the two boundaries. The "surface" temperature T_0 was held fixed at 130 K, appro-

priate to the boundary between insulating regolith and subjacent compact, thermally annealed ice (Passey and Shoemaker 1982); the model results were in any case found to be insensitive to variation of T_0 of the order of 10 K. The bottom temperature T_4 and pressure P_4 depend only on P_5 (and the phase diagram). Between these limits we divide the mantle into four regions, with the temperature distribution in each approximated by a linear function of pressure. From the surface inward, these are (Figure 2.2): an immobile cap in which heat is transported by conduction, the upper boundary layer of the convective region (also conductive), an adiabatic region of vigorous convection, and a lower boundary layer (Turcotte and Oxburgh 1967). The pressure-temperature coordinates of the three interfaces between these regions are chosen to obey the following constraints. First, the total heat conducted through the cap and each of the two boundary layers must be equal (ignoring the small difference due to cooling of the ice I itself). Because these regions occupy less than the outermost 5% of the body's radius, the flux *per unit area* F may be taken as approximately constant, as may the gravitational acceleration g. The variation of thermal conductivity k with temperature is important, however. We take as an approximate fit to data in Hobbs (1974) $k_\mathrm{I} = \frac{k_0}{T}$ with $k_0 = 650\,\mathrm{W\,m^{-1}}$. Then the heat flux is:

$$F = \begin{cases} \dfrac{k_0 \rho_\mathrm{I} g}{P_1} \ln\left(\dfrac{T_1}{T_0}\right), & \text{in the lithosphere,} \quad (2.11) \\ \dfrac{k_0 \rho_\mathrm{I} g}{P_2 - P_1} \ln\left(\dfrac{T_2}{T_1}\right), & \text{in the upper boundary layer,} \quad (2.12) \\ \dfrac{k_0 \rho_\mathrm{I} g}{P_4 - P_3} \ln\left(\dfrac{T_4}{T_3}\right), & \text{in the lower boundary layer.} \quad (2.13) \end{cases}$$

Next, the thermal gradient in the middle region is constrained to be adiabatic:

$$\frac{T_3 - T_2}{P_3 - P_2} = \frac{\alpha_\mathrm{I} T}{\rho_\mathrm{I} C_\mathrm{I}}, \quad (2.14)$$

where α_I is the volume coefficient of thermal expansion and C_I the heat capacity at constant pressure of ice I. The adiabatic gradient is so much smaller than that in the

conductive regions that its variation with temperature is unimportant; the righthand side of (2.14) was evaluated at the fixed temperature of 250 K.

The remaining constraints involve the Rayleigh numbers in the boundary layers. The Rayleigh number in each boundary was assigned the critical value for convection:

$$Ra_{12} \equiv \frac{\alpha_{\rm I}(T_2 - T_1)(P_2 - P_1)^3}{\rho_{\rm I}^2 g^2 \kappa_{\rm I} \eta_{\rm I}} = Ra^*, \tag{2.15}$$

and

$$Ra_{34} \equiv \frac{\alpha_I(T_4 - T_3)(P_4 - P_3)^3}{\rho_{\rm I}^2 g^2 \kappa_{\rm I} \eta_{\rm I}} = Ra^*. \tag{2.16}$$

Here $\kappa_{\rm I}$ is the thermal diffusivity and $\eta_{\rm I}$ is the dynamic viscosity of ice I, evaluated at the mean temperature of the boundary layer (Booker 1976). The critical Rayleigh number Ra^* depends on the geometry of the convecting region and the viscosity variation across it but is typically of the order of 10^3 (Chandrasekhar 1961; Booker and Stengel 1978). A value of 10^3 was used throughout, but note that any uncertainty in Ra^* may be factored into η, the effect of varying which we have investigated in detail.

Finally, the Rayleigh number of the upper boundary layer was assumed to be a maximum with respect to its thickness:

$$\left.\frac{\partial Ra_{12}}{\partial P_1}\right|_{F,P_2} = 0. \tag{2.17}$$

Physically, this means that convection cannot become any more vigorous (Ra cannot increase) by penetrating closer to the surface. This assumption is similar in spirit if not equivalent to that used by Reynolds and Cassen (1979). We consider it physically more reasonable than the widely used criterion of a fixed "cut-off" viscosity at the base of the lithosphere (*e.g.*, Ellsworth and Schubert (1983) for the Saturnian satellites, and numerous papers on the terrestrial planets), which has been shown experimentally to be invalid for large viscosity variations (Nataf and Richter 1981).

Equations (2.11) to (2.17) must be completed by a prescription for the viscosity in the two boundary layers. We discuss the available information about the rheology of ice in the Appendix and conclude that the dominant mechanism of creep in the upper boundary layer is volume diffusion of vacancies (Nabarro-Herring creep). The resulting viscosity can be represented by a homologous-temperature formulation (A.1) with a pre-exponential factor η_0 dependent on the ice grain size d but not on the stress (numerical values are given in Table VI). In the lower boundary layer $\frac{T}{T_m} \simeq 0.96$, but the creep is enhanced by melting at grain boundaries, and hence it is reasonable to let $\eta = \eta_0$ there. We attempt in the Appendix to estimate d, and hence L_0, but the results are at best conjectural. We will therefore present a range of models with $10^{12}\,\mathrm{Pa\,s} \lesssim \eta_0 \lesssim 10^{15}\,\mathrm{Pa\,s}$; our best estimates lie in the middle of this range.

Together with the temperature-dependent viscosity (A.1), equations (2.11) to (2.17) completely constrain the heat flux F and the (P,T) coordinates of the three tiepoints on the Ganymedotherm in the ice I layer, given P_4 and T_4 corresponding to a given value of P_5. With F a uniquely if implicitly determined function of P_5, equation (2.1) represents a nonlinear first-order ordinary differential equation for P_5 as a function of time. This equation was integrated numerically, equations (2.11) through (2.17) being solved by iteration at each timestep.

2.4 Convection Across the Residual Ocean

Here, we describe only the modifications that must be made to the energy balance equation (2.1) when the remaining liquid water ocean is sufficiently thin that it is no longer a barrier to throughgoing solid-state convection. Details of the transition to throughgoing convection — its time of occurrence, extreme rapidity of onset, and consequences — will be addressed in Section 3 below.

Not all the terms in (2.1) are affected when a single solid-state convective cell is set up. The core heat output given by (2.4) is entirely unchanged. The rates

of release of sensible and latent heat are still formally described by (2.7) and (2.8) but the latter is now sufficiently small to be ignored, inasmuch as the amount of remaining liquid is small and nearly constant. The shape of the Ganymedotherm is qualitatively different, however, changing the explicit formulation of (2.7) as well as the parameterized convection model. The lithosphere and boundary layer beneath it are as before, but beneath them is now a single region of solid-state convection extending adiabatically down to the core (Reynolds *et al.* 1981). It is important to note that this adiabat is *not* a curve of continuous thermal gradient. Because of the latent heats of transition between solid phases, adjacent solid phases with the same entropy must differ in temperature. The Ganymedotherm is thus offset along the phase boundaries it crosses — in particular, the bottom of the ice I is now of the order of 10 K warmer than the subjacent ice III. In contrast, when convection proceeded separately in the two layers, the adiabatic region of ice I was ~ 8 K colder than the ice III across the boundary layer from it.

Once again, the Ganymedotherm is fully specified by a single subsurface tie-point — conveniently chosen now as T_e, the temperature of the ice I adiabat extrapolated to zero pressure. With this parameterization, (2.7) can be written in the form:

$$\left.\frac{dE}{dt}\right|_{sens} = g(T_e)\frac{dT_e}{dt}. \tag{2.18}$$

The function $g(T_e)$ is obtained explicitly under the same assumptions used for equation (2.9).

As before, the bottom of the lithosphere lies at (P_1, T_1) and the bottom of the boundary layer at (P_2, T_2). Including F, we thus have five unknowns for given T_0 and T_e. They are constrained by equations (2.11),(2.12), (2.15), (2.17), and

$$\frac{T_2 - T_e}{P_2} = \frac{\alpha_{\rm I} T}{\rho_{\rm I} C_{\rm I}}, \tag{2.19}$$

the requirement that (P_2, T_2) lie on an adiabat through T_e.

Equation (2.1) is now to be construed as an ordinary differential equation for T_e as a function of time. The initial condition on T_e is the requirement that the total thermal energy of Ganymede be the same immediately before and after the transition to through-going convection

$$\int_0^{P_{core}} C(P) T_{P_5}(P, t_{pulse}) \frac{dM}{dP} dP = \int_0^{P_{core}} C(P) T_{T_e}(P, t_{pulse}) \frac{dM}{dP} dP \qquad (2.20)$$

where t_{pulse} is the time at which the transition occurs and the subscripts on T indicate the two parameterizations of the Ganymedotherm. The gravitational potential energy released by the redistribution of warm and cold ice, giving rise to a net warming of the order of 0.03 K, is entirely negligible.

2.5 Results

Thermal evolution scenarios for a range of ice viscosities η_0 from 10^{12} to 10^{15} Pa s are represented in Figure 2.4. We plot surface heat flux, or equivalently surface thermal gradient, versus time for models with the nominal total radionuclide budget under two assumptions about its distribution. The models in Figure 2.4a assume that all radionuclides are sequestered in the core, whereas in 2.4b 30% are dissolved or suspended in the ocean. The dotted curves indicate the radiogenic heating rate (expressed as an equivalent heat flux over the surface area), which rises first conductively, then due to the onset of core convection in 2.4a, but is buffered by the oceanic contribution in 2.4b to a gentle, nearly monotonic decline. We will discuss only Figure 2.4a in detail; the curves in 2.4b are similar but less complicated. The shapes of the curves of heat flux versus time reflects the stages of evolution, and, with reference to the schematic version (Figure 2.3, inset) and its associated Ganymedotherms, the internal structure may be read from them. The case with $\eta_0 = 10^{12}$ Pa s is the most similar to the schematic version, and may be divided into the following stages:

1) Cooling by conduction through the ice I layer, until the onset of convection at 3.3×10^6 y (corresponding to Figure2.3b), outside the range of the plot.

This early conductive phase is of course the same for all viscosities, but its duration varies. The onset of convection in the models with $\eta_0 = 10^{13}$ and 10^{14} Pa s show clearly as breaks in the slope of the heat flux curve at 0.3 and 1.6×10^8 y, respectively. The change in slope is due to the weaker dependence of conductive heat flux on layer thickness compared to that for conduction, rather than to a change in the rate of freezing.

2) Thickening of the ice I layer with convective heat transport, until 2.3×10^8 y (Fig. 2.3c). The subtle breaks in slope at 0.6 and 1.8×10^8 y mark the times at which the dense ice phase freezing out at the bottom of the ocean changes. The consequent changes in L and $\frac{dT_m}{dP}$ cause a discontinuity in the function f of equation (2.9), and hence in the rate of thickening.

3) A sudden enhancement of the heat flux at 2.3×10^8 y, which we refer to as the "heat pulse," due to the reorganization of convection when the ocean becomes thin. Comparison of Figure 2.3c (before) and d (after) shows that the boundary layer at the base of the ice I has been eliminated and all the ices now lie on a single adiabat, convecting as a single cell. For given total internal heat (equation 2.20), this arrangement leads to a warmer subsurface and a higher heat flux. The heat pulse will be discussed in the next section, where we will show that its onset is essentially instantaneous on the timescale of Figure 2.4.

4) Decline of the heat flux towards equilibrium with the output of the core (Fig. 2.3e). This decline in heat flux is much more rapid than that which occurred before the heat pulse, because cooling is no longer buffered by the freezing of the ocean.

5) A rapid increase in the heat flux following the onset of core convection at 1.5×10^9 y, after which both the core and surface heat fluxes are nearly in

Figure 2.4a. Heat flux histories : surface heat flux or thermal gradient vs. time. All heat sources in the core (note onset of core convection at $\sim 1.5 \times 10^9$ y). Solid curves: Results of thermal evolution model, labeled by η_0 in Pa s. Dotted curve: Surface flux corresponding to the core flux with nominal radionuclide abundances (Table III). Dashed curve: Thermal gradient inferred from degree of viscous relaxation of craters (based on Passey 1982, Fig. 39, converted back to a thermal gradient using his assumed ice conductivity). Stated uncertainty is $\pm 0.5\,\mathrm{K\,km^{-1}}$.

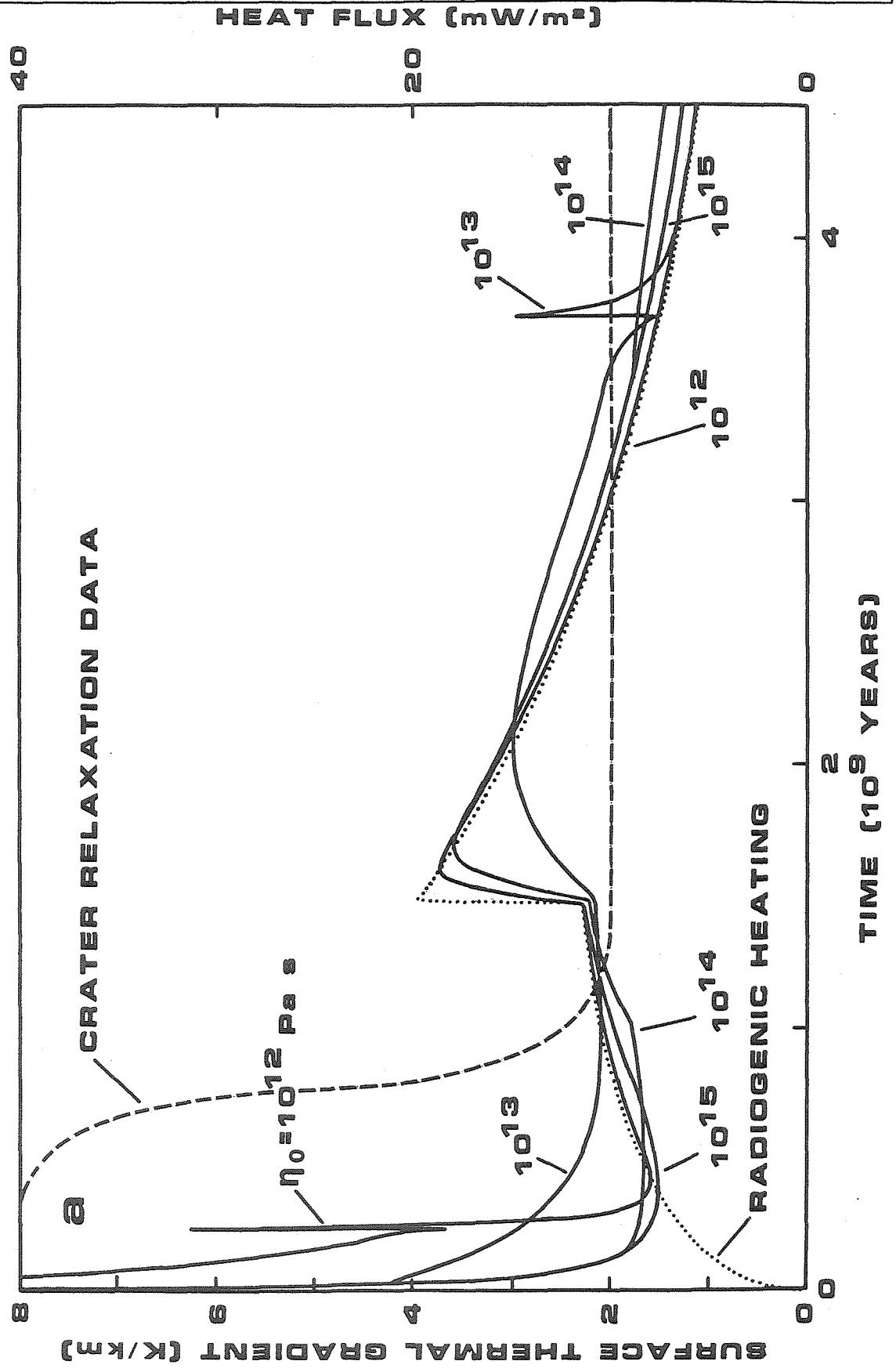

HEAT FLUX (mW/m²)
40
20
0
10^{13}
10^{14}
10^{15}
10^{12}
CRATER RELAXATION DATA
$\eta_0=10^{12}$ Pa s
a
10^{13}
10^{14}
10^{15}
RADIOGENIC HEATING
TIME (10^9 YEARS)
0
2
4
SURFACE THERMAL GRADIENT (K/km)
8
6
4
2
0

Figure 2.4b. Heat flux histories. As in part (a), but assuming 30% of all radiogenic heat sources to be dissolved or suspended in the ocean. Onset of core convection is suppressed.

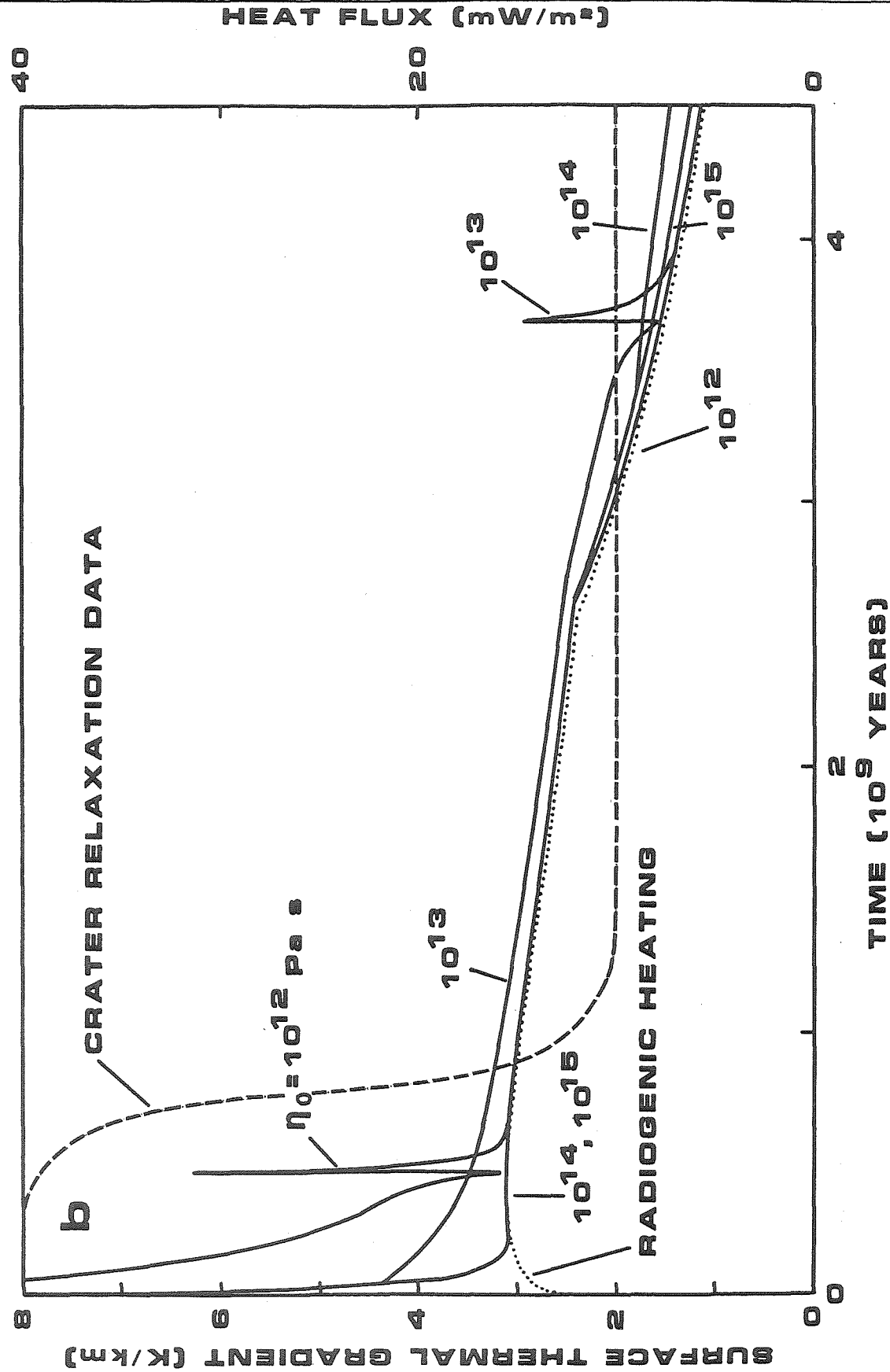

HEAT FLUX (mW/m²)
40
20
0
10^{13}
10^{14}
10^{15}
10^{12}
CRATER RELAXATION DATA
$\eta_0 = 10^{12}$ Pa s
10^{13}
$10^{14}, 10^{15}$
RADIOGENIC HEATING
b
8
6
4
2
0
SURFACE THERMAL GRADIENT (K/km)
0
2
4
TIME (10^9 YEARS)

Figure 2.5a. Timing of heat pulse (principal contour interval 0.5×10^9 y; secondary contour interval 0.1×10^9 y) as a function of radiogenic heating and η_0 (or ice grain size d). All radionuclides are assumed to be in the core. Heavy contour marks the transition from an early pulse (before peak radiogenic heating) to a late pulse (after substantial radioactive decay). Probable ranges of parameter values are $0.8 \lesssim$ core heating $\lesssim 1.2$ and $10^{13} \lesssim \eta_0 \lesssim 10^{14}$ Pa s.

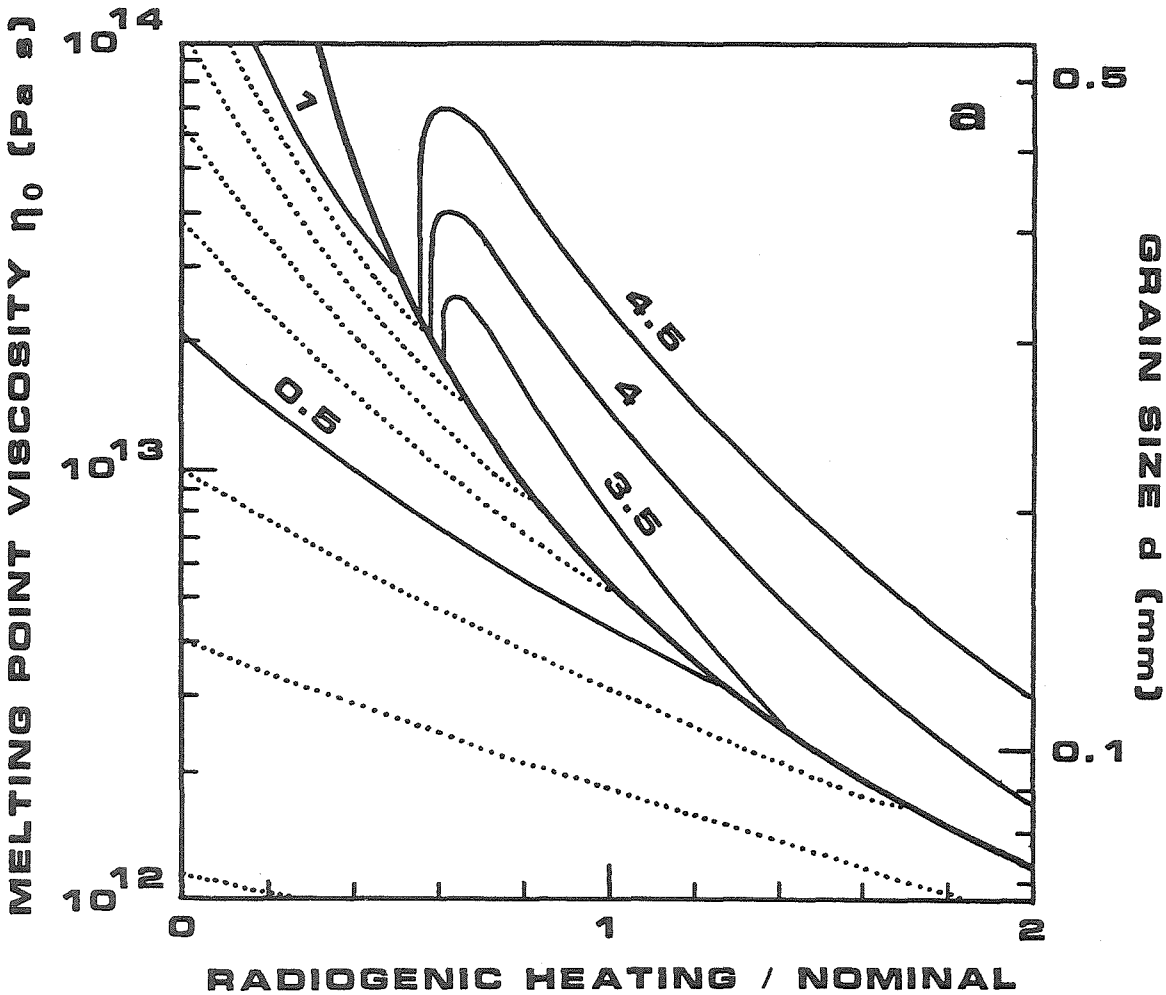
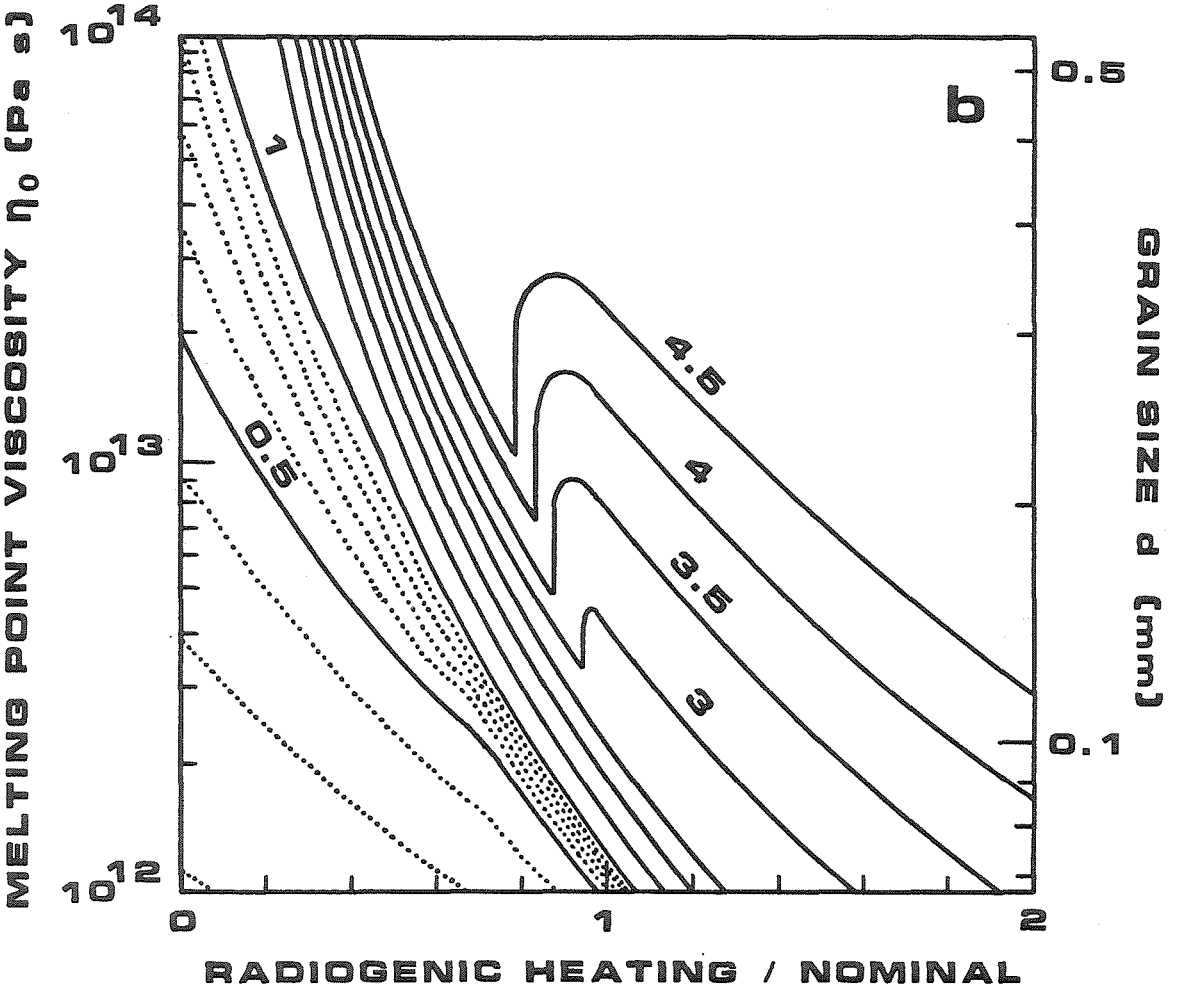

MELTING POINT VISCOSITY η_0 (Pa s)
10^{14}
10^{13}
10^{12}
GRAIN SIZE d (mm)
0.5
0.1
a
1
0.5
4.5
4
3.5
0
1
2
RADIOGENIC HEATING / NOMINAL

Figure 2.5b. Timing of heat pulse. As in part (a), but 30% of all radionuclides are assumed to be dissolved or suspended in the ocean. Radiogenic contribution to the heat budget (dotted curve) declines near-monotonically, broadening the singular contour into a transition zone between early and late pulse.


b
MELTING POINT VISCOSITY η_0 [Pa s]
10^{14}
10^{13}
10^{12}
GRAIN SIZE d [mm]
0.5
0.1
1
0.5
4.5
4
3.5
3
0
1
2
RADIOGENIC HEATING / NOMINAL

equilibrium with the decaying rate of radiogenic energy release. The heat flux history for $\eta_0 = 10^{13}$ Pa s is qualitatively different from that just described only in that, because the body cools more slowly, the heat pulse is delayed until *after* the rise and subsequent decline of the radiogenic heating. The response of the surface flux to the onset of core convection is slowed dramatically by the buffering effect of the still substantial ocean.

For $\eta_0 = 10^{14}$ Pa s, the heat pulse has not yet occurred (but will do so when the heating rate has declined sufficiently). Indeed, the rewarming due to radiogenic heating is so great as to stop convection in the ice I layer at 1.0×10^9 y. It does not resume until 3.4×10^9 y. . For viscosities of 10^{15} Pa s and greater, the thermal history is purely conductive.

Figure 2.5 summarizes a wide range of thermal evolution scenarios, with contours indicating the time at which the heat pulse (if any) occurs as a function of radiogenic heating and ice viscosity. As in Figure 2.4, models in part (a) assume all radionuclides are in the core, while those in (b) assume 30% of the total are in the ocean. The probable ranges of core heating ($\pm 20\%$) and η_0 (10^{13}–10^{14} Pa s) occupy the center of the figure. The location of the contours was found to be negligibly affected by changes of the order of 10% of the less important parameters such as T_0 and x_0.

The heavy contour in Figure 2.5a indicates a discontinuous transition from an early heat pulse, before the core heat output has reached its maximum value, to a late one, after substantial radioactive decay. The abruptness of this transition results both from the abrupt increase of radiogenic heating when the core begins to convect, and from the requirement that the residual ocean thin to a given extent before the heat pulse can occur. Increasing viscosity thus leads to a "'double bind": the residual ocean thickness is increased (temperature is increased) for any given heat flux, and

the thermal evolution is slowed, so that the heat flux comes close to equilibrium with the core output later and at a higher value. Placing some of the heat sources in the ocean allows their heat to be released earlier and hence flattens the radiogenic heating curve, broadening the transition from an early to a late heat pulse into a finite band of viscosities.

3. The Heat Pulse

The most striking feature of the thermal evolution scenarios presented in Section 2.6 is the existence in some cases of a sudden and dramatic enhancement of surface heat flow which we have termed the "heat pulse." As we have indicated above, the heat pulse is a consequence of the proposed transition from separate convection cells in the upper and lower ice mantles to a single cell convecting across the residual liquid layer. In this section we discuss the existence, nature, timing, and consequences of the transition.

3.1 Mechanism and Timing of the Heat Pulse

The horizontally averaged Ganymedotherms of Figure 2.3 show evidence of the convective reorganization of the heat pulse in the disappearance of boundary layer at the base of the ice I between (c) and (d). This fact serves as a starting point for discussion of the pulse, but to fully understand how it occurs we will have to take into account horizontal temperature variations. Consider first for simplicity the case (shown in Figure 2.3) of a pure H_2O ocean, which — provided the radiogenic heating is small enough — freezes completely at a well-defined time. The lower ice I boundary layer then comes in contact with ice III. The entropy of the bulk of the ice I above the boundary layer is much less than that of the ice III (not only is it colder, but heat is absorbed when ice I transforms to ice III) so that one might naively expect an overturn, bringing the suboceanic ice with its higher potential temperature upward

to replace supraoceanic ice until a single convective adiabat is established.

Bercovici *et al.* (1986) have recently suggested on the basis of a linearized stability analysis of the ice I-III system that such an overturn will not in fact occur. They calculate the required temperature gradient for the onset of convection separately in the ice I and III fields and compare it to that for a single cell convecting across the phase boundary. The presence of a phase transition is known on the basis of studies of the Earth's mantle (Schubert and Turcotte 1971; Schubert *et al.* 1975) to affect convection in three ways: through thermal buoyancy generated by the release of latent heat, by vertical migration of the phase boundary in response to this temperature change, and by similar migration due to advection of heat along the preexisting thermal gradient. When $\frac{dT_m}{dP}$ is negative, the first effect is destabilizing, but those involving phase boundary migration are stabilizing; for ices I and III the net stabilizing effect offsets the advantage of a single large convective cell, and convection first becomes possible separately in the ice I and III layers.

We argue, however, that this linearized analysis of the ice I-III boundary is not relevant to the problem of the heat pulse. First, the Ganymedean mantle is highly supercritical with respect to *either* of the stability criteria calculated by Bercovici *et al.* (1986), so that, *if it can be initiated*, convection in a single cell may certainly persist. (As η_0 is increased, the equilibrium residual ocean thickness becomes too large to allow a heat pulse long before the linearized stability limit for a hypothetical single convective cell is reached.) This large supercriticality also makes possible convection through the region of adverse thermal gradient at the I-III interface. Second, we believe that a mechanism for triggering the overturn, involving additional phases, may exist. We outline this multistep mechanism, illustrated in Figure 3.2, below.

We believe that overturn may be initiated by the formation of ice II in the cold descending plumes of the upper mantle and hence may depend critically on the

horizontal temperature variations due to *finite amplitude* convection. The instability we envision, with the ice II rich plumes descending through the boundary layer, across the ocean, and on into the lower mantle (ice II is denser than ice III) is complex and novel, and we have treated it only in an approximate way. Because the background state is one of finite amplitude convection, numerical modeling would be needed to answer conclusively whether ice II formation leads to transoceanic flow or merely increases the vigor of convection. We have addressed only the much simpler problem of buoyant force balance, comparing the density of a column of ice with some initial thermal structure in which ice II is forming (possibly aided by an initial downward displacement δz) with that of a reference column following the horizontally averaged Ganymedotherm (including solute effects). A *necessary* condition for instability with respect to transoceanic flow is then that the test column be heavier than the reference column (and that this negative buoyancy increase with continued downward flow). Whether this condition is also *sufficient* for instability depends on the nature of the viscous stresses, hence the need for numerical modeling. We make the following assumptions regarding the viscous stresses: first, if no ice II forms, clearly the cold plume does not flow down through the ocean. Its purely thermal buoyancy is resisted by the viscous stresses associated with ordinary convection, and hence may be removed from consideration. Second, we suppose (optimistically) that the much larger negative buoyancy due to phase changes cannot be supported by ordinary convection, and hence must lead to a new, ocean-crossing flow. Under these assumptions, given the residual ocean thickness z_r (which completely specifies the reference column structure; cf. Section 2.3), and a plume of initial temperature T_p, it is straightforward to calculate the distance δz that the plume must be perturbed downward before instability is possible.

We perform the perturbation adiabatically, offering in justification the follow-

ing self-consistency argument. First, although we have not considered the dynamics of the instability in detail, its rate will be determined by a balance between viscous and buoyant forces, so we expect a timescale similar to that for a Rayleigh-Taylor instability (Chandrasekhar 1961):

$$\tau_{RT} = \frac{8\pi\eta}{\Delta\rho g\lambda} \tag{3.1}$$

(or half this if the fluid on one side of the interface is inviscid), where $\Delta\rho$ is the density contrast and λ a typical horizontal dimension of a disturbance of the boundary. For thermal conduction to be unimportant in reducing $\Delta\rho$ we require that τ_{RT} be less than the conductive timescale $\tau_c \sim \frac{\lambda^2}{4\kappa}$ (this requirement may also be expressed in terms of a Rayleigh number for the instability). Taking $\Delta\rho = 58\,\mathrm{kg\,m^{-3}}$, appropriate to ice II in the ocean, and $T = 230\,\mathrm{K}$, we obtain the requirement that $\lambda \gtrsim 1.0\left(\frac{\eta_0}{10^{13}\,\mathrm{Pa\,s}}\right)^{1/3}$ km. The cold mantle plumes that will first become unstable are of similar thickness to the lower boundary layer, $\sim 2\left(\frac{\eta_0}{10^{13}\,\mathrm{Pa\,s}}\right)^{1/3}$ km, so that loss of the dense ice II by conductive warming is unimportant. The timescale (3.1) for the downward instability is $\simeq 10^3\,\mathrm{y}$.

Figure 3.1 shows the perturbation δz required for instability as a function of T_p and z_r; of especial interest is the contour $\delta z = 0$. A plume with given T_p will (under our assumptions about the instability process) spontaneously give rise to a flow into the lower mantle once the residual ocean thins to this value of z_r. In the thermal histories presented in this paper we took the crossing of this contour as the criterion for onset of the heat pulse. The temperature of the cold plume just above the lower boundary layer was estimated from the horizontally averaged temperatures by comparison with numerical convection calculations by Jarvis (1984) which suggest that over a wide range of Rayleigh numbers one has approximately:

$$T_p \simeq T_3 - \frac{T_2 - T_1}{3}. \tag{3.2}$$

The value of this expression depends on z_r through the parameterized convection scheme outlined in Section 2.3. The dotted curves in Figure 3.1 illustrate $T_p(z_r)$ trajectories, assuming $\eta_0 = 10^{15}$, 10^{14}, and 10^{13} Pa s, respectively.

The possibility of an instability triggered by the finite amplitude perturbation due to a large impact deserves comment, especially since it played a central part in our earlier description of the heat pulse based on an erroneous analysis of the instability of the ice I-III boundary (Kirk and Stevenson 1983). We argued on the basis of an analysis of the viscous relaxation eigenmodes of the floating ice I layer that relaxation of large impact craters rapidly creates substantial isostatically compensated topography on the underside of the upper mantle, then smooths it away on a longer timescale. Essentially the full 85% of the top-surface topography dictated by isostasy is transferred to the bottom for craters with diameters $D \gtrsim 500$ km, falling off to 10% for a 350 km crater. Under our previous assumptions, such impact-induced finite amplitude instability was the only way of triggering the heat pulse. We still envision this mechanism as operating, but it will have only a small and quantitative effect on the calculated pulse time. As Figure 3.1 indicates, for (say) $\eta_0 = 10^{13}$ Pa s spontaneous instability occurs when $z_r \simeq 7$ km; the largest plausible crater, with $D = R_{Ganymede}$, perturbs the bottom of the mantle by only 9.5 km, triggering the heat pulse if $z_r \lesssim 14$ km. In most cases, the residual ocean thins from 14 to 7 km very rapidly. The impact mechanism permits a significant hastening of the heat pulse only for those special choices of ice viscosity and radiogenic heating that lead to a prolonged state of quasi-equilibrium with z_r in this intermediate range.

It is important to understand that the instability discussed so far leads only to a downward flow of ice from the upper mantle as illustrated in Figure 3.2a. As yet there is no return flow of ice made possible by instabilities in the lower mantle, so the residual ocean is lifted from its equilibrium position. This rather remarkable

Figure 3.1. Criterion for occurrence of the heat pulse . Solid curves are contours of downward perturbation δz of the ice I layer (in km) required to trigger instability by ice II formation, as a function of cold mantle plume temperature T_p and residual ocean thickness z_r. Bend in contours indicates formation of ice III (stabilizing) at high T_p and low z_r. Large impacts could lead to $\delta z \lesssim 9.5$ km. Dotted curves are trajectories of T_p vs. z_r in the parameterized convection model, assuming (left to right) $\eta_0 = 10^{14}$, 10^{13}, and 10^{12} Pa s. Time of heat pulse in our models (Figs. 2.4, 2.5) was taken as the time of crossing the contour $\delta z = 0$.

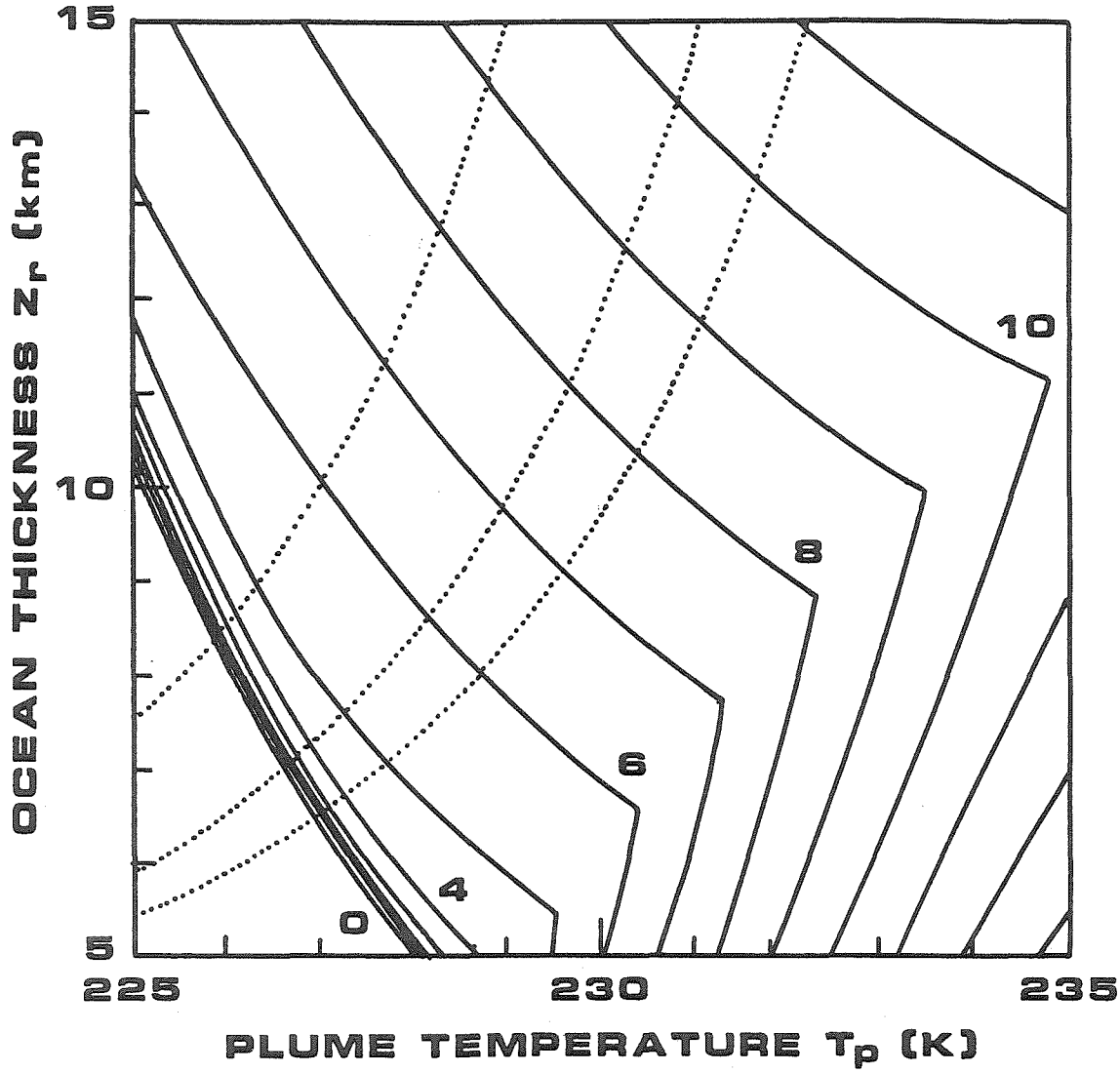

15
10
5
OCEAN THICKNESS z_r (km)
0
4
6
8
10
225
230
235
PLUME TEMPERATURE T_p (K)

Figure 3.2a. Cartoon history of the heat pulse mechanism : the onset of the pulse. Roughly to scale (with 2 : 1 vertical exaggeration and planetary curvature removed. Ellipse corresponds to a circle of 10 km radius in Ganymede.) Gap indicates 90 km omitted from the column for clarity. Heavy arrows indicate convective flow in the upper and lower mantles. Dashed line marks the base of the thermal lithosphere (including boundary layer). A cold descending ice I plume has reached the ice I-II phase boundary, leading to instability and runaway downward flow through the ocean and lower mantle.

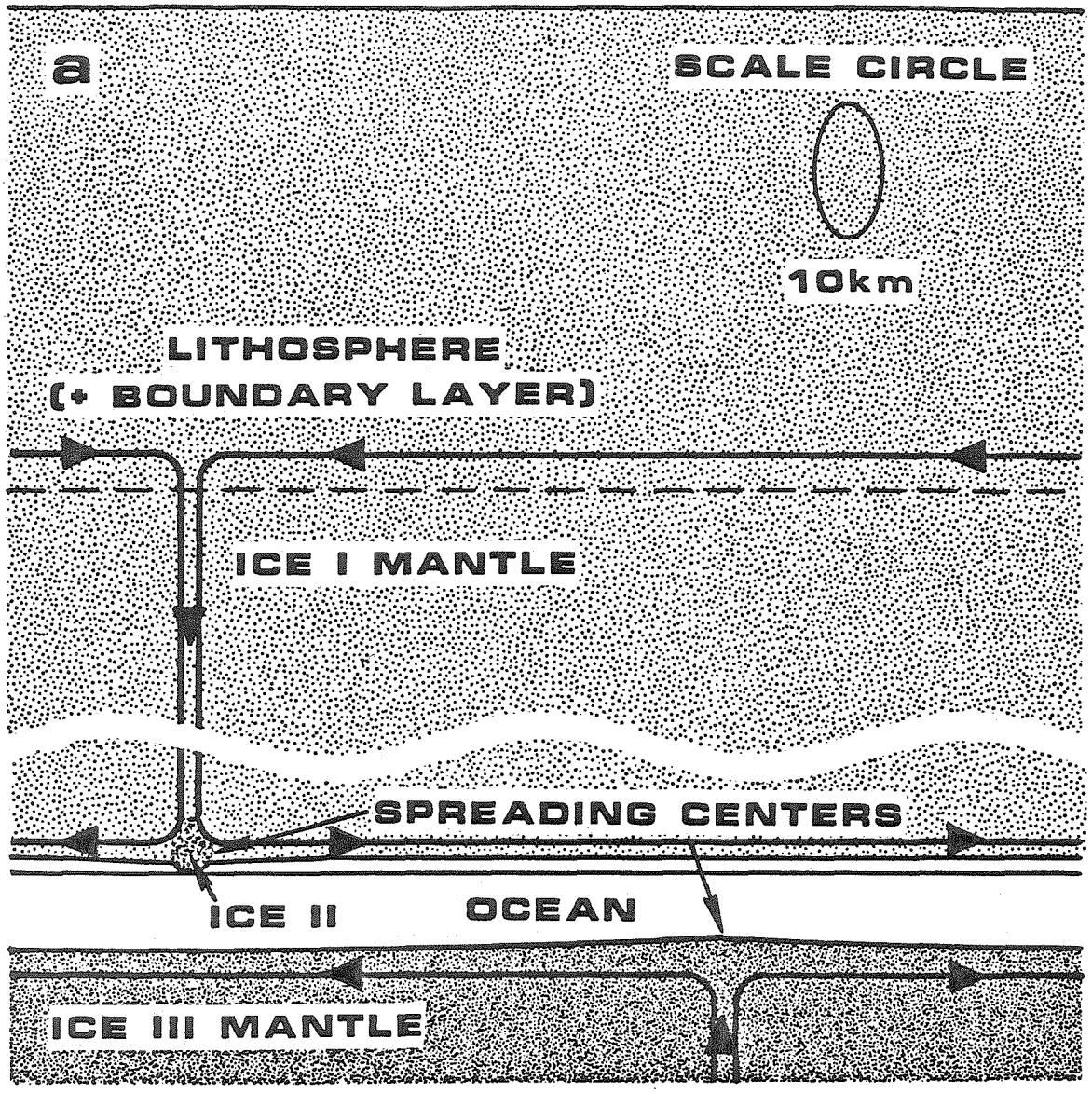

a
SCALE CIRCLE
10km
LITHOSPHERE
(+ BOUNDARY LAYER)
ICE I MANTLE
SPREADING CENTERS
ICE II
OCEAN
ICE III MANTLE

Figure 3.2b. Cartoon history of the heat pulse mechanism: triggering of return flow. After $\sim 10^5$ y the top of the ice III is lifted to the L-I-III triple point and converts to ice I plus melt, buoyant in the ocean.

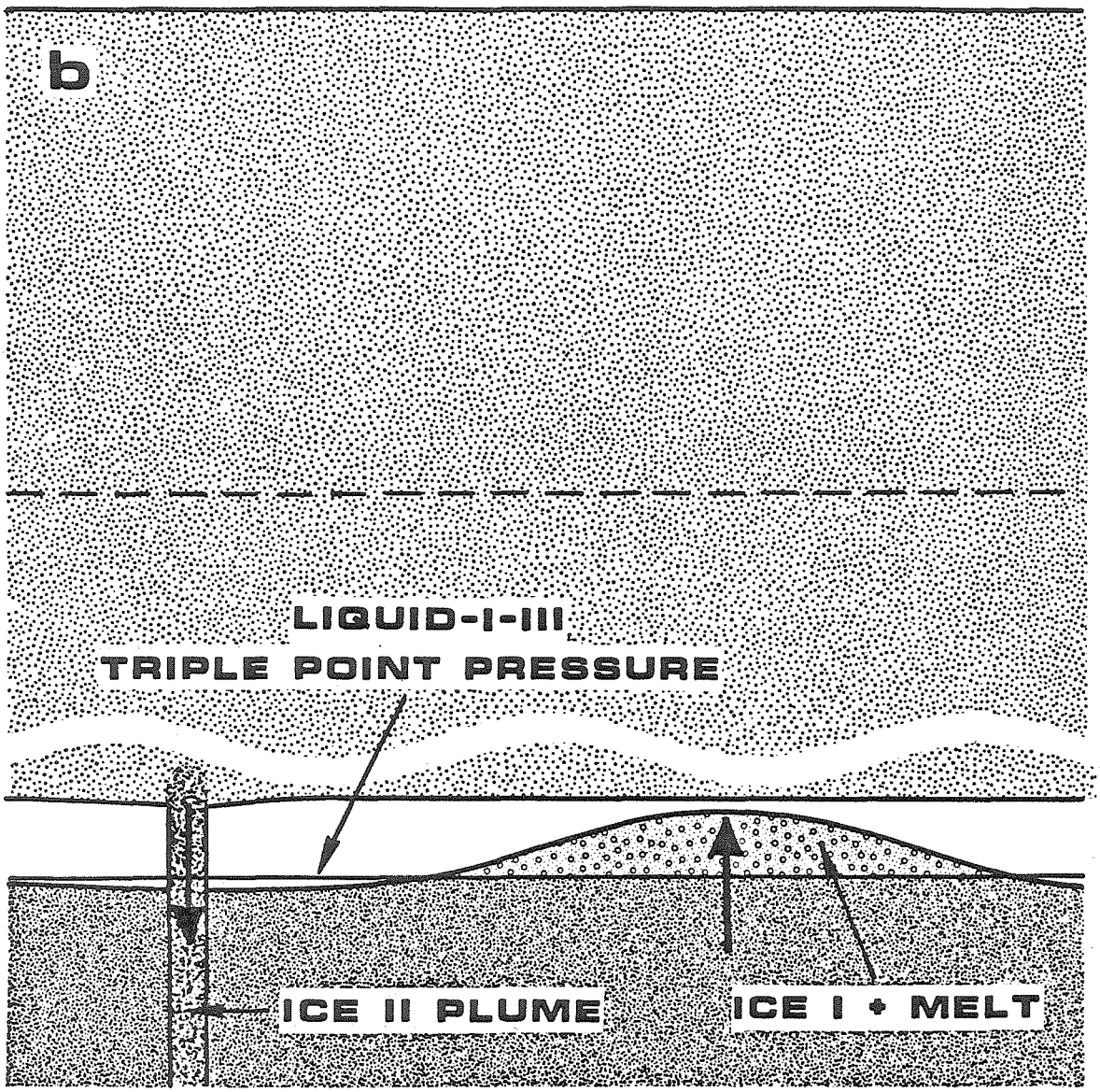

b
LIQUID-I-III
TRIPLE POINT PRESSURE
ICE II PLUME
ICE I + MELT

Figure 3.2c. Cartoon history of the heat pulse mechanism: initial warm ice diapirism. The weight of the displaced ocean pushes the ice plus melt diapir upwards to the base of the lithosphere in $\sim 10^3$ y. Loading and flexure of the lithosphere causes extension fracturing at the surface. The upward flow drops the ocean back toward its equilibrium position.

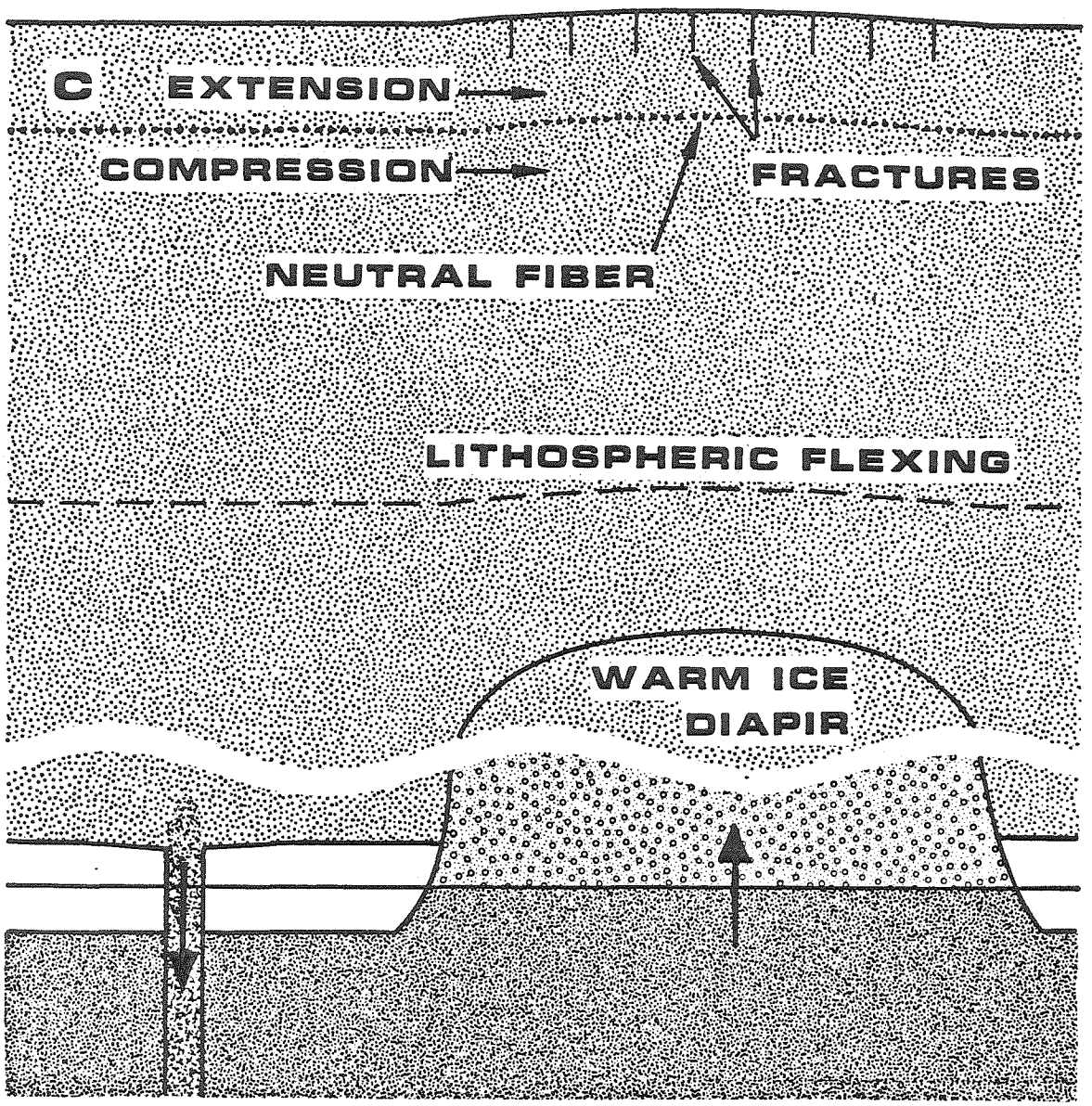

C
EXTENSION
COMPRESSION
FRACTURES
NEUTRAL FIBER
LITHOSPHERIC FLEXING
WARM ICE
DIAPIR

Figure 3.2d. Cartoon history of the heat pulse mechanism: initial resurfacing. The diapir reaches the near-surface by thermal softening in $\sim 10^6$ y and is released onto the surface by a small impact crater.

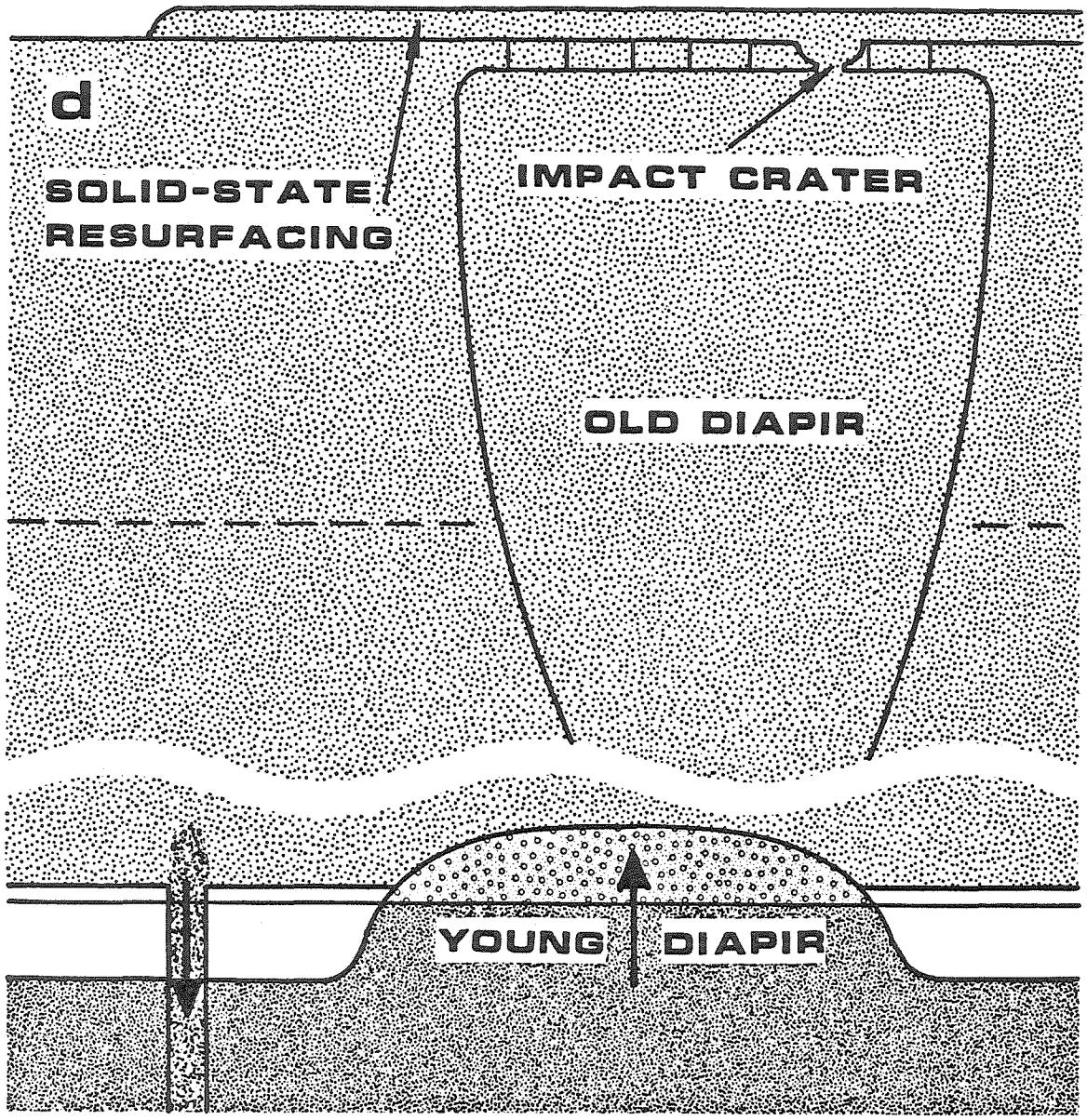
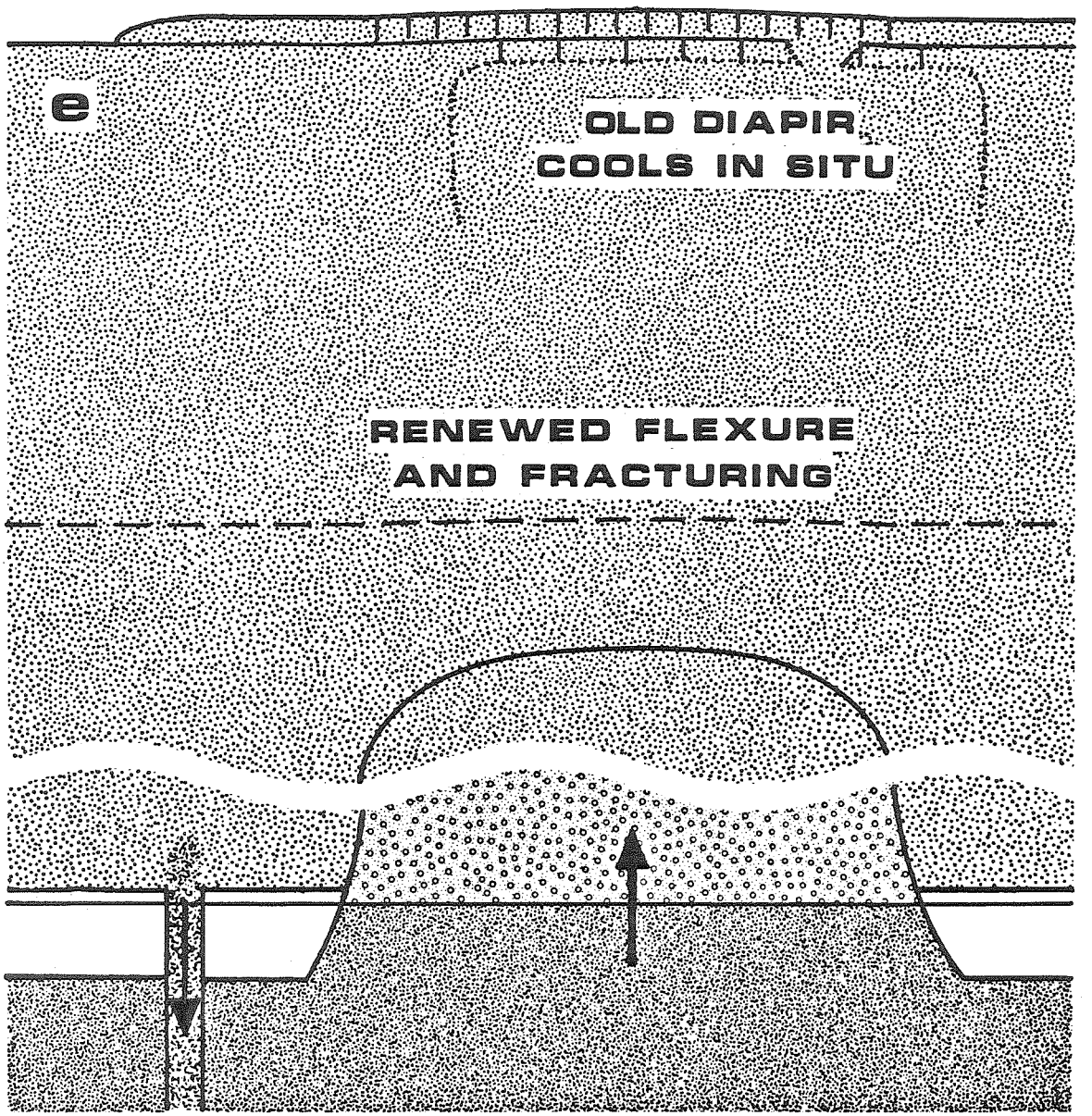

d
SOLID-STATE
RESURFACING
IMPACT CRATER
OLD DIAPIR
YOUNG
DIAPIR

Figure 3.2e. Cartoon history of the heat pulse mechanism: fracture of the resurfaced terrain by a younger diapir. Episodic diapirism, resurfacing, and fracturing continue for some time, subsiding into steady-state convection across pockets of residual ocean.


e
OLD DIAPIR
COOLS IN SITU
RENEWED FLEXURE
AND FRACTURING

assertion requires justification. We therefore demonstrate that melting at the base of the ocean and freezing at its top cannot restore it to equilibrium, by comparing the upward and downward mass fluxes. Consider first a downward-going plume of ice of width x, with an excess density $\Delta\rho$. The velocity v at which it descends is governed by a balance between viscous and buoyancy forces and hence we expect it to be similar to that for Poiseuille flow:

$$v \simeq v_p = \frac{\Delta\rho g x^2}{12\eta}. \tag{3.3}$$

Each convective cell of the upper mantle, with width of the order of X, will thus give rise to a downward mass flux of the order of $\rho v x X$. The upward flux is limited by the power available at the base of the ocean to drive melting. Conduction of heat downward through the ocean is inefficient, so the melting rate reaches a maximum when convection in the ocean ceases, and the entire heat flux F from below is converted to latent heat. If the heat of fusion is L, the upward flux (integrated over one convective cell) is $\frac{X^2 F}{L}$. Comparing the two, we find that the ocean will be lifted if:

$$x \gtrsim \left(\frac{\eta F X}{12\Delta\rho g \rho L}\right)^{\frac{1}{3}}. \tag{3.4}$$

For $T = 250\,\mathrm{K}$, $F = 20\,\mathrm{mW\,m^{-2}}$, $\Delta\rho = 58\,\mathrm{kg\,m^{-3}}$, and $X = 160\,\mathrm{km}$, this becomes $x \gtrsim 0.1\left(\frac{\eta_0}{10^{13}\,\mathrm{Pa\,s}}\right)^{1/3}$ km. As indicated above, the plumes are ~ 20 times wider than this, and they may widen by entrainment of additional ice, so the criterion is well satisfied.

The next stage of the heat pulse begins when, after $\sim 10^5\,\mathrm{y}$ (or less if the descending plumes entrain additional ice) the top of the lower mantle has been lifted to the level of the liquid-ice I-ice III triple point (Figure 3.2b). The topographically highest parts of the ice III, the warm spreading ridges, now convert to ice I (plus possibly melt) and rise buoyantly through the ocean on a Rayleigh-Taylor-like timescale. With $\Delta\rho = 140\,\mathrm{kg\,m^{-3}}$ for ice I in liquid, $\eta = 10^{13}\,\mathrm{Pa\,s}$, and $\lambda = 160\,\mathrm{km}$ (once the

ascending ice begins to interact with the $\sim 160\,\mathrm{km}$ thick upper mantle, the dominant horizontal scale of overturn will be close to this value; we use it now for illustrative purposes) this is only $\sim 0.1\,\mathrm{y}$. The subsequent diapiric rise of the suboceanic ice is slower, depending on its buoyancy with respect to the upper mantle due to temperature and melt content, and hence on the thermodynamic state of the ice immediately below the ocean. We therefore discuss very briefly the nature of convection in the lower mantle.

In the early stages of cooling, the oceanic solute concentration is low and the ocean-mantle boundary lies only slightly below the liquidus temperature ($\Delta T_{sol} \ll 10\,\mathrm{K}$ in equation 2.10). The usual thermal plumes and boundary layers are then not possible in the lower mantle; warm ice from the core-mantle boundary layer must encounter the liquidus and run along it as a "wet adiabat," leading to a plume and upper boundary layer distinguished by their melt fraction rather than temperature. A $10\,\mathrm{K}$ temperature drop across the lowermost boundary layer (roughly to be expected if the viscosity of ice VI is similar to that of the other phases) leads to $x_L \simeq 0.1$ at the top of the ice III. At an opposite extreme, if concentrations yielding $\Delta T_{sol} \gtrsim 20\,\mathrm{K}$ could be reached, both a cold boundary layer and a hot, melt-free plume could exist. The state of affairs at the time of the heat pulse is intermediate, with $\Delta T_{sol} \simeq 10\,\mathrm{K}$, leading us to expect a partially molten plume, but a cold boundary layer, and the core of the suboceanic ice (from which the diapirs will primarily be derived) with $T \simeq T_{\mathrm{L-I-III}} = 251\,\mathrm{K}$ and $x_L = 0$. This material converts to ice I plus melt at the triple point, so that $T = T_{\mathrm{L-I-III}}$ and $x_L \simeq 0.09$ at the base of the diapirs. The resultant density is greater than that of the cold upper mantle at $T_3 \simeq 230\,\mathrm{K}$, but the weight of the displaced ocean — equivalently, buoyancy of the diapir with respect to the liquid layer — drives the ice upward farther. If x_L were constant, the diapirs would relax *toward* an equilibrium depth $\frac{z_r}{x_L} \simeq 70\,\mathrm{km}$ above the top of the displaced ocean.

Intergranular flow reduces the melt content only in a compaction boundary layer of negligible thickness (Richter and McKenzie 1984), but pressure-release freezing reduces x_L by $1.8 \times 10^{-3}\,\mathrm{km}^{-1}$, restoring positive buoyancy to a parcel by the time it is $\sim 30\,\mathrm{km}$ above the ocean. The diapir as a whole is thus positively buoyant at all times. We model its ascent in detail in the next section; it suffices here to note that it rises to the base of the lithosphere (Figure 3.2c) in only $\sim 300\,\mathrm{y}$ for $\eta_0 = 10^{13}\,\mathrm{Pa\,s}$. The total timescale for rise of the surface heat flux to its peak value is thus dominated by the $\sim 10^5\,\mathrm{y}$ needed to lift the residual ocean.

The upward flow of ice allows the ocean to drop towards the position dictated by thermodynamic equilibrium and hence reduces the "boost" applied to the partially molten diapir from below. We thus expect a *fluid*-dynamic equilibrium to be established, with the ocean displaced just enough to drive an upward flow equal to the downward flow in the cold plumes. With time, cooling of the lower mantle will reduce the melt fraction in the diapirs and hence the magnitude of this required displacement. When it reaches zero, the diapirs will be melt-free, but still at the triple point temperature. As we show below, they will then require $\sim 3000\,\mathrm{y}$ to rise through a 230 K mantle. This less vigorous diapirism in turn will grade imperceptibly into ordinary convection in a single multiphase cell, with pockets of residual ocean separated by ascending and descending plumes of ice.

3.2 Implications of the Heat Pulse

The most obvious significance of the heat pulse is that it leads to a high but rapidly declining heat flux relatively late in Ganymede's history, in qualitative accord with Passey's (1982) analysis of the cratering record. His reconstruction of the surface thermal gradient versus time, based on the degree of viscous relaxation of craters as a function of size and inferred age, appears in Figure 2.4. Although aspects of Passey's interpretation of the cratering record — for example, his assumptions about

the rheology of ice I, the unrelaxed shape of large craters in ice, and the interaction of a deep crater with the local temperature distribution — are admittedly problematical, his finding of a late, rapid decay of the heat flow is nonetheless suggestive. Relatively high heat fluxes can be produced at times of several $\times 10^8$ y by invoking a large additional energy source such as strong tidal dissipation (Cassen *et al.* 1982). For the heat flow to decline sharply at late times, however, requires that it do so smoothly from an untenably large early value, that the energy source maintaining the high heat flux "shut off" abruptly, or that the effective heat capacity of the body (*i.e.*, its rate of cooling, for a given heat flux) decrease abruptly. Just such a change in heat capacity occurs in our model when the latent heat reservoir of the ocean is used up. Indeed, the post-pulse heat flux declines somewhat faster than the 10^8 y timescale estimated by Passey. The rapid *onset* of the heat pulse is not necessarily in conflict with his results, since his constant high thermal gradient before $\sim 6 \times 10^8 y$ is purely conjectural. The record of this period has been effectively erased, and the strongest conclusion that can be drawn is that the period of high heat flux lasted long enough to relax away all older craters resolvable by the Voyager cameras.

Given the large uncertainty in thermal gradient ($\pm 0.5\,\mathrm{K\,km^{-1}}$) quoted by Passey and the aforementioned reservations about his analysis, we have not attempted to fit his thermal history with our model. Based on our results as summarized in Figures 2.4 and 2.5, however, a scenario with ~ 0.8 times the nominal core heat output and $\eta_0 \simeq 9 \times 10^{12}\,\mathrm{Pa\,s}$ would match both the time at which Passey's thermal gradient declines and the quasi-steady state value to which it tends, although the peak gradient during the heat pulse would be less than his $8\,\mathrm{K\,km^{-1}}$. The required viscosity is close to the threshold for transition to a much later heat pulse, and is therefore *a priori* somewhat unlikely. This difficulty could in principle be surmounted by compressing the cratering timescale, so that a given crater density corresponds to an earlier abso-

lute time than that assumed by Shoemaker and Wolfe (1982), but the present results hardly argue conclusively for such a change. Note, however, that even more drastic tampering with the timescale is required if one attempts to reproduce Passey's results with a monotonically declining heat flux.

A much more exciting possibility (and one that is to some extent mutually exclusive, since Passey finds evidence of the late decline of heat flux in both the cratered and grooved terrains) is that the heat pulse is directly connected with the formation of the grooved terrain. The diapiric rise of warm ice during the pulse could have potentially provided both the driving force for tectonism and a source of clean, buoyant H_2O for resurfacing (cf. Figure 3.2). In the remainder of this section we describe calculations of the ascent of warm ice diapirs, concentrating on the buoyant loading of the lithosphere from below and the ability of the diapir to penetrate nearly to the surface. We will derive a criterion for extensional fracturing in terms of the instantaneous loading rate (independent of the assumed cause of loading) in Section 4, and in Section 6 we will discuss the problem of evolving diapiric ice onto the surface from a shallow depth.

We have investigated numerically the rise of both the earliest, most vigorous diapirs, which are driven by the weight of the displaced residual ocean (see the previous section) and those at a later stage driven only by the thermal buoyancy due to their being $\Delta T \simeq 20\,\mathrm{K}$ warmer than the core of the ice I mantle. For purposes of scaling we consider only the latter case, which is both easier to deal with and more conservative. The maximum buoyancy force will be $\sigma_b \simeq \rho g \alpha \Delta T Z$ where $Z = 160\,\mathrm{km}$ is the approximate thickness of the ice I layer. With these values $\sigma_b \simeq 0.6\,\mathrm{MPa}$. As we show in Section 4, tensional stresses $\gtrsim 3\,\mathrm{MPa}$ are probably required to fracture the lithosphere; we also show that this level of horizontal stress can be achieved in a thin lithosphere provided σ_b is applied quickly enough. We therefore wish to estimate the

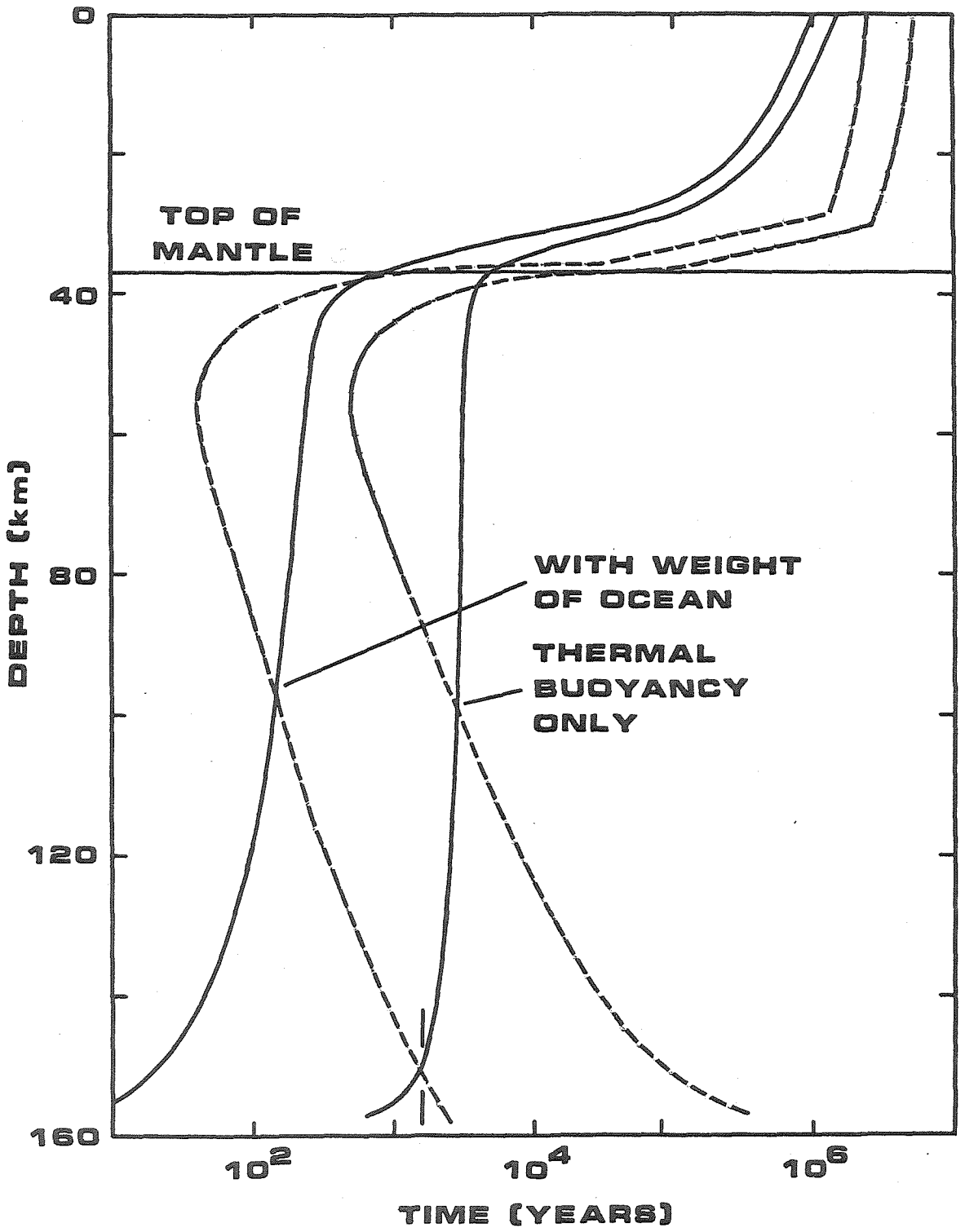
rates of rise and of loading of the lithosphere by a diapir. For historical reasons we will not express the latter result directly in terms of the loading rate $\sigma_{b,t}$ (the comma denotes a partial derivative) but in terms of the time constant of an error function loading history whose peak loading rate is $\sigma_{b,t}$. Writing this equivalent loading history as $\sigma_b(t) = \rho g \alpha \Delta T Z \frac{1}{2}(1 + \mathrm{erf}(ht))$, we seek $h^{-1} = \frac{\rho g \alpha \Delta T Z}{\sqrt{\pi}\sigma_{b,t}}$.

Consider a diapiric body of radius R ($\simeq \frac{Z}{4}$) and let $z(t)$ be the instantaneous depth of its top, while $z_2 = \frac{P_2}{\rho g}$ is the depth to the base of the upper boundary layer (cf. Fig. 2.2a). As the diapir begins its ascent, $z - z_2 \gg R$ and the flow is similar to Stokes flow, with deformation of the isoviscous mantle occurring in a layer of thickness $\sim R$ about the diapir. The stresses associated with this flow are comparable to the convective stress (A.3), so that the appropriate viscosity is that for diffusion creep, not for the nonlinear creep invoked in the lithosphere. The more viscous upper layers do not affect the flow significantly, nor is there any load on them.

As the diapir approaches the base of the boundary layer ($z - z_2 \lesssim R$, Figure 3.2c and e), the Stokes-flow pressure at the boundary rises. At the same time, the nature of the flow alters, increasing the drag and slowing the diapir. For $z \leq z_2$, the top of the diapir will be flattened, and the drag will be similar to that for pressing a flat plate into an exponential viscosity gradient (or equivalently into an isoviscous fluid but with an upper plate a distance $a = \sqrt[3]{6}L$ away, where L is the viscosity scale depth — the "toothpaste tube syndrome" flow of Section 5). In this limit the full buoyancy force on the diapir is transmitted to the "upper plate"; the lithosphere is fully loaded.

As z decreases still further, the increasing ambient viscosity slows the ascent, until eventually softening of the surroundings by heat diffusing out of the diapir becomes important (Morris 1982). The "toothpaste tube syndrome" has become a "china syndrome" (albeit upside-down), stopping only when the thermal boundary

Figure 3.3. Numerical models of rising warm ice diapirs for $\eta_0 = 10^{13}$ Pa s and surface thermal gradient 2 K km^{-1}. Solid curves: Time for diapir to rise to given depth. Dashed curves: Equivalent strain time h^{-1} for loading of the lithosphere (cf. Fig. 4.1). Right pair of curves assume constant $\Delta T = 20$ K between diapir and mantle, no melt. Left pair include melt ($x_L = 0.09$ at base, decreasing upwards because of freezing), thermal buoyancy, and the weight of a displaced residual ocean 7 km thick. Vertical line indicates the Rayleigh-Taylor timescale (3.1).


0
40
80
120
160
DEPTH (km)
TOP OF MANTLE
WITH WEIGHT OF OCEAN
THERMAL BUOYANCY ONLY
10^2
10^4
10^6
TIME (YEARS)

layer interacts with the surface, at $z \simeq 2\,\mathrm{km}$. The full diapiric load is transferred to the overlying ice by this flow mechanism as well, so that for $z < z_2$ the lithospheric loading increases only fractionally with time because of the increasing hydrostatic head of the rising diapir. The maximum loading rate must occur, therefore, when $z - z_2 \lesssim R$, and we can estimate it by using the velocity and stress gradient for Stokes flow. In terms of the equivalent strain time

$$h^{-1} = \frac{\rho g \alpha \Delta T Z}{\sqrt{\pi} |\partial \sigma_{rr}/\partial r|_{Stokes} U_{Stokes}}. \tag{3.5}$$

Evaluating the Stokes velocity and stress gradient for an inviscid sphere (Landau and Lifshitz 1959, pp. 63–70) of radius $R = \frac{Z}{4}$ and density anomaly $\alpha\Delta T$ in a medium whose viscosity is determined by diffusion creep at 230 K, we obtain the numerical result $h^{-1} \simeq 900 \left(\frac{\eta_0}{10^{13}\,\mathrm{Pa\,s}}\right)$ y. Figure 3.3 shows the results of a numerical integration of the rise of a cylindrical diapir (which has a slightly larger buoyancy) with drag based on Stokes flow in the mantle and Morris's (1982) calculations for a flat plate in the lithosphere. The minimum of h^{-1} is indeed $\sim 500\,\mathrm{y}$ for $\eta_0 = 10^{13}\,\mathrm{Pa\,s}$; it is achieved when $z - z_2 \simeq 0.5R$. The diapir reaches this depth, which is about 50 km for a surface thermal gradient of $2\,\mathrm{K\,km^{-1}}$, at a few times the Rayleigh-Taylor timescale (3.1). These values of h^{-1} may be compared with the fracture criterion derived in Section 4 and illustrated in Figure 4.1b. Under the fairly conservative rheologic assumptions embodied there, fracture to a depth of $\sim 1\,\mathrm{km}$ is possible for a thermal gradient of $2\,\mathrm{K\,km^{-1}}$.

The time for the diapir to reach the surface is much longer, being dominated by the time required to cross most of the boundary layer and lithosphere at the thermal softening velocity

$$U_{soft} \simeq \left(\frac{16\rho g \alpha \Delta T Z}{\eta(T_d) R^2} \left(\frac{\kappa}{\Theta}\right)^3\right)^{\frac{1}{4}}, \tag{3.6}$$

where $\Theta \equiv (T_d - T(z)) \frac{\partial \ln \eta}{\partial T}$ evaluated at the diapir temperature T_d. Using $T_d = 250\,\mathrm{K}$

and 180 K as an average value for $T(z)$, we find $U_{soft} \simeq 25 \left(\frac{\eta_0}{10^{13}\,\mathrm{Pa\,s}}\right)^{1/4}$ km My^{-1}. The warm ice thus reaches the surface at a time on the order of 10^6 y (Figure 3.2d).

Also shown in Figure 3.3 is a similar calculation of the rise of an early diapir partially driven by a 7 km residual ocean displaced from its equilibrium position. The base of the diapir has a melt fraction of 0.09 and a temperature of 251 K, and the total buoyancy is calculated by keeping track of pressure-release freezing of the melt (and its small associated temperature rise). The rise time to the base of the lithosphere and the minimum loading timescale are both about an order of magnitude smaller than for the thermally driven diapir, leading us to expect extension fracturing to a depth of perhaps 2 km.

The conclusion that diapiric loading can lead to fracture of the uppermost ~ 1 km of lithosphere is fairly robust. Both the expected value of h^{-1} and the value required for fracture scale as R^{-2}, so the poorly known quantity R does not affect the conclusion. Dispersion hardening (or more drastic inhibition of nonlinear creep), a surface colder than the assumed subregolith temperature of 130 K, and pre-existing weaknesses all act to increase the expected depth of fracture. What is less clear is the relationship between lithospheric fracturing, resurfacing, and groove formation. We will have more to say about this in Section 6.

4. Criteria for Lithospheric Fracture

In this section we derive and present criteria for extension fracturing of Ganymede's icy lithosphere by both global expansion and local deformation due to buoyant loading from below. Although second of these results applies directly to the possible warm ice diapirism during the heat pulse described above, we present them in a separate section to emphasize their independence from any particular assumptions about the *cause* of the lithospheric deformation. We show below that (subject only to reasonable assumptions about the rheology — plastic and brittle — of ice), for either local or

global deformation to lead to fracture, it must take place rapidly (*i.e.* on a timescale $\lesssim 10^3$–10^4 years). Warm ice diapirism satisfies this criterion; global expansion due to cooling and differentiation do not.

We consider first the simpler case of global expansion. Our analysis is similar to that of Squyres (1982) but draws on recent measurements of the rheology of ice at low temperature by Durham *et al.* (1983; 1984). They obtain a brittle failure strength in uniaxial compression of $\sim 30\,\mathrm{MPa}$ at temperatures of 113–158 K. We therefore adopt a uniaxial tensile strength $\sigma_t = 3\,\mathrm{MPa}$, based on the compressive strength and generalized Griffith failure theory (Jaeger and Cook 1979). The *macroscopically averaged* strength may of course be lower due to pre-existing fractures, but the relative simplicity of tectonic patterns on Ganymede have been used to argue for the "healing" of such zones of weakness (McKinnon and Parmentier 1986). The failure criterion for extension fracturing is $\sigma'_{\theta\theta} = -(\rho g z + \sigma_t)$, where $\sigma'_{\theta\theta} = \sigma'_{\phi\phi}$ is the additional horizontal stress at a depth z due to the expansion. We do not consider normal faulting, for which larger shear *and* confining stresses are required. We model the Ganymedean lithosphere as an incompressible nonlinear-viscoelastic shell surrounding an interior whose radius R varies with time. The horizontal strain rate is then $\dot{\epsilon}_{\theta\theta} = \dot{\epsilon}_{\phi\phi} = 2\frac{\dot{R}}{R}$, independent of depth. For consistency with Squyres (1982), we adopt an error function strain history: $R(t) = R_0 + \Delta R \frac{1}{2}(1 + \mathrm{erf}(ht))$, with a range of time constants h^{-1} and a radial strain $\frac{\Delta R}{R} = 3\%$ (Squyres 1980a). Including both diffusion and nonlinear creep, we obtain

$$\dot{\epsilon}_{\theta\theta} = -\frac{2\Delta R h}{\sqrt{\pi} R} e^{-h^2 t^2} = \frac{\dot{\sigma}'_{\theta\theta}}{2G} + \frac{\sigma'_{\theta\theta}}{2\eta_v(T)} + \sum_i A_i(T)(\sigma'_{\theta\theta})^{n_i}. \tag{4.1}$$

$G \simeq 10\,\mathrm{GPa}$ is the shear modulus, η_v is the volume-diffusion viscosity (see Appendix), and we have expressed the three different creep mechanisms elucidated by Durham *et al.* (1983; 1984) in the form $\dot{\epsilon} = A(T)\sigma'^n$, with A calculated for the flow geometry of interest here and hence differing from their uniaxial-stress parameter by a

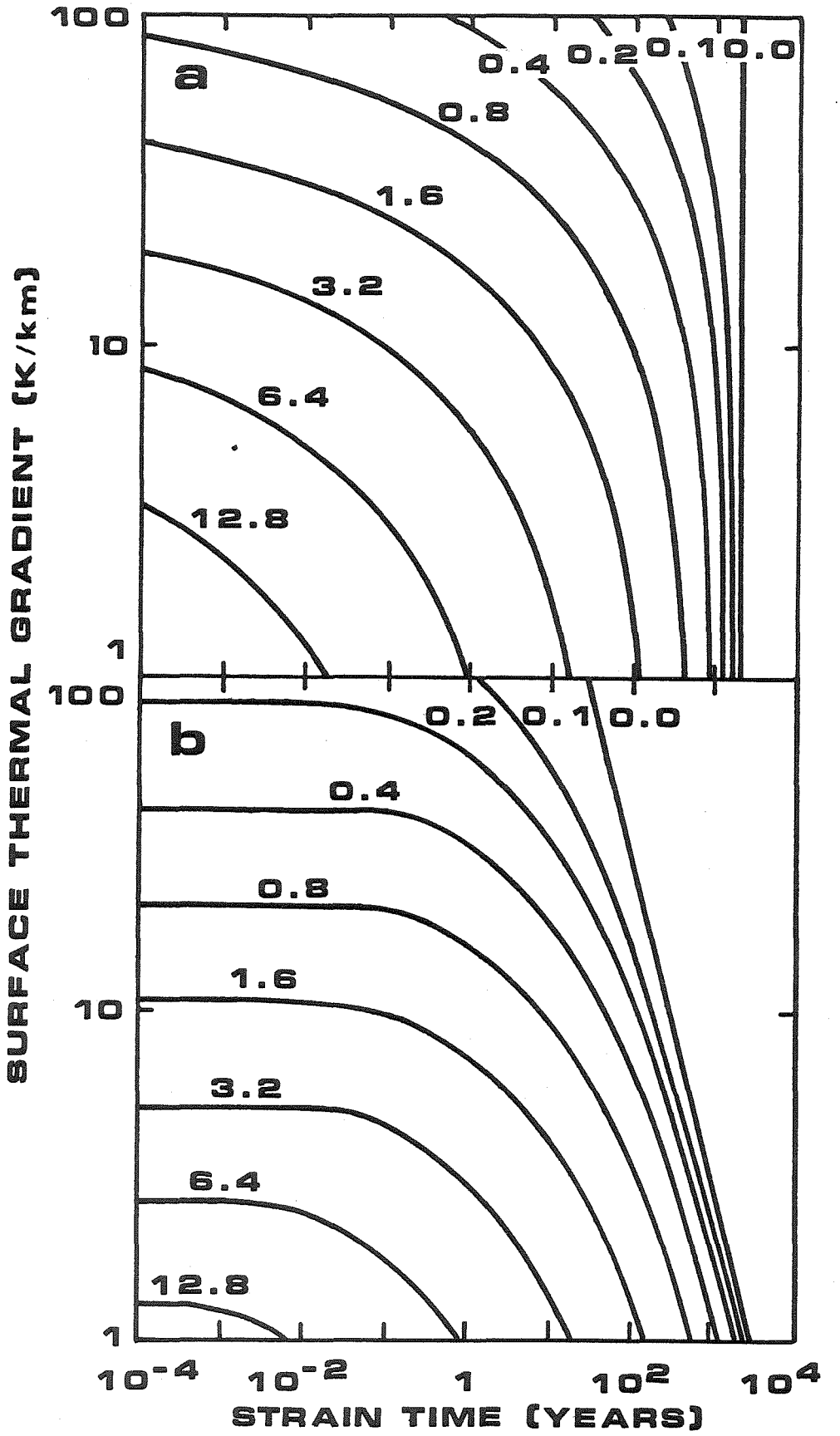
numerical factor. At any given temperature, one of these mechanisms will dominate the others. Term-by-term comparisons (with $T = 130\,\mathrm{K}$ and $\sigma'_{\theta\theta} = -\sigma_t$ appropriate to fracture at the surface) indicate that the elastic strain rate is unimportant for $h^{-1} \gtrsim \frac{2G}{A}\left(\frac{4G\Delta R}{R_0}\right)^{n-1} \simeq 10^{-9}\,\mathrm{y}$ (the Maxwell time at the elastic stress), while diffusion creep is unimportant for $h^{-1} \lesssim \left(\frac{4\eta_v \Delta R}{R_0 \sigma_t}\right)^{-1}\left(\frac{A\sigma_t^{n-1}}{2\eta_v}\right)^{1/(n-1)} \simeq 10^{12}\,\mathrm{y}$. In short, the stress is always instantaneously relaxed by nonlinear creep, and we can write the criterion for tension fracturing at a depth z directly in terms of the strain time h^{-1}:

$$h^{-1} \lesssim \frac{2\Delta R}{\sqrt{\pi} R_0 A(T(z))(\sigma_t + \rho g z)^n}. \tag{4.2}$$

Figure 4.1a shows the relationship between strain time, depth of failure, and surface thermal gradient implied by this equation. Even at the surface, $h^{-1} \lesssim 2000\,\mathrm{y}$ is required for failure; this value may be increased slightly by the effects of silicate inclusions. Durham *et al.* (1983) report that for their lowest-temperature creep mechanism the peak stress attained at low strains is $\sim 20\%$ greater than the steady-state stress at the same strain rate. The effect of dispersion hardening should be similar, inasmuch as it operates by inhibiting recrystallization and thus extending the small-strain regime. A strain time of $\sim 5000\,\mathrm{y}$ would then suffice for surface fracture. This result is of course sensitive to the assumed surface temperature. The value of 130 K used was based on the assumption of an insulating regolith (Passey and Shoemaker 1982); using instead $T_0 = 110\,\mathrm{K}$ based on solar equilibrium calculations (Squyres 1980*c*) without a substantial regolith allows fracture for strain times approximately 130 times longer. Since proposed sources of global volume change such as core differentiation and freezing of the ocean operate on timescales of $\gtrsim 10^8\,\mathrm{y}$, the probability of their involvement in lithospheric fracturing is remote.

This conclusion is weakened but not invalidated by a consideration of the implications of the nonlinear creep relation for crater relaxation. An analysis including the variation of viscosity with depth due to *both* stress and temperature in an ap-

Figure 4.1. Depth of extension fracturing (contours in km) as a function of strain time h^{-1} (strain $\propto$ erf(ht)) and surface thermal gradient for two strain geometries. Fracture stress is 3 MPa above confining pressure and corresponding strain rate is based on Durham *et al.* (1983; 1984) and a surface temperature of 130 K for both models. (a) Global expansion of 3% in radius. (b) Lithospheric flexing by a diapir of width 40 km, peak buoyancy 0.6 MPa. Initial fracture only, before migration of the neutral fiber in response to weakening of the failed zone.


a
b
0.4
0.2
0.1
0.0
0.8
1.6
3.2
6.4
12.8
100
10
1
SURFACE THERMAL GRADIENT (K/km)
10⁻⁴
10⁻²
1
10²
10⁴
STRAIN TIME (YEARS)

proximate way (Brennen 1974) leads to the conclusion that craters larger than about 3–5 km will be substantially relaxed after 3×10^9 y at a thermal gradient of $2\,\mathrm{K\,km^{-1}}$, leading us to suspect that the creep mechanisms reported by the Livermore group are suppressed to some unknown extent in the lithosphere of Ganymede. Nonetheless, if we assume a linear viscoelastic rheology and require the preservation of 50–100 km craters, extension fracturing by expansion on a 10^8 y timescale is only marginally possible.

Let us now consider instead the stresses set up by a warm ice diapir. As we showed in Section 3.2, the peak buoyant load is $\sigma_b = \rho g \alpha \Delta T Z \simeq 0.6\,\mathrm{MPa}$, much less than the tensile failure stress. In a thin lithosphere, however, geometric factors may lead to horizontal stresses that are much larger. Ignoring viscous relaxation for the moment, we illustrate this effect by the following standard result from elastic thin plate theory (Landau and Lifshitz 1970, pp. 44–43): if a plate of thickness z_p is loaded with a sinusoidally varying stress of wavelength λ, the amplitude of the surface fiber stress will be $6\left(\frac{\lambda}{2\pi z_p}\right)^2$ times that of the load. Taking $\lambda = Z = 160\,\mathrm{km}$, we find that tensional stresses as great as $\sim 25\,\mathrm{MPa}$ could be generated in a 10 km thick elastic lithosphere. Including isostatic compensation reduces this value to $\sim 18\,\mathrm{MPa}$.

Now consider a more realistic model, retaining the thin-plate approximation but recognizing that the fiber stresses will be determined not by elasticity, but by nonlinear creep as the plate changes shape under an increasing load. The equation of equilibrium is

$$M(x,t)_{,xx} = \sigma_b(x,t) - \rho g d(x,t), \tag{4.3}$$

where $\sigma_b(x,t)$ is the applied load as a function of horizontal coordinate x and time t, d is the plate deflection, and the comma indicates a partial derivative. The bending moment M is given by

$$M(x,t) = \int_0^\infty \left(\frac{\dot{\epsilon}_{xx}(x,z,t)}{A(T(z))}\right)^{\frac{1}{n}} (z_0 - z)\, dz. \tag{4.4}$$

Here, z_0 is the depth of the neutral fiber and the upper limit of integration has been extended to infinity (rather than z_p) without harm, since the viscosity decreases quasi-exponentially with depth. The flow geometry differs from that for global expansion, so the function $A(T)$ will differ from that in equation (4.2) by a numerical factor. For small curvatures, the fiber strain rate is $\dot{\epsilon}_{xx} = -(z_0 - z)d_{,xxt}$ and we get

$$M = I_1(-d_{,xxt})^{\frac{1}{n}}, \tag{4.5a}$$

where

$$I_1 \equiv \int_0^\infty A(T(z))^{-\frac{1}{n}}(z_0 - z)^{\frac{(n+1)}{n}}\, dz, \tag{4.5b}$$

subject to the condition

$$I_2 \equiv \int_0^\infty A(T(z))^{-\frac{1}{n}}(z_0 - z)^{\frac{1}{n}}\, dz = 0. \tag{4.5c}$$

Equation (4.5c) is the requirement that the plate be under no net tension; for a given thermal structure it determines z_0. Numerical solution shows that, for a wide range of thermal structures, the neutral fiber is 16.6 K warmer than the surface. At least initially, the depth of the neutral fiber is an upper limit on the depth of fracture, since for $z > z_0$ the fiber stresses are compressive. If and when a near-surface layer fractures, however, its contribution to both I_1 and I_2 is reduced or eliminated. The effective top surface of the lithospheric plate now lies below $z = 0$. In consequence, the neutral fiber must also move deeper into the planet, allowing failure to extend to a greater depth. This additional failure may in turn be followed by further migration of the neutral fiber, and so on, in a diminishing sequence. Substituting (4.5a) into (4.3) we obtain

$$((d_{,xxt})^{\frac{1}{n}})_{,xx} = \frac{\sigma_b - \rho g d}{I_1}, \tag{4.6}$$

to be solved for d given the loading history σ_b. We simplify the problem by assuming a load varying parabolically with x (and, more importantly, neglecting end effects!). It is then consistent to let $d = \frac{\sigma_b}{\rho g}$ since M will be constant and the entire load will be

isostatically supported. We can thus obtain a relationship between the instantaneous loading rate $\sigma_{b,t}$ and the fiber stress $\sigma_{xx}(z)$. Although only the instantaneous loading rate is important, to facilitate comparison with the global expansion case we express the result in terms of the equivalent strain time $h^{-1} \equiv \frac{\rho g \alpha \Delta T Z}{\sqrt{\pi}\sigma_{b,t}}$ for an error-function loading history with the same peak loading rate. The criterion for tension fracture at a depth z is then

$$h^{-1} \lesssim \frac{2(z_0 - z)\rho g \alpha \Delta T Z}{\sqrt{\pi} R^2 A(T(z))(\sigma_t + \rho g z)^n}, \tag{4.7}$$

for $z < z_0$, and where R is the half-width of the parabolic load distribution, *i.e.*, $\sigma_b \propto (R^2 - x^2)$. The initial depth of failure, with z_0 held fixed, is shown in Figure 4.1b for a variety of thermal gradients. Progressive fracture and migration of the neutral fiber leads to significantly different results only for $h^{-1} \lesssim 0.1$ y, the maximum depth of failure in this regime no longer being limited by z_0 for a given thermal gradient. We present these results in terms of a thermally driven diapir ($\Delta T = 20$ K, $R = 40$ km) as discussed in Section 3.2, but they may be applied to *any* source of local lithospheric loading by noting that, for fracture to any given depth, the required h^{-1} scales as $\frac{\sigma_b}{R^2}$. The required strain times are similar to those for global expansion because the total strains are similar, and the strong stress-dependence of the viscosity washes out the numerical factors due to differing geometries; as before, dispersion hardening and lower surface temperatures may permit fracture at somewhat larger h^{-1}. The significant difference is that diapiric loading on such short timescales is indeed possible.

5. Core Formation

The existence, nature, and timing of core formation in large icy satellites, especially Ganymede, are of interest for several reasons. First, core formation might trigger resurfacing either through global expansion (Squyres 1980a) or through core-driven

diapiric or plume activity of water or ice (Parmentier and Head 1979; McKinnon 1981). Second, the nature and timing of core formation might affect the thermal history of the overlying ice mantle, thereby influencing the evolution scenarios discussed in this paper. Third, the existence and nature of a core may influence a number of potentially measurable properties of the satellite and thereby serve to discern between competing models. Observations of gravitational moments and any magnetic field are particularly relevant.

Despite the importance of these issues, this section is relatively brief and follows the bulk of the modeling effort in this paper. The reason is that the approximate calculations of core formation discussed below all indicate a timescale of at most several hundred million years for near-completion of the process. This leads us to doubt that core formation plays a direct role in explaining the resurfacing of Ganymede. Consequently, a lengthy and precise quantification of the models seems pointless (even assuming it were possible!). Here, we adopt the "conservative" approach of identifying models that prolong core formation as much as possible.

Our analysis focuses on the postaccretional state illustrated in Figure 2.1 and characterized by a potentially unstable configuration of silicates overlying a primitive ice-rock mixture. It is conceivable, of course, that the formation of a core occurred contemporaneously with accretion, as seems likely for the earth (Stevenson 1981). Although possible in principle, it is unlikely that the retention of impact energy was sufficiently efficient (Schubert *et al.* 1981; Lunine and Stevenson 1982). It is also possible that these bodies accreted heterogeneously by first growing a rock core and then accreting ice. Although superficially attractive because of the much lower condensation temperature of ice, heterogeneous accretion models pose unresolved dynamic problems because the relevant dynamic timescales are much shorter than cooling times. Our subsequent discussion assumes the "initial" state illustrated in

Figure 2.1.

Although gravitationally unstable, a layer of silicates overlying an ice-rock mixture may persist for a substantial time because both layers are cold and have high viscosities. Both the rock layer and the underlying ice-rock mixture have very uncertain rheologies. During accretion, the rock layer will accumulate at the base of the water ocean, compacting and squeezing out most of the interstitial water by a combination of processes, probably dominated by crushing of rock (sintering) on large lengthscales and stress-driven chemistry (depositional closure of pores and cracks) on subgrain lengthscales. Existing relevant laboratory data are at higher temperatures, but estimated activation energies suggest that cementation can occur in water-permeated silicates on less than a geologic timescale. For example, an extrapolation of the work of Smith and Evans (1984) on water-mediated crack healing in quartz indicates that even at temperatures as low as 300 K, the process could occur on a timescale of 10^4 years at most. However, it is not clear whether significant *creep* can occur at these temperatures, even in the presence of plenty of water. Indeed, enormous (and hence unreliable) extrapolations of existing data (Jaoul *et al.* 1984; Blacic and Christie 1984; and other papers in the same special issue of *Journal of Geophysical Research*) suggest viscosities of perhaps as great as 10^{29} Pa s at 300 K. The increased ionicity of water with pressure may reduce this estimate. We shall adopt the conservative hypothesis (meaning the one most likely to prolong core formation) in which the rock layer becomes well cemented and forms a rigid "carapace" over the soft center.

With this admittedly extreme hypothesis, we are then confronted with the problem of rupturing and disaggregating the carapace. This can be achieved by the tensional stress that builds up as underlying ice melts, producing less dense water. Neglecting the small thermal expansion of the rock as it heats up, we can estimate the required melting by use of the equation for tensional stresses in a thick shell,

$R_{ic} < r < R_c$, subjected to an excess pressure Δp at the inner surface (Landau and Lifshitz 1970, p. 21):

$$\sigma_{\theta\theta} = \sigma_{\phi\phi} = \frac{\Delta p R_{ic}^3}{R_c^3 - R_{ic}^3}\left(1 + \frac{R_c^3}{r^3}\right). \tag{5.1}$$

In the thick shell limit, it is approximately valid to set $\Delta p = K\frac{\Delta v}{v}$, where K is the bulk modulus of water, Δv is the volume change of the inner core (due to partial melting) that would occur if the entire body were hydrostatic, and v is the total inner core volume. We adopt $\sigma_{\theta\theta}$, $\sigma_{\phi\phi} \gtrsim 0.2\,\mathrm{GPa}$ as our criterion for carapace rupture (suggested by the results of Blacic and Christie 1984); this leads to the requirement $\frac{\Delta v}{v} \gtrsim 0.02$, implying approximately 20% melting of the inner core, since the volume difference between water and ice VII at 3 GPa is $\sim 10\%$ (Hobbs 1974).

Proceeding conservatively, we assume that radiogenic heating is the only available energy source for melting. Available gravitational energy is small at this depth so a "runaway" core formation cannot occur (compare this with Friedson and Stevenson 1983, where a runaway was possible because the differentiation was initiated at a large radius). Tidal heating is possible but highly uncertain. Generalizing equation (2.4), the temperature rise prior to melting is given by

$$\Delta T(t) = \frac{\epsilon}{\overline{C}} \sum_i X(q_i) A(q_i) \frac{1 - e^{-\lambda_i t}}{\lambda_i}, \tag{5.2a}$$

where

$$\overline{C} \equiv \epsilon C_c + (1 - \epsilon) C_{ice} \tag{5.2b}$$

is the mass averaged specific heat of an ice-rock mixture, ϵ is the mass fraction of rock, and the other symbols have their usual meanings (Tables II–V). For $\epsilon = 1$, $\Delta T(t) \simeq 1200t$ where t is measured in units of 10^9 years (assuming $t \ll 0.5$). For $\epsilon = 0.4$ (the primordial core), $\Delta T(t) \simeq< 280t$. If the rock layer and the primordial core were intimately mixed, then $\epsilon \simeq 0.9$ and $\Delta T(t) \simeq 1000t$. A simple estimate for

the actual temperature in the inner core, prior to any heat redistribution, is

$$T(r,t) = 150 + 225\left(\frac{r}{R_{ic}}\right)^2 + 280t, \tag{5.3}$$

where the value at $r = 0$, $t = 0$ is an estimate for the ambient formation conditions of Ganymede (Lunine and Stevenson 1982), and the quadratic dependence on r is that predicted by simple accretion models, assuming $T(R_{ic}, 0) = T_m$, the appropriate high pressure melting point:

$$T_m(r) \simeq 450 - 75\left(\frac{r}{R_{ic}}\right)^2. \tag{5.4}$$

It is found that the time of melting onset is $\tau_m(r) \simeq 10^9\,\mathrm{y}\,(1 - \frac{r^2}{R_{ic}^2})$. Once melting is initiated, the subsequent fractional melting of ice, f, is given by

$$f(r,t) = \frac{\epsilon}{L(1-\epsilon)} \sum_i X(q_i) A(q_i) \frac{e^{-\lambda_i \tau_m} - e^{-\lambda_i t}}{\lambda_i}, \tag{5.5}$$

where L is the latent heat. To satisfy our rupture requirement, $\overline{f} \gtrsim 0.2$ where

$$\overline{f}(t) \equiv \int_{R_{min}}^{R_{ic}} \frac{3r^2 f(r,t)\,dr}{(R_{ic}^3 - R_{min}^3)}, \tag{5.6}$$

and $\tau_m(R_{min}) = t$. It is found that $\overline{f} \simeq 1.5t$ and the rupture is initiated after $\sim 1.4 \times 10^8$ years. In reality, fluid-filled cracks begin to nucleate and propagate through the carapace at lower stress levels than the tensile failure strength (Weertman 1971; Stevenson 1982). A network of cross-cutting fissures should develop, causing disaggregation of the rock layer.

The subsequent downward migration of rock fragments through soft (or slushy) ice can be modeled in one of two ways. It can be treated either as a Rayleigh-Taylor instability of an "effective medium" (a rock-rich icy layer overlying a less rock-rich icy layer) or as a Stokes flow problem for individual rock fragments. The former has the advantage that we need no knowledge of rock fragment sizes and is probably closer to a realistic description. However, it offers no way of understanding

the extrusion of ice from between rock fragments (the "toothpaste tube" problem, discussed below). The Stokes approach is simpler and can be extended to consider the extrusion question. Since the growth time of Rayleigh-Taylor instabilities in the nonlinear regime is the same (to within a numerical factor) as the "Stokes time" (the time it takes a density inhomogeneity to sink a distance equal to its own size), we choose a Stokesian formulation for convenience. Consider a spherical density enhancement of density $\rho + \Delta\rho$ and radius R, sinking through a uniform medium of dynamic viscosity L. If r is the distance of the blob from the center of the satellite, then

$$\frac{dr}{dt} \simeq -\frac{2g(r)\Delta\rho R^2}{9\eta(T(r,t))}. \tag{5.7}$$

This equation can be cast in the nondimensional form

$$\frac{dx}{dt} = -x \exp\left\{-A\left(\frac{T_m(x)}{T(x,t)} - 1\right)\right\}, \tag{5.8}$$

where $g = \frac{4\pi}{3}G\rho r$ is assumed, $x \equiv \frac{r}{R_{ic}}$, time is measured in units of $\tau_{Stokes} = \frac{27\eta_0}{8\pi G\rho\Delta\rho R^2}$, η_0 is the viscosity at the melting point, and a homologous temperature form for viscosity is assumed (c.f. equation A.1). For $\eta_0 = 10^{15}$ Pa s and $R = 10^2$ km, we find $\tau_{Stokes} \simeq$ 10–100 years. A number of numerical integrations of equation (5.8) were performed, subject to equations (5.3), (5.4), and $x = 1$ at $t = 0$. In other words, the time to rupture the carapace was neglected. In fact, the choice of initial condition is largely irrelevant since in all calculations the sinking is initially very rapid until a "self regulation" is reached, in which subsequent progress is mediated by the timescale for radiogenic heating of the surrounding mixture. (Enhanced heating due to the greater rock content of the sphere was ignored in this conservative calculation, but see below). To be precise, $\frac{T_m(x)}{T(x,t)}$ tends to a constant whose value is simultaneously consistent with equations (5.3), (5.4), and (5.8). Typically, this constant is about 1.7, and the sinking velocity is then ~ 0.2 cm per year. This state is usually achieved at around $r \simeq 0.5R_{ic}$ to $0.6R_{ic}$ (at which point most of the core formation is complete,

volumetrically speaking). *It is important to realize that this asymptotic state is highly insensitive to the rheology because it is dictated by the radiogenic heating timescale.* In fact, the result depends only on the logarithm of η_0 and only linearly on A. These models all predict that for $R \simeq 10^2\,\mathrm{km}$, the time taken to go from $x = 1$ to $x \leq 0.6$ is $4 \times 10^8\,\mathrm{y}$ or less. Recalling that R should be interpreted as the characteristic wavelength of a Rayleigh-Taylor instability, rather than the size of an individual rock fragment, we see that the evolution is rapid.

Nevertheless, the extrusion of ice from between the infalling rock fragments can be slower. This is of interest if it leads to the formation of liquid water at a later stage. This is the "toothpaste tube syndrome" (the well-known inability to extrude the last bit of toothpaste[1]), ameliorated in our case by the ability of the rock to lubricate the ice by the injection of radiogenic heat. The following simple model illustrates the basic physics.

Consider two planar rock surfaces separated by a horizontal distance δ and with vertical extent L. We assume that the intervening space is filled with ice which is extruding upward, driven by an excess pressure gradient $g\Delta\rho$. We also assume a continuous flow of heat into the ice from the hot rock and neglect $\frac{d^2}{\kappa}$ relative to the extrusion timescale, where d is a characteristic rock fragment size. The resulting Poiseuille flow is characterized by an average velocity

$$\overline{v} = \frac{g\Delta\rho\delta^2}{12\eta(T_0 + \alpha t)}, \tag{5.9}$$

where $T = T_0 + \alpha t$ is the ice temperature, α is a heating rate, and a continuity condition

$$\frac{d}{dt}\Big(L\delta\Big) = -\overline{v}\delta \tag{5.10}$$

is imposed. The resulting nondimensionalized equation has the form

$$\frac{dx}{dt} = -x^3 \exp\left\{-A\left(\frac{T_m}{T} - 1\right)\right\}, \tag{5.11}$$

[1] Was this definition worth waiting for, or what?

where $x = \frac{\delta}{\delta_0}$, δ_0 is the initial spacing between the rock walls, and the unit of time is $\tau_p = \frac{12\eta_0}{g\Delta\rho\delta_0^2}$. Unlike the previous problem, we are interested in the behavior as $x \to 0$ and no "self-regulation" is possible. In fact, T reaches T_m at some finite value of $x = x_m$. Numerical integration shows that $0.005 \lesssim x_m \lesssim 0.01$ for plausible parameter choices ($L \simeq 10^2\,\mathrm{km}$, $\delta_0 \simeq 10\,\mathrm{km}$, $\eta_0 \simeq 10^{14}\,\mathrm{Pa\,s}$). We conclude that most of the H_2O is extruded as ice; only a small amount is trapped and eventually escapes as water. The elapsed time to reach $x = x_m$ is $\lesssim 7 \times 10^8$ years. This small amount of water will *not* escape all the way to the surface of the satellite.

We conclude this section with a summary description of the probable sequence of events accompanying core formation.

1) Sedimentation and cementation of rock at the base of the primordial ocean but overlying a primordial core of ice and rock. This is contemporaneous with accretion ($t \simeq 0$).
2) Rupture of the rock carapace by the expansion induced as ice melts because of radiogenic heating. Disaggregation of the rock layer proceeds ($t \simeq 1 \times 10^8\,\mathrm{y}$).
3) Downward migration of rock fragments, initially very rapid. On a volumetric basis, most of the primordial core is displaced (or mixed with the downgoing rock) before $t \simeq 4 \times 10^8\,\mathrm{y}$.
4) Extrusion of ice from between rock fragments is 90% complete even at $4 \times 10^8\,\mathrm{y}$; only a very small amount ($\lesssim 1\%$) is trapped for so long that it melts ($t \simeq 7 \times 10^8\,\mathrm{y}$).
5) Sintering of the rock mass to form a coherent rock core (95% complete at $4 \times 10^8\,\mathrm{y}$).
6) Subsequent expulsion of water of hydration may occur (McKinnon 1981) at $t \simeq 10^9\,\mathrm{y}$.
7) Onset of convection in the core at $t \simeq 1.5 \times 10^9\,\mathrm{y}$.

6. Discussion

Our thermal evolution models indicate that a large, initially hot and differentiated ice-rock body will cool by subsolidus convection on a timescale of several times 10^8 years. For what we argue are plausible values of ice viscosity, radiogenic heating, and oceanic solute concentration, this cooling will reduce the thickness of the liquid layer to the point where the upper and lower ice mantles can interact. The overturn of the unstably thermally stratified mantles that may then take place is the most dramatic thermal and tectonic event in the model history. We propose in the light of our models that such an overturn took place in the evolution of Ganymede and that it was responsible for the formation of the resurfaced terrain. Accordingly, this final section is devoted to some of the most important questions raised by such a scenario: How was the resurfaced terrain emplaced? What was the mechanism of groove formation? How did Callisto escape resurfacing? At the moment we can only speculate about the answers to these questions. We conclude with a brief listing of the prospects for resolving some of these issues in the near future.

Emplacement of the Ganymedean resurfaced terrain in the solid state is, contrary to intuition, not a difficulty if we identify individual flows with the $\sim 100\,\mathrm{km}$ sized "structural cells" of the grooved terrain (Shoemaker *et al.* 1982). (The global pattern of resurfacing on a $\sim 10^3\,\mathrm{km}$ scale could, in our model, plausibly be associated with the pattern of convective spreading centers on the dense ice mantle, since, we argue in Section 3.1, it is from their topography that the upward diapiric flow of the heat pulse is derived.) Once the warm diapiric ice reaches the surface, the problem is essentially one of glacier flow (Paterson 1981, pp. 85–88). The basal shear stress in an ice slab of thickness z_m flowing down a slope s is $\tau_b = \rho g s z_m$, and on the isostatically uplifted arch of lithosphere over a diapir (cf. Section 3.2) we expect the slope to be of the order of $s \simeq \frac{\alpha \Delta T Z}{R} \simeq 4\alpha\Delta T \simeq \frac{1}{80}$. Hence $\tau_b \simeq 4\rho g \alpha \Delta T z_m$, which is even

less than the convective and diapiric stresses ($z_m \ll Z$). The appropriate viscosity is therefore that for volume diffusion creep (see Appendix, Table VI). The horizontal velocity profile in the slab is parabolic, and the mean velocity is

$$\overline{U} = \frac{\rho g s z_m^2}{3\eta}. \tag{6.1}$$

In the time $\tau_c \simeq \frac{z_m^2}{4\kappa}$ before the basal layer cools and rigidifies significantly (cooling of the upper layers is unimportant, as they will be "rafted" on the warmer ice below) the flow can be extended a distance

$$x_m \simeq \overline{U}\tau_c \simeq \frac{\rho g s z_m^4}{12\eta}. \tag{6.2}$$

Choosing $x_m = 10^2\,\mathrm{km}$ and using a temperature of 250 K, we obtain a required thickness for flooding of a structural cell of $z_m \gtrsim 1.9\left(\frac{\eta_0}{10^{13}\,\mathrm{Pa\,s}}\right)^{1/4}$ km. Given the uncertainties in obtaining this figure, it is not inconsistent with Schenk and McKinnon's (1985) estimate of a thickness of 1.0–1.5 km for the thickness of the grooved terrain in Uruk Sulcis, based on the minimum size of dark halo craters (assumed to have excavated subjacent ancient material).

The fact that the young terrains on Ganymede appear to have formed from "very fluid" material (Shoemaker *et al.* 1982) can also be reconciled with resurfacing by ice. Although a glacier-like flow would have a steep lobate margin during emplacement, a comparison of the characteristic strain time $\frac{2\eta}{\rho g z}$ and the thermal diffusion time $\frac{z^2}{4\kappa}$ indicates that a "toe" of height $z \gtrsim \left(\frac{8\eta}{\rho g}\right)^{1/3}$ will relax significantly before cooling. With $\eta_0 = 10^{13}\,\mathrm{Pa\,s}$ and $T = 250\,\mathrm{K}$ as usual, this is $\sim 100\,\mathrm{m}$. In addition to such relaxation during cooling, the results of Durham *et al.* (1984) indicate that *even at* 130 K a scarp higher than $\sim 500\,\mathrm{m}$ will viscously relax on a timescale of 10^9 years, though this figure will be increased if nonlinear creep is inhibited, as appears to be the case. The smoothness of the contacts between the ancient and resurfaced terrains at the $\gtrsim 1\,\mathrm{km}$ per line pair resolution of Voyager is therefore not surprising.

The greatest difficulty with solid-phase resurfacing is that of evolving the warm ice onto the surface. In contrast to the competing model of resurfacing by liquid water (which of course runs into difficulties in making the liquid sufficiently buoyant), the likelihood of lithospheric fracturing (Section 4) has little bearing on the eruption of ice. That the warm ice reaches the surface $\sim 10^6$ y after the fracturing occurs is not a problem; in this time interval, creep at 130 K will close fractures only to a depth of about 2 km. Rather, the anticipated tension fractures are too narrow to permit flows of significant thickness to be erupted. Glacier flow and flow of ice through a tension fracture are both special cases of Poiseuille flow, and inasmuch as the stress gradients driving each will be similar, the surface flow will be approximately as thick as the fracture through which it is fed. We expect fractures of width $\Delta x_f \simeq \frac{x_f z_0 \alpha \Delta T Z}{R^2}$ where x_f is their separation, probably a few times the depth of fracturing. As an *extreme* case, let $x_f = 3z_0$, and take $\frac{dT}{dz} = 2\,\mathrm{K\,km^{-1}}$, so that $z_0 \simeq 8\,\mathrm{km}$; then $\Delta x_f \simeq 60\,\mathrm{m}$. Although the tension fractures thus cannot supply flows of any significant size, they may nonetheless be of great importance in providing structural control of surface flooding (Golombek and Allison 1981). We argue below that fracturing by diapiric loading is also likely to be involved in groove formation.

It is possible to envision resurfacing *without* the eruption of clean ice, in which the proximity of a diapir warms and softens the ice above to the point where the more silicate-rich surface layers sink in an overturn similar to a Rayleigh-Taylor instability. Our calculations show, however, that such an overturn is strongly suppressed both by the proximity of a free surface and by the steep viscosity gradient, both of which force the horizontal wavelength of greatest instability to be small. Removal of the silicate-rich layer by subduction of slabs with a more advantageous width-to-depth ratio may be possible, but this cannot in any case be the whole story, given the evidence that the dark, ancient surface lies at modest depths below much of the grooved terrain

(Schenk and McKinnon 1985).

Perhaps the most promising possibility is that the formation of impact craters of modest size opens channels for the warm ice to reach the surface. Based on the thickness of the softened layer above the ascending diapir, the warm ice will probably reach a depth of $\sim 2\,\mathrm{km}$, so that a 10 km diameter crater would suffice to remove the overlying lithosphere. The topographic domes seen on the grooved terrain (Squyres 1980*b*) may be marginal cases of such impact-created channels, in which eruption of warm ice was limited to local uplift because the diapir was at too great a depth when the crater formed. (Subduction of limited regions of ancient terrain is another possible mechanism for producing such a wide orifice.) If the width of the eruptive channel greatly exceeds the depth to the diapir, the thickness, and hence areal extent, of surface flooding will be governed by the hydrostatic head $\rho\Delta TZ$ that the warm ice can achieve (we have argued that the lithosphere is bowed up by this amount, but once it is breached it will subside, allowing flooding). For a thermally driven diapir at 250 K, this is only $\sim 500\,\mathrm{m}$ but may be substantially greater for the earliest diapirs, so that extensive flooding by this mechanism may be possible. Modeling of the diapir-surface interaction and of the combined heat- and mass-flow problem of eruption from a large orifice needs to be done to remove these ideas from the realm of speculation.

We turn now to the problem of groove formation, in the restricted sense: accepting the hypothesis that grooves originated as extension fractures and/or graben, subsequently modified by viscous relaxation and mass wasting (Squyres 1980a), what was the source of the fracturing stress? Necking instability of a "plastic" surface layer has also been proposed as an origin for the multiple groove sets (Fink and Fletcher 1981) but the required discontinuity in rheology, from $n \lesssim 10$ to $n \gtrsim 10^4$ in equation (A.2) (Fletcher and Hallet 1983), is difficult to justify except as a macroscopic

consequence of the transition to brittle behavior near the surface. The requirements for groove formation are thus the same as in the case in which fracturing is explicitly invoked. As in Section 4, for any given flow geometry we can calculate the strain rate required to achieve tensional failure as a function of depth.

Motivated by the occurrence of grooves predominantly in the resurfaced terrain, we first consider the magnitude of the thermal stresses in a layer of ice emplaced at (say) 250 K and cooling towards a lower equilibrium temperature. Implicit in this model is the assumption that the ice is laterally confined, *i.e.*, that no horizontal motion takes place. The shortening of the "natural" length of the ice due to thermal contraction must thus be offset by viscous or elastic elongation, with an associated tensile stress (Turcotte and Schubert 1982). The temperature of a cooling half-space is a known function of time and depth (Carslaw and Jaeger 1959), and through it viscosity and strain rate are also both known. A simplified estimate of the stress history ignoring elasticity led us previously to doubt the importance of thermal stresses. Based on the instantaneous strain rate and the creep mechanisms of Durham *et al.* (1983; 1984), the peak stress will exceed the tensile strength only in the uppermost 50 m of a layer cooling towards 130 K. Elastic effects will, however, act to increase the maximum stress substantially. Numerical solution of the differential equation for stress in a nonlinear Maxwell fluid leads to the prediction of failure to a depth of 1 km for $T_0 = 130$ K. This figure depends strongly on the equilibrium surface temperature; for a value of 110 K (plausible for clean, bright ice on which no regolith has formed) it increases to nearly 10 km. Actually, of course, failure due to thermal stresses is limited to the thickness of the emplaced layer. Dispersion hardening may also act to increase the suceptibility to thermal fracture. Assuming that the typical horizontal lengthscale of deformation is three to four times the maximum depth of fracture (a result obtained for brittle-ductile necking instability and likely to apply roughly to

extension fracture also), we can thus reproduce the $\sim$ 6–8 km modal groove spacing obtained by Grimm and Squyres (1985).

Interestingly, thermal fracture to a substantial depth is also predicted for rheologies and temperatures appropriate to the Saturnian and Uranian satellites, hinting at a possible common mechanism of formation for at least some of the groovelike features observed on a variety of icy bodies. The rheology of $NH_3 \cdot 2H_2O$ glass is highly uncertain, but for plausible estimates (Stevenson and Lunine 1986) failure to 1 km will occur upon cooling from a magma at the eutectic temperature to $\sim$ 85–105 K, well in excess of typical equilibrium temperatures at Saturn. Pressure-solution creep in the presence of cryogenic pore fluids has been proposed to explain mobilization of ice on the Uranian satellites (Stevenson and Lunine 1986). The very low viscosity expected precludes significant thermal fracture in the early stages of cooling, but upon freezing of the pore fluid the effective viscosity must increase enormously to that of the ice matrix. Thermal fracture of the entire emplaced region is then possible even for a temperature drop on the order of one Kelvin.

There are nonetheless several problems with the thermal fracturing model of groove formation. Most fundamentally, the generation of tensile stress is critically dependent on the assumption of lateral confinement. If the ice is able to contract laterally, not only will the coldest, near-surface layers shrink to maintain their "natural" length without stress, they will subsequently be placed under compression as the ice beneath them cools. The result is analogous to the familiar process in which glass is tempered by quenching of the surface followed by slower cooling of the interior. In this case extension fracturing would not be possible, though grooves could conceivably form as compressional features if the near-surface layer were to buckle. Whether the net stress in a cooling ice layer of realistic dimensions is tensile or compressive can probably only be answered by numerical modeling. Second, even if thermal contrac-

tion leads to tensile stresses, the one-dimensional stress calculation outlined above does not indicate whether the observed morphology of multiple parallel grooves (let alone groove pairs and single grooves) will result. Indeed, one might rather expect the formation of a polygonal network of fractures, similar to mud cracks. A final problem with thermal fracturing is that grooves should be equally likely to form on all of the resurfaced areas on Ganymede, whereas some areas of young, high-albedo terrain are observed to be smooth. We therefore turn to fracturing of the lithosphere by diapirism in the mantle beneath.

The depth of fracturing due to diapiric loading calculated in Section 4 can also lead to multiple grooves of roughly the observed spacing, at least provided we invoke dispersion hardening to obtain $z_f \simeq 2\,\mathrm{km}$. This mechanism is also apparently more consistent with the observed morphology and distribution of the grooves. If overturn of sufficient vigor to cause crustal fracturing persists for longer than the $\sim 10^6\,\mathrm{y}$ diapiric rise time, then we can envision a sequence of events similar to that deduced by Golombek and Allison (1981) from crosscutting relationships:

1) Rise of the earliest warm-ice diapirs to the base of the boundary layer (Figure 3.2c) results in crustal fracturing, including the formation of "primary and secondary grooves."
2) After $\sim 10^6\,\mathrm{y}$ these diapirs reach the surface (Figure 3.2d). Resurfacing occurs, structurally confined by the primary and secondary grooves.
3) The thin resurfaced units cool on a timescale of $\sim 10^5\,\mathrm{y}$. Away from the feeding diapir the thermal gradient is similar to its original value. Any extension occurring before this will be viscously relaxed.
4) Ascent of younger diapirs nearby (Figure 3.2e) leads to fracturing of the resurfaced units, forming multiple groove sets or "tertiary grooves." Adjacent resurfaced units are structurally separated by their bounding grooves, so that their

groove sets are independent.

5) Resurfacing and fracturing continue, but once tertiary grooves form on a given unit, subsequent episodes of diapiric loading will (usually) reactivate them rather than create intersecting grooves. The thermal and stress history, and hence the groove morphology, of each unit will be homogeneous but different from that of its neighbors.

6) The final regions to be resurfaced may escape tertiary groove formation, becoming smooth terrain.

In this scenario, the strong but not total correlation between grooving and resurfacing occurs not because the resurfaced areas are intrinsically susceptible to groove formation, but because both resurfacing and lithospheric fracturing occurred above the regions of diapiric activity (which in turn are probably the regions above the ascending convective plumes of the lower mantle). The expected degree of correlation is unfortunately not quantifiable, so that we can only assert the plausibility of obtaining what is actually observed on Ganymede by this means. The implication is clear, however (for what it is worth) that the ancient surface underlying the grooved terrain is itself grooved.

There are two other difficulties with the sequence of events just proposed. First, although we believe that diapiric loading can produce tension fractures of up to 2 km depth, and hence tertiary groove sets with wavelengths of 6–8 km, the origin of some of the larger primary and secondary grooves is problematic. In particular, the morphology of groove pairs makes their interpretation as viscously relaxed graben attractive (Squyres 1982), but graben formation requires failure at confining stresses greater than $3\sigma_t$ (Jaeger and Cook 1979), *i.e.*, at depths greater than about 7 km. Second, while ongoing diapirism for several million years as required for our groove formation scenario seems plausible (the lower mantle contains enough excess heat to

supply $\sim 10^3$ diapirs), the 7×10^8 y range of crater density ages on grooved terrain (Shoemaker and Wolfe 1982) is very difficult to explain. Our parameterized convection model predicts that the heat flux enhancement is over in one tenth this time for $\eta_0 = 10^{13}$ Pa s; any resurfacing must come to an end even more rapidly. We know of no means for breaking up the heat pulse into episodes separated by 10^8 or more years. The maximum delay results if the bidirectional flow described in Section 3.1 is not immediately established. Alternating episodes of predominantly downward and upward diapirism will then take place, lasting roughly the time required to transport the mass of the residual ocean: $\sim 3 \times 10^6$ y, if one uses diapirs of 40 km radius. The wide range of apparent ages of the grooved terrain is in fact a stumbling block for any theory of the resurfacing of Ganymede. It is difficult to see how *any* resurfacing process can be vigorous enough to cause crustal fracture, while lasting the better part of a billion years.

We have offered plausibility arguments in this section for resurfacing and groove formation on Ganymede based on the results of our thermal evolution model. A complete understanding of the icy Galilean satellites also requires an explanation of the *absence* of resurfacing on Callisto, despite its similar radius and density. The difficulty therein unfortunately lies not in proposing an explanation, but in verifying its validity. This is as true of our model as it is in general. If we ccept for the moment that the heat pulse phenomenon was responsible for the resurfacing of Ganymede, there are several possible explanations for the lack of a heat pulse in Callisto. We list them in order of what we consider to be decreasing probability.

1) Models of accretion (Schubert *et al.* 1981; Lunine and Stevenson 1982) suggest that differentiation was limited to a much thinner outer region in Callisto than in Ganymede. Subsequent differentiation due to radiogenic heating also seems less likely in Callisto (Friedson and Stevenson 1983). If, as therefore seems

likely, the base of the primordial Callistian ocean lay at a pressure of less than 0.2 GPa, only ice I would have formed upon subsequent cooling. There would thus have been no reservoir of stored heat to supply a heat pulse once the residual ocean became thin.

2) It is possible, depending on the thermal structure of the protojovian nebula, that ammonia was able to condense at the orbit of Callisto, and was incorporated into that body in significant quantities, but not into Ganymede (Lunine and Stevenson 1982). With a mole fraction of NH_3 of the order of 1%, differentiation would of necessity have been extensive, but subsequent thinning of the ocean to the point where convective overturn would become possible would require cooling to nearly the H_2O–$NH_3 \cdot 2H_2O$ eutectic temperature. This is impossible if radiogenic heat is to be lost by conduction or by convection at any plausible viscosity. Thus, if it accreted condensed ammonia hydrate, Callisto could still have a substantial liquid ocean.

3) We argued in Section 2.5 that ice grain size and hence viscosity is controlled by stirring of oceanic sediments into the crust by large impacts. Starting from the same differentiated state, Callisto, which receives a lower impact flux than Ganymede because of gravitational focusing, would be expected to have a slightly more viscous ice crust. Given the sensitivity of the thermal history to η_0 (Figure 2.5), this difference could have substantially delayed (possibly even prevented) the Callistian heat pulse. The lower silicate content of Callisto (based on mean density) would, however, act to make an early heat pulse more likely. This last possibility also predicts a differentiated Callisto, with a small residual ocean today.

To what extent can future work or observations be expected to clarify the issues raised here? Certainly, it would be desirable to understand whether the proposed

finite amplitude instability responsible for the heat pulse can occur. This would require finite amplitude modeling of convection.

Tighter constraints on the parameters of our thermal evolution are unfortunately unlikely. Although Durham *et al.* (1983; 1984; 1985) have immensely extended our knowledge of the rheology of ice at outer-solar-system temperatures in recent years, measurements at the strain rates of interest as well lie beyond the limit of human patience and would in any case almost certainly be moot. Any uncertainties about the theoretical diffusion creep viscosity are overwhelmed by the strong dependence on grain size, an unknown and planet-dependent quantity.

Beyond the Jovian system, the Voyager encounter with Triton in 1989 could provide an additional test of our heat pulse scenario for resurfacing. Although smaller and less silicate-rich than Ganymede, Triton may have been extensively melted by tidal dissipation during capture by Neptune (McKinnon 1984). Its subsequent cooling could then have resembled that of Ganymede, including the occurrence of a heat pulse. Evidence of Tritonian resurfacing would be of great interest and would help clarify the influence of size, heat supply, impact flux, and abundance of minor consitituents on the thermal evolution. It is of course possible that such evidence, even if originally present, has been buried or modified beyond recognition by the presence of volatile species such as CH_4 and N_2 (Cruikshank *et al.* 1984). In particular, resurfacing by methane may depend on the high-pressure thermodynamics of clathrate (Lunine and Stevenson 1985).

The ability to discriminate between competing models of Ganymede's structure and evolution will, however, be significantly increased by the Galileo mission. Perhaps most importantly, the close flybys may allow moderately accurate measurements of the gravitational moments, and hence estimation of the degree of central condensation of the Galilean satellites (Hubbard and Anderson 1979). The relation-

ship between differentiation and resurfacing will, we hope, be clarified; current models range from a differentiated-but-unresurfaced Callisto (*e.g.*, our 2) and 3) above) to an undifferentiated-but-resurfaced Ganymede (Croft 1985), while the most widely held view is that resurfacing is a consequence of differentiation.

Spectroscopic detection of a residual N_2 atmosphere by Galileo is possible and would help set limits on the initial incorporation of ammonia in the icy satellites. This is, of course, highly relevant to the problem of planetary nebular structure, but if Callisto should prove to be centrally condensed, the role of ammonia in its history will be of especial interest.

Spectral data from the Galileo NIMS instrument may help constrain the grain size and silicate content of at least the outermost layers of the Jovian satellites, although extrapolation to the deep interior for the purposes of estimating the mantle viscosity will still be risky.

Galileo will, of course, make images of Ganymede and the other Jovian satellites at higher resolution than those obtained by Voyager 1 and 2. Such close-up views may shed light on the mechanisms of resurfacing and groove formation (as well as on crater morphology and other important icy-surface phenomena), but if past experience is a guide, they will be complex and open to multiple interpretations, raising more questions than they resolve.

Acknowledgements

We thank R. Reynolds and G. Schubert for pointing out potential difficulties with our heat pulse proposal, which led to significant revision and improvement of the analysis. Supported by NASA grant NAGW–185.

Appendix A: The Rheology of H_2O Ice

The greatest obstacle to constructing a definitive thermal evolution scenario for Gan-

ymede — and, for that matter, to answering many questions about outer solar system bodies — is our incomplete knowledge of the rheology of ice. In this Appendix we attempt to assess both the mechanisms of creep important to convection in Ganymede and the values of physical parameters (notably stress and ice grain size) controlling the resulting viscosity. The deformation behavior of water ice is complicated (see Weertman 1973; Hooke 1981; Goodman *et al.* 1981; Weertman 1983; Poirier 1982 for reviews, the last with especial reference to icy satellites). It depends not only on temperature and deviatoric stress, but also on pressure (Jones and Chen 1983), strain history (Mellor and Cole 1983; Ashby *et al.* 1978), grain size and fabric (Baker 1978; 1981; Lile 1978), suspended particles (Hooke *et al.* 1972; Baker and Gerberich 1979; Friedson and Stevenson 1983), dissolved impurities (Jones and Glen 1969; Goodman *et al.* 1976), and of course the ice polymorph being deformed. We are concerned with the steady-state creep at large strains of ices III, V, VI and in particular I_h, frozen from the Ganymedean ocean. At its maximum concentration, the ocean contains perhaps 1% NH_3, leading to a maximum incorporation of 3 ppm in the ice (Hobbs 1974), which results in an utterly negligible increase in viscosity (Jones and Glen 1969). As we shall see below, included silicate particles affect the viscosity only by determining the grain size of the ice. We thus concentrate first on the properties of the pure ices.

At very high shear stresses (~ 100 MPa) and homologous temperatures $\frac{T}{T_m} \simeq 0.95$, ice VI is perhaps 10^5–10^8 times stiffer than ice I (Poirier *et al.* 1981) — a result not widely appreciated, since the viscosity of $\sim 10^{12}$ Pa s obtained is often quoted without reference to the stress level; nothing is yet known about the stress-dependence of ice VI viscosity. Only preliminary measurements of the viscosity of ice V have been made (Durham *et al.* 1985), but at a strain rate of $3.5 \times 10^{-4}\,s^{-1}$ and temperatures of 230–250 K the viscosities obtained are very similar to those of ice I. The viscosity of ice III has a very strong dependence on both stress and temperature

($n = 5.5$ and $A = 156$ in equations A.1, A.2 below) and is lower than that of ice I at the same temperatures for stresses $\gtrsim 1$ MPa (Durham *et al.* 1984; 1985). As we show in the case of ice I below, additional deformation mechanisms may lead to much lower viscosities at the low stresses of interest here than would be predicted on the basis of these results. If the viscosity of ice VI remains anomalously high at low stresses of the order of 10 kPa, the ice VI mantle may form a separate convective cell. Otherwise, the exact viscosity of the dense ices is not critical to our model. Deep in Ganymede where they occur, the homologous temperatures are high. Ice I_h, on the other hand, extends to the cold exterior, and it is the increase of its viscosity from the ocean to the surface — by a factor of perhaps 10^{12}, dwarfing the differences between phases — that is the "bottleneck" controlling the rate of thermal evolution. For this same reason, we concentrate on the conditions in the upper (cold) boundary layer. Deformation in the lower boundary layer will most probably proceed by the same mechanism(s) as in the upper, but enhanced by partial melting along grain boundaries. For the purposes of our thermal model, we approximate this enhancement crudely but conveniently by using the same viscosity law (A.1) in both boundary layers, but setting $\frac{T}{T_m} = 1.0$ in the warm layer, rather than the actual value ~ 0.96.

A variety of mechanisms contribute to the steady creep of ice I at stresses below those that cause fracture (Goodman *et al.* 1981; Duval *et al.* 1983): diffusion within grains and along grain boundaries, and dislocation glide, which may in turn be controlled by proton rearrangement, kink nucleation, and defect formation. Recrystallization, partial melting, and even solid-solid phase transitions ("transformational plasticity" due to the ice I-II transition has been observed by Durham *et al.* 1983) can further modify creep. The strong temperature dependence of most of these mechanisms is commonly represented by an Arrhenius form $\eta \propto \exp\left\{\left(\frac{E^* + pV^*}{k_B T}\right)\right\}$, where E^* is an activation energy for the mechanism, V^* is an activation volume, and

k_B is Boltzmann's constant. We adopt the approximately equivalent formulation in terms of the *homologous temperature* $\frac{T}{T_m}$, where T_m is the pressure melting point (Weertman 1970):

$$\eta = \eta_0 \exp\left\{A\left(\frac{T_m}{T} - 1\right)\right\}. \qquad (A.1)$$

Here η_0 is the viscosity extrapolated to (*not* evaluated at) $T = T_m$, and the dimensionless constant $A \simeq 18$–35 for a wide variety of materials and mechanisms. In both experimental and theoretical work, additional dependences of the viscosity are commonly limited to proportionality of η_0 to powers of the temperature, the grain size d, and the equivalent shear stress $\sigma_e \equiv (\frac{1}{2}\sigma'_{ij}\sigma'_{ij})^{1/2}$:

$$\eta_0 = B\left(\frac{T}{T_0}\right)^k \left(\frac{d}{d_0}\right)^m \left(\frac{\sigma_e}{\sigma_0}\right)^{1-n}, \qquad (A.2)$$

(where B, k,m, and n are constants depending on the mechanism), although there is theoretical (Lile 1978) and experimental (Baker 1981) support for additional dependences on fabric and on the third stress invariant. In this paper we use as reference conditions for B the values $T_0 = 220\,\mathrm{K}$, $d_0 = 1\,\mathrm{mm}$, and $\sigma_0 = 1\,\mathrm{MPa}$. We make the assumption that an equilibrium grain size d is reached in the flow; as shown below the variation of T over the course of evolution is not important in the pre-exponential factor. Furthermore, numerical calculations (Parmentier *et al.* 1976) indicate, at least for temperature-*in*dependent viscosity, that convection with a stress-dependent viscosity ($n \neq 1$) obeys essentially the same Ra–heat flow relationship as convection with an appropriately chosen Newtonian ($n = 1$) viscosity (equal to the viscosity averaged with respect to the square of the strain rate). Our parameterized convection model implicitly based on Newtonian flow, may thus be applied for creep mechanisms with $n \neq 1$, provided η_0 is chosen in an appropriate way, self-consistent with the resulting strain rates. Our problem is thus to choose η_0 so that it will satisfy (A.2) with the boundary-layer temperatures and stresses to which it leads in the thermal model — and assuming the flow parameters A, B, k, m, and n appropriate to the creep

mechanism that dominates at those temperatures and stresses. The achievement of self-consistency is potentially complicated by uncertainties in the values of the flow law parameters, and by the fact that grain size (and fabric) are difficult to constrain.

We used an iterative approach to obtaining self-consistent rheologic parameters, first running the thermal evolution model with four sets of rheologic parameters ranging somewhat more broadly than that in the icy satellite literature (Passey 1982; Reynolds and Cassen 1979): $A = 18$ and 24 with $\eta_0 = 10^{13}$ and 10^{14} Pa s. We examined the convective stress, estimated from boundary layer theory (Turcotte and Oxburgh 1967):

$$\sigma_e \simeq 0.1\alpha_I(T_3 - T_2)(P_3 - P_2), \tag{A.3}$$

and the mean temperature in the boundary layer, $\overline{T} \equiv \frac{1}{2}(T_1 + T_2)$. In all cases, $0.76 \lesssim \frac{\overline{T}}{T_m} \lesssim 0.85$ and $5\,\mathrm{kPa} \lesssim \sigma_e \lesssim 20\,\mathrm{kPa}$, with temperatures decreasing and stresses increasing with time. These temperatures and stresses were then used to refine the choice of rheologic parameters. They lie in the diffusion creep regime of published deformation maps (Goodman *et al.* 1981), which are, however, based on extrapolations of data at higher temperatures. We therefore compared the viscosity for diffusion creep with the nonlinear creep mechanisms recently measured at low temperatures (albeit relatively high stresses) by Durham *et al.* (1983; 1984).

Claims of the observation of diffusion creep based on stress (Bromer and Kingery 1968) or grain-size (Baker 1978) dependence are controversial (Mellor and Testa 1969), but it is well understood theoretically (Nabarro 1948; Herring 1950; Coble 1963). The viscosity is Newtonian and is given by:

$$\eta_{diff} = c\frac{k_B T d^2}{42\Omega}\left\{D_V + \frac{\pi\delta}{d}D_B\right\}^{-1}, \tag{A.4}$$

where Ω is the atomic volume, δ is the grain-boundary thickness, D_V and D_B are the coefficients for lattice and boundary diffusion, respectively, and c is an enhancement factor, which is unity for infinitesimal strain but of order 0.4 for large strains (Raj

and Ashby 1971). At the temperatures of interest, volume (Nabarro-Herring) diffusion dominates diffusion at grain boundaries for $d \gtrsim 10^{-4}$ mm. The values of the viscosity parameters for diffusion creep given in Table VI are based on crystallographic and diffusivity data quoted in Goodman *et al.* (1981).

Table VI. Rheologic Parameters of Ice I_h[a]

Mechanism		T/T_m	A	B (Pa s)	k	m	n
Volume diffusion[b]		$\lesssim 0.9$	26.2	2.3×10^{14}	1	2	1
Livermore[c]	low T	0.58–0.71	12.8	1.2×10^{15}	[d]	[d]	4.8
	med. T	0.71–0.89	26.9	1.4×10^{11}	[d]	[d]	4.0
	high T	0.89–0.93	40.1	1.5×10^{10}	[d]	[d]	4.0

[a] Cf. equations (A.1), (A.2).
[b] Theoretical (input quantities from Goodman *et al.* 1981).
[c] Experimental (Durham *et al.* 1983; 1984).
[d] Zero by assumption.

Also appearing in Table VI are parameters for three mechanisms observed in the uniaxial compression of jacketed polycrystalline ice samples by Durham *et al.* (1983; 1984). The second of these dominates the others at the temperature of the upper boundary layer; whether it will dominate Nabarro-Herring creep as well depends on the relative magnitudes of $\eta_0(\sigma_e, d)$, but not on temperature, since A is nearly the same for the two mechanisms.

Two pieces of information from our preliminary models are useful at this point. First, at given heat flux, $\sigma_e \propto \eta_0^{1/3}$ approximately (exact proportionality follows from boundary layer theory if $\eta \neq \eta(T)$) — a constraint added to our expression $\eta_0 = \eta_0(\sigma_e, d)$ for the sum of the two mechanisms. Subject to this constraint, volume diffusion will be the dominant creep mechanism for $d \lesssim 6$ mm (corresponding to $\eta_0 \lesssim 8 \times 10^{15}$ Pa s). Second, if $\eta_0 \gtrsim 10^{15}$ Pa s, the Ganymedean ice I layer never becomes convective. Thus we conclude that, if it occurs, subsolidus convection in Ganymede will be controlled by volume diffusion creep.

Diffusion creep has very convenient theoretical properties: unlike glide creep,

it is both Newtonian and isotropic. Isotropy leads not only to the absence of fabric effects on viscosity, but also to the absence of dispersion hardening by suspended particles. During glide creep, substantial local concentration of stress occurs in unfavorably oriented ice grains, and this stress is ultimately accommodated by recrystallization (Duval *et al.* 1983). The mechanism of dispersion hardening is believed to be the inhibition of recrystallization by foreign particles, which "pin down" grain boundaries. Thus diffusive flow, which is not dependent on recrystallization because internal stresses are distributed uniformly among grains, will not suffer dispersion hardening. (The weaker viscosity enhancement due simply to the effect of large inclusions as "obstacles" to the flow (Friedson and Stevenson 1983) will, of course, be present.)

The controlling factor for the viscosity is thus the grain size, and it is here that suspended silicates become important. Surface tension drives the growth of ice grains even in the absence of locally concentrated stresses. Measurements of the growth rate as a function of temperature and pressure (Azumo and Higashi 1983) indicate that over the age of the solar system, grain sizes in excess of 1.5 m could be achieved for $T \gtrsim 195\,\mathrm{K}$ in *pure* ice. One might expect an equilibrium diameter to be reached, at which severe straining disrupts grains as fast as they grow, but this would not in fact occur. Recrystallization and Nabarro-Herring creep both depend on the kinetics of vacancy diffusion and hence have the same dependence on both grain size and temperature. For reasonable stresses, d^2 can always double in a time much less than $\dot{\epsilon}^{-1}$, while $\dot{\epsilon}^{-1}$ *more* than doubles because of the increasing viscosity and decreasing vigor of convection. Grain size thus grows without limit in pure ice, up to and beyond the point where convection becomes impossible. Even a small concentration of silicate particles will inhibit grain boundary movement enough, however, to limit d to a value for which convection is still possible. The remainder of this Appendix is an attempt

to estimate the silicate-controlled ice grain size. While admittedly conjectural, and subject to uncertainties at several key points, it nonetheless raises (and attempts to answer) the key questions concerning d: How do inclusions control grain size? and how (hence in what quantity) do silicates make their way into the ice I layer?

The theory of Zener (Smith 1948, but note typographic error) relates the maximum ice grain diameter to inclusion radius r and volume fraction ϕ by

$$d = \frac{8}{3}\frac{r}{\phi}. \qquad (A.5)$$

For polydisperse inclusions one may use (A.5) with the effective radius defined by

$$r_{eff} = \frac{\int r^3\, dn}{\int r^2\, dn}. \qquad (A.6)$$

The particle size distribution $n(r)$ is unfortunately an unknown. We will assume a distribution with a power law dependence typical of collisional processes: $n(m) \propto m^{-1}$, or $dn \propto r^{-4}\, dr$ for $r_{min} \leq r \leq r_{max}$. Then equation (A.6) yields

$$r_{eff} = r_{min}\frac{\ln(r_{max}/r_{min})}{1 - r_{min}/r_{max}}. \qquad (A.7)$$

Also, if the distribution of silicate particles originally accreted ranged from $r_{0\,max} \geq r_{max}$ to $r_{0\,min} \leq r_{min}$ with a total volume fraction ϕ_0, and if only a fraction f of the particles at each radius between r_{min} and r_{max} (and none outside this range) find their way into the ice, then

$$\frac{\phi}{\phi_0} = f\frac{\ln(r_{max}/r_{min})}{\ln(r_{0\,max}/r_{0\,min})}. \qquad (A.8)$$

Combining (A.6), (A.7), and (A.8), and assuming $r_{max} \gg r_{min}$,

$$d \simeq \frac{8}{3}\frac{r_{min}}{\phi_0 f}\ln\left(\frac{r_{0\,max}}{r_{0\,min}}\right). \qquad (A.9)$$

Using the assumed silicate density $\rho_c = 3000\,\mathrm{kg\,m^{-3}}$ leads to $\phi_0 \simeq 0.36$ based on the bulk density of Ganymede. The values of $r_{0\,max}$ and $r_{0\,min}$ are not critical; we

take 10^3 m and 10^{-7} m, respectively, as used in Friedson and Stevenson (1983). Then $d \simeq \frac{170 r_{min}}{f}$.

Estimation of r_{min} and f depends on a consideration of how silicate material accreted with the Ganymedean ocean can find its way into the ice mantle. Clearly, the largest rock fragments will settle out and cannot enter the ice I layer. Mixing length theory yields an estimate of convective velocities in the ocean of $\sim 10^{-2}\,\mathrm{m\,s^{-1}}1$ based on typical heat fluxes (Schubert *et al.* 1981), so that particles with $r \gtrsim 100\,\mu$m will surely settle. The remaining material constitutes $\sim 30\%$ of the total mass of silicates, and will remain suspended unless flocculation of clay minerals causes accumulation into larger particles (the topography on the ice-water interfaces is quite subdued, so that only a small fraction of the larger suspended grains will enter the turbulent boundary layer where they can settle out). The flocculation process depends on cationic concentrations as well as on the clay mineral species (Whitehouse *et al.* 1959). At the low ionic concentrations probable in the Ganymedean ocean ($\sim 4 \times 10^{-3}\,M$ if all available chloride in the suspended sediment is leached) most clays, including montmorillonites, do not flocculate, though kaolinite may flocculate when exposed to only $10^{-9}\,M$ of Mg^{++}. Lacking knowledge of the clay mineralogy in the ocean, we will assume tentatively that the fine silicate particles remain dispersed and in suspension.

How can suspended silicates with $r \lesssim 100\,\mu$m become entrapped in the ice? Experiment (Corte 1962) indicates that grains of this size will be pushed ahead of a planar ice surface and excluded (even against gravity) unless the solidification front moves faster than $\sim 10^{-7}\,\mathrm{m\,s^{-1}}$. The ice I mantle thickens by $\sim 100\,$km in $3 \times 10^8\,$y, giving an interfacial velocity on the order of $10^{-11}\,\mathrm{m\,s^{-1}}$. Silicates will thus enter the ice only if the interface becomes *non*planar, so that they can become physically entrapped.

Constitutional supercooling can give rise to a highly structured freezing sur-

face in an impure melt (Harrison and Tiller 1963), but consideration of the effects of stirring (Burton *et al.* 1953) by convection indicates this will not occur in Ganymede. Mechanical disruption of the freezing layer can, however, lead to entrapment of silicates. At the onset of freezing, the upper few cm of the ocean will be boiling vigorously into near-vacuum, leading to the formation of a slush of ice fragments known as frazil ice. As the ice thickens, boiling will cease, but the layer will continue to be disrupted and stirred into the ocean by those impacts that completely penetrate it. We refer to this large-scale stirring as the formation of *megafrazil* and estimate that it can lead to the introduction of significant amounts of oceanwater (and, hence, suspended silicates) into the forming ice. (The silicates contained in the impacting bodies themselves may be neglected here because the impactor volume is a miniscule fraction of the volume of the crater.) Our calculation is based on the cratering rates of Shoemaker and Wolfe (1982) as a function of time t (measured from 4.55×10^9 y ago) and crater diameter D given by:

$$\frac{\partial^3 n}{\partial D \partial A \partial t} = -\frac{\gamma}{D_0}\left(\frac{D}{D_0}\right)^{\gamma-1}[R_0 e^{-\lambda(t-t_0)} + R_1], \qquad (A.9)$$

with $\gamma = -2.2$, $D_0 = 10\,\mathrm{km}$, $\lambda = \ln 2 (10^8\,\mathrm{y})^{-1}$, $t_0 = 1.25 \times 10^9\,\mathrm{y}$, $R_0 = 263 (10^6\,\mathrm{y})^{-1}$ $(10^6\,\mathrm{km}^2)^{-1}$, and $R_1 = 115 (10^6\,\mathrm{y})^{-1} (10^6\,\mathrm{km}^2)^{-1}$. We also assume a conductively thickening crust with thickness $Z \propto t^{1/2}$, and assume that for any crater with $D \geq 5Z$ when it forms, the true crater penetrates into the ocean. If the breccia lens, which will occupy most of the true crater, has a porosity p, a volume of oceanwater $\sim \frac{\pi}{4} D^2 Z p$ will rise hydrostatically to fill the void space. Integrating over t and D and comparing the amount of oceanwater introduced to the total crustal volume yields the silicate incorporation efficiency f:

$$f = \frac{\pi}{4Z(t)} \int_0^t \int_{5Z(t')}^{\infty} D^2 Z(t') p \frac{\partial^3 n}{\partial D \partial A \partial t}\, dD\, dt'. \qquad (A.10)$$

Substitution of (A.9) into (A.10) yields a result in terms of the incomplete gamma function, but we are primarily interested in the asymptote at early times: $f \sim t^{0.9}$,

falling off after 10^8 years because of the decaying cratering rate. By as early as 10^7 years each spot on the crust has been punctured an average of once; the incorporated silicates will thus be fairly uniformly stirred after this time, leading to a uniform viscosity.

Combining (A.8) and (A.10), with $r_{max} = 100\,\mu$m, $r_{min} = r_{0\,min} = 0.1\,\mu$m, we obtain an estimate of the ice grain size and hence viscosity during conductive crustal growth. Comparing this steadily decreasing viscosity with the threshold viscosity for convection exhibited by the thermal model, we can estimate the time and viscosity at which convection begins as a function of p. The asymptotic forms for $p \lesssim 0.5\%$ are

$$t_{onset} \sim 7 \times 10^7 p^{-0.7}\,\mathrm{y}, \qquad (A.11a)$$

$$d_{onset} \sim 0.6\,p^{-0.5}\,\mathrm{mm}, \qquad (A.11b)$$

$$\eta_{0\,onset} \sim 9 \times 10^{13} p^{-0.9}\,\mathrm{Pa\,s}, \qquad (A.11c)$$

where p is expressed in percent. Clearly, p is not well known, but values of a few percent are plausible and lead to ice viscosities in the range $10^{13} \lesssim \eta_0 \lesssim 10^{14}$ Pa s. As we showed in Section 2.5, this is precisely the range of viscosities for which the predicted thermal history of Ganymede is most interesting. It should be noted, however, that additional and potentially considerable uncertainties in the results (A.11) stem from the facts that the cratering rate at the early times of interest ($\gtrsim 4.4 \times 10^9$ y ago) may have differed greatly from the assumed rate (A.9) recorded at later times, and that the silicate particle size distribution assumed is also conjectural.

References

ANDERS, E., AND M. EBIHARA (1982). Solar System Abundances of the Elements. *Geochim. Cosmochim. Acta.* **46**, 2363–2380.

ASHBY, M. F., G. H. EDWARDS, J. DAVENPORT, AND R. A. VERRALL (1978). Applications of Bound Theorems for Creeping Solids and Their Application to Large Strain Diffusional Flow. *Acta Metall.* **26**, 1379–1388.

AZUMO, N., AND A. HIGASHI (1983). Effects of Hydrostatic Pressure on the Rate of Grain Growth in Antarctic Polycrystalline Ice. *J. Phys. Chem.* **87**, 4060–4064.

BAKER, R. W. (1978). The Influence of Ice-Crystal Size on Creep. *J. Glaciol.* **21**, 485–500.

BAKER, R. W. (1981). Textural and Crystal-Fabric Anisotropies and the Flow of Ice Masses. *Science.* **211**, 1043–1044.

BAKER, R. W., AND W. W. GERBERICH (1979). The Effect of Crystal Size and Dispersed Solid Inclusions on the Activation Energy for Creep of Ice. *J. Glaciol.* **24**, 179–194.

BERCOVICI, D., G. SCHUBERT, AND R. T. REYNOLDS (1986). Phase Transitions and Convection in Icy Satellites. *Geophys. Res. Lett.* **13**, 448–451.

BIRCH, F., ED. (1942). *Handbook of Physical Constants*, Geol. Soc. Amer. Special Paper 36.

BLACIC, J. D., AND J. M. CHRISTIE (1984). Plasticity and Hydrolytic Weakening of Quartz Single Crystals. *J. Geophys. Res.* **89**, 4223–4239.

BOOKER, J. R. (1976). Thermal Convection With Strongly Temperature Dependent Viscosity. *J. Fluid Mech.* **76**, 741–754.

BOOKER, J. R., AND K. C. STENGEL (1978). Further Thoughts on Convective Heat Transport in a Variable-Viscosity Fluid. *J. Fluid Mech.* **86**, 289–291.

BRENNEN, C. (1974). Isostatic Recovery and the Strain Rate Dependent Viscosity of the Earth's Mantle. *J. Geophys. Res.* **79**, 3993–4001.

BROMER, D. J., AND W. D. KINGERY (1968). Flow of Polycrystalline Ice at Low

Stresses and Small Strains. *J. Appl. Phys.* **39**, 1699–1691.

BURTON, J. A., R. C. PRIM, AND W. P. SLICHTER (1953). The Distribution of Solute in Crystals Grown from the Melt. Part I. Theoretical. *J. Chem. Phys.* **21**, 1987–1991.

CARSLAW, H. S., AND J. C. JAEGER (1959). *Conduction of Heat in Solids*, 2nd ed. Clarendon Press, Oxford, p. 59.

CASSEN, P. M., S. J. PEALE, AND R. T. REYNOLDS (1980). On the Comparative Evolution of Ganymede and Callisto. *Icarus.* **41**, 232–239.

CASSEN, P. M., S. J. PEALE, AND R. T. REYNOLDS (1982). Structure and Thermal Evolution of the Galilean Satellites. In *Satellites of Jupiter* (D. Morrison, Ed.), Univ. of Arizona Press, Tucson, pp. 93–128.

CHANDRASEKHAR, S. (1961). *Hydrodynamic and Hydromagnetic Stability*, Clarendon Press, Oxford.

COBLE, R. L. (1963). A Model for Boundary Diffusion Controlled Creep in Polycrystalline Solids. *J. Appl. Phys.* **34**, 1679–1682.

CONSOLMAGNO, G. J., AND J. S. LEWIS (1976). Structural and Thermal Models of Icy Galilean Satellites. In *Jupiter* (T. Gehrels, Ed.), Univ. of Arizona Press, Tucson, pp. 1035–1051.

CORTE, A. E. (1962). Vertical Migration of Particles in Front of a Moving Freezing Plane. *J. Geophys. Res.* **67**, 1085–1090.

CROFT, S. K. (1985). A New Scenario for Differentiation of Ganymede and Callisto: Beauty is Only Skin Deep. *LPSC Abstracts.* **XVI**, 152–153.

CRUIKSHANK, D. P., R. H. BROWN, AND R. N. CLARK (1984). Nitrogen on Triton. *Icarus.* **58**, 293–305.

DURHAM, W. B., H. C. HEARD, AND S. H. KIRBY (1983). Experimental Deformation of Polycrystalline H_2O Ice at High Pressure and Low Temperature:

Preliminary Results. *Proc. LPSC* **XIVth**, in *J. Geophys. Res.* **88** (Suppl.), B377–B392.

DURHAM, W. B., S. H. KIRBY, AND H. C. HEARD (1984). Flow and Fracture of H_2O Ices I_h, II and III: Latest Experimental Results. *LPSC Abstracts.* **XV**, 234–235.

DURHAM, W. B., S. H. KIRBY, AND H. C. HEARD (1985). Rheology of the High Pressure H_2O Ices II, III, and V. *LPSC Abstracts.* **XVI**, 198–199.

DUVAL, P., M. F. ASHBY, AND I. ANDERMAN (1983). Rate-Controlling Processes in the Creep of Polycrystalline Ice. *J. Phys. Chem.* **83**, 4066–4074.

ELLSWORTH, K., AND G. SCHUBERT (1983). Saturn's Icy Satellites: Thermal and Structural Models. *Icarus.* **54**, 490–510.

FINK, J. H., AND R. C. FLETCHER (1981). A Mechanical Analysis of Extensional Instability on Ganymede (abstract). In *Reports of Planetary Geology Program*, pp. 51–53. NASA TM 84211.

FLETCHER, R. C., AND B. HALLET (1983). Unstable Extension of the Lithosphere: A Mechanical Model for Basin-and-Range Structure. *J. Geophys. Res.* **88**, 7457–7466.

FRIEDSON, A. J., AND D. J. STEVENSON (1983). Viscosity of Rock-Ice Mixtures and Applications to the Evolution of Icy Satellites. *Icarus.* **56**, 1–14.

GANAPATHY, R., AND E. ANDERS (1974). Bulk Composition of the Moon and Earth, Estimated from Meteorites. *Geochim. Cosmochim. Acta.* **5** (Suppl.), 1181–1206.

GOLOMBEK, M. P., AND M. L. ALLISON (1981). Sequential Development of Grooved Terrain and Polygons on Ganymede. *Geophys. Res. Lett.* **8**, 1139–1142.

GOODMAN, D. J., H. J. FROST, AND M. F. ASHBY (1976). The Effect of Impu-

rities on the Creep of Ice I_h and its Illustration by the Construction of Deformation Maps. IASH Publication No. 118, pp. 17–22.

GOODMAN, D. J., H. J. FROST, AND M. F. ASHBY (1981). The Plasticity of Polycrystalline Ice. *Phil. Mag. A.* **43**, 665–695.

GRIMM, R. E., AND S. W. SQUYRES (1985). Spectral Analysis of Groove Spacing on Ganymede. *J. Geophys. Res.* **90**, 2013–2021.

HARRISON, J. D., AND W. A. TILLER (1963). Controlled Freezing of Water. In *Ice and Snow* (W. D. Kingery, Ed.), pp. 215–225. MIT Press, Cambridge, MA.

HERRING, C. (1950). Diffusional Viscosity of a Polycrystalline Solid. *J. Appl. Phys.* **21**, 437–445.

HOBBS, P. V. (1974). *Ice Physics*, Clarendon Press, Oxford.

HOOKE, R. LE B. (1981). Flow Law for Polycrystalline Ice in Glaciers: Comparison of Theoretical Predictions, Laboratory Data, and Field Measurements. *Rev. Geoph. Space Phys.* **19**, 664–672.

HOOKE, R. LE B., B. B. DAHLIN, AND M. T. KAUPER (1972). Creep of Ice Containing Dispersed Fine Sand. *J. Glaciol.* **11**, 327–336.

HUAUX, A. (1951). Sur un Modèle de Satellite en Glace. *Bull. Acad. Roy. Sci. Belgique.* **37**, 534–539.

HUBBARD, W. B., AND J. D. ANDERSON (1978). Possible Flyby Measurements of Galilean Satellite Interior Structure. *Icarus.* **33**, 336–341.

JAEGER, J. C., AND N. G. W. COOK (1979). *Fundamentals of Rock Mechanics*, 3rd ed. Chapman and Hall, London.

JAOUL, O., J. TULLIS, AND A. KRONENBERG (1984). The Effect of Varying Water Contents on the Creep Behavior of Heavitree Quartzite. *J. Geophys. Res.* **89**, 4298–4312.

JARVIS, G. T. (1984). Time-Dependent Convection in the Earth's Mantle. *Phys. Earth Planet. Inter.* **36**, 305–327.

JOHNSON, M. L., A. SCHWAKE, AND M. NICHOL (1985). Partial Phase Diagram for the System NH_3–H_2O: The Water-Rich Region. In *Ices in the Solar System* (J. Klinger, D. Benest, A. Dollfus and R. Smoluchowski, Eds.), D. Reidel, Netherlands, pp. 39–47.

JONES, S. J., AND H. A. M. CHEN (1983). Creep of Ice as a Function of Hydrostatic Pressure. *J. Phys. Chem.* **87**, 4064–4066.

JONES, S. J., AND J. W. GLEN (1969). The Effect of Dissolved Impurities on the Mechanical Properties of Ice Crystals. *Phil. Mag.* **19**, 13–14.

KAULA, W. M. (1968). *An Introduction to Planetary Physics: The Terrestrial Planets*, John Wiley and Sons, NY, p. 111.

KIRK, R. L., AND D. J. STEVENSON (1983). Thermal Evolution of a Differentiated Ganymede and Implications for Surface Features. *LPSC Abstracts.* **XIV**, 373–374.

LANDAU, L. D., AND E. M. LIFSHITZ (1959). *Fluid Mechanics*, Pergamon Press, NY.

LANDAU, L. D., AND E. M. LIFSHITZ (1970). *Theory of Elasticity*, Pergamon Press, NY.

LEWIS, J. S. (1971*a*). Satellites of the Outer Planets: Thermal Models. *Science.* **172**, 1127–1128.

LEWIS, J. S. (1971*b*). Satellites of the Outer Planets: Their Physical and Chemical Nature. *Icarus.* **15**, 174–185.

LILE, R. C. (1978). The Effect of Anisotropy on the Creep of Polycrystalline Ice. *J. Glaciol.* **21**, 475–483.

LUNINE, J. I., AND D. J. STEVENSON (1982). Formation of the Galilean Satellites

in a Gaseous Nebula. *Icarus.* **52**, 14–39.

LUNINE, J. I., AND D. J. STEVENSON (1985). Thermodynamics of Clathrate Hydrate at Low and High Pressures with Applications to the Outer Solar System. *Astrophys. J. Suppl.* **58**, 493–531.

MCKINNON, W. B. (1981). Tectonic Deformation of Galileo Regio and Limits to the Planetary Expansion of Ganymede. *Proc. LPSC* **XIIth**, 1585–1597.

MCKINNON, W. B. (1984). On the Origin of Triton and Pluto. *Nature.* **311**, 355–358.

MCKINNON, W. B., AND E. M. PARMENTIER (1986). Ganymede and Callisto. In *Satellites* (J. A. Burns and M. S. Matthews, Eds.), Univ. of Arizona Press, Tucson, pp. 718–763.

MELLOR, M., AND D. M. COLE (1983). Stress/Strain/Time Relations for Ice under Uniaxial Compression. *Cold Regions Sci. and Tech.* **6**, 207–230.

MELLOR, M., AND R. TESTA (1969). Creep of Ice under Low Stress. *J. Glaciol.* **8**, 147–152.

MORRIS, S. (1982). The Effects of a Strongly Temperature Dependent Viscosity on Slow Flow Past a Hot Sphere. *J. Fluid Mech.* **124**, 1–26.

NABARRO, F. R. N. (1948). Deformation of Crystals by the Motion of Single Ions. In *Strength of Solids* (N. F. Mott, Ed.), Physical Society, London, p. 75.

NATAF, H. C., AND F. M. RICHTER (1981). Convection Experiments in Fluids with Highly Temperature-Dependent Viscosity and the Thermal Evolution of the Planets. *Proc. NATO A.S.I.*, Early Evolution of the Planets and Their Satellites.

PARMENTIER, E. M., AND J. W. HEAD (1979). Internal Processes Affecting Surfaces of Low Density Satellites: Ganymede and Callisto. *J. Geophys. Res.* **84**, 6263–6276.

PARMENTIER, E. M., D. L. TURCOTTE, AND K. E. TORRANCE (1976). Studies of Finite-Amplitude Non-Newtonian Thermal Convection with Application to Convection in the Earth's Mantle. *J. Geophys. Res.* **81**, 1839–1846.

PASSEY, Q. R. (1982). *Viscosity Structure of the Lithospheres of Ganymede, Callisto, and Enceladus, and of the Earth's Upper Mantle*, Ph.D. Thesis (Unpubl.), Caltech.

PASSEY, Q., AND E. M. SHOEMAKER (1982). Craters and Basins on Ganymede and Callisto: Morphological Indicators of Crustal Evolution. In *Satellites of Jupiter* (D. Morrison, Ed.), Univ. of Arizona Press, Tucson, pp. 379–454.

PATERSON, W. S. B. (1981). *The Physics of Glaciers*, 2nd ed. Pergamon Press, NY.

POIRIER, J. P. (1982). Rheology of Ices: A Key to the Tectonics of the Ice Moons of Jupiter and Saturn. *Nature.* **299**, 683–688.

POIRIER, J. P., C. SOTIN, AND J. PEYRONNEAU (1981). Viscosity of High-Pressure Ice VI and Evolution and Dynamics of Ganymede. *Nature.* **292**, 225–227.

RAJ, R., AND M. F. ASHBY (1971). On Grain Boundary Sliding and Diffusional Creep. *Trans. Met. Soc. AIME.* **2**, 1113–1127.

REYNOLDS, R. T., C. ALEXANDER, A. SUMMERS, AND P. CASSEN (1981). Solid State Convection in Icy Satellites: Effects of Phase Transitions Upon Stability (abstract). In *Reports of Planetary Geology Program*, pp. 59–61. NASA TM 84211.

REYNOLDS, R. T., AND P. M. CASSEN (1979). On the Internal Structure of the Major Satellites of the Outer Planets. *Geophys. Res. Lett.* **6**, 121–124.

RICHTER, F. M., AND D. MCKENZIE (1984). Dynamical Models for Melt Segregation from a Deformable Matrix. *J. Geol.* **92**, 729–740.

SCHENK, P. M., AND W. B. MCKINNON (1985). Dark Halo Craters and the Thickness of Grooved Terrain on Ganymede. *Proc. LPSC* **XVth**, in *J. Geophys. Res.* **90** (Suppl.), C775–C783.

SCHUBERT, G., AND D. L. TURCOTTE (1971). Phase Changes and Mantle Convection. *J. Geophys. Res.* **76**, 1424–1432.

SCHUBERT, G., D. A. YUEN, AND D. L. TURCOTTE (1975). Role of Phase Transitions in a Dynamic Mantle. *Geophys. J. Roy. Astr. Soc.* **42**, 705–735.

SCHUBERT, G., T. SPOHN, AND R. T. REYNOLDS (1986). Thermal Histories, Compositions and Internal Structures of the Moons in the Solar System. In *Satellites* (J. A. Burns and M. S. Matthews, Eds.), Univ. of Arizona Press, Tucson, pp. 224–292.

SCHUBERT, G., D. J. STEVENSON, AND K. ELLSWORTH (1981). Internal Structures of the Galilean Satellites. *Icarus.* **47**, 46–59.

SHOEMAKER, E. M., AND R. F. WOLFE (1982). Cratering Timescales for the Galilean Satellites. In *Satellites of Jupiter* (D. Morrison, Ed.), Univ. of Arizona Press, Tucson, pp. 277–339.

SHOEMAKER, E. M., B. K. LUCCHITTA, J. B. PLESCIA, S. W. SQUYRES, AND D. E. WILLIAMS (1982). The Geology of Ganymede. In *Satellites of Jupiter* (D. Morrison, Ed.), Univ. of Arizona Press, Tucson, pp. 435–520.

SMITH, C. S. (1948). Grains, Phases, and Interphases: An Interpretation of Microstructure. *Trans. Met. Soc. AIME.* **175**, 15–51.

SMITH, D. L., AND B. EVANS (1984). Diffusional Crack Healing in Quartz. *J. Geophys. Res.* **89**, 4125–4135.

SQUYRES, S. W. (198a). Volume Changes in Ganymede and Callisto and the Origin of Grooved Terrain. *Geophys. Res. Lett.* **7**, 593–596.

SQUYRES, S. W. (1980*b*). Topographic Domes on Ganymede: Ice Vulcanism or

Isostatic Upwarping. *Icarus.* **44**, 472–480.

SQUYRES, S. W. (1980*c*). Surface Temperatures and Retention of H_2O Frost on Ganymede and Callisto. *Icarus.* **44**, 502–510.

SQUYRES, S. W. (1982). The Evolution of the Tectonic Features on Ganymede. *Icarus.* **52**, 545–559.

STEVENSON, D. J. (1981). Models of the Earth's Core. *Science.* **214**, 611–619.

STEVENSON, D. J. (1982). Migration of Fluid-Filled Cracks: Applications to Terrestrial and Icy Bodies. *LPSC Abstracts.* **XIII**, 768–769.

STEVENSON, D. J., AND J. I. LUNINE (1986). Mobilization of Cryogenic Ice in Outer Solar System Satellites. *Nature.* **323**, 46–48.

THURBER, C. H., A. T. HSUI, AND M. N. TOKSOZ (1980). Thermal Evolution of Ganymede and Callisto: Effects of Solid-State Convection and Constraints from Voyager Imagery. *Proc. LPSC* **XIth**, 1957–1977.

TURCOTTE, D. L., AND E. R. OXBURGH (1967). Finite Amplitude Convective Cells and Continental Drift. *J. Fluid Mech.* **28**, 29–42.

TURCOTTE, D. L., AND G. SCHUBERT (1982). *Geodynamics*, John Wiley and Sons, NY, pp. 178–182.

WASSERBURG, G. J., G. J. F. MACDONALD, F. HOYLE, AND W. A. FOWLER (1964). Relative Contributions of Uranium, Thorium, and Potassium to Heat Production in the Earth. *Science.* **143**, 465–467.

WEAST, R., ED. (1976). *Handbook of Chemistry and Physics*, 56th ed. Chemical Rubber Company, Chicago.

WEERTMAN, J. (1970). The Creep Strength of the Earth's Mantle. *Rev. Geophys. Space Phys.* **8**, 145–148.

WEERTMAN, J. (1971). Theory of Water Filled Crevasses in Glaciers Applied to Vertical Magma Transport Beneath Ocean Ridges. *J. Geophys. Res.* **76**,

1171–1183.

WEERTMAN, J. (1973). Creep of Ice. In *Physics and Chemistry of Ice* (E. Whalley, S. J. Jones, and L. W. Gold, Eds.), Roy. Soc. of Canada, Ottowa, pp. 320–337.

WEERTMAN, J. (1983). Creep Deformation of Ice. *Ann. Rev. Earth Planet. Sci.* **11**, 215–240.

WHITEHOUSE, U. G., L. M. JEFFREY, AND J. D. DEBBRECHT (1959). Differential Settling Tendencies of Clay Minerals in Saline Waters. *Proc. Conf. on Clays and Clay Minerals.* **7**, 1–74.

PAPER II

Hydromagnetic Constraints on Deep Zonal Flow in the Giant Planets

The nature of the world which, motionless
At core, the wheeling of the rest maintains,
Starteth from here the running of the race...

— Dante Alighieri *Paradiso*, XXVII, 106–108

Hydromagnetic Constraints on Deep Zonal Flow in the Giant Planets

R. L. KIRK AND D. J. STEVENSON

Division of Geological and Planetary Sciences

California Institute of Technology

Pasadena, California 91125

Published in modified form in *Astrophys. J.*

March, 1987

Contribution number 4337 from the Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125.

Abstract

The observed zonal flows of the giant planets will, if they penetrate below the visible atmosphere, interact significantly with the planetary magnetic field outside the metalized core. The appropriate measure of this interaction is the Chandrasekhar number $Q = \frac{H^2}{4\pi\rho\nu\alpha^2\lambda}$ (where H = radial component of the magnetic field, ν = eddy viscosity, λ = magnetic diffusivity, α^{-1} = lengthscale on which λ varies); at depths where $Q \gtrsim 1$ the velocity will be forced to oscillate on a small lengthscale or decay to zero. We estimate the conductivity due to semiconduction in H_2 (Jupiter, Saturn) and ionization in H_2O (Uranus, Neptune) as a function of depth; the value $\lambda \simeq 10^{10}\,\mathrm{cm^2\,s^{-1}}$ needed for $Q = 1$ is readily obtained well outside the metallic core (where $\lambda \simeq 10^2\,\mathrm{cm^2\,s^{-1}}$).

These assertions are quantified by a simple model of the equatorial zonal jet in which the flow is assumed uniform on cylinders concentric with the spin axis, and the viscous and magnetic torques on each cylinder are balanced. We solve this "Taylor constraint" simultaneously with the dynamo equation to obtain the velocity and magnetic field in the equatorial plane. With this model we reproduce the widely differing jet widths of Jupiter and Saturn (though not the flow at very high or low latitudes) using $\nu = 2500\,\mathrm{cm^2\,s^{-1}}$, consistent with the requirement that viscous dissipation not exceed the specific luminosity. A model Uranian jet consistent with the limited Voyager data can also be constructed, with appropriately smaller ν, but only if one assumes a two-layer interior. We tentatively predict a wide Neptunian jet.

For Saturn (but not Jupiter or Uranus) the model has a large magnetic Reynolds number where $Q = 1$ and hence exhibits substantial axisymmetrization of the field *in the equatorial plane.* This effect may or may not persist at higher latitudes. The one-dimensional model presented is only a first step. Variation of the velocity and magnetic field parallel to the spin axis must be modeled in order to answer several

important questions, including: 1) What is the behavior of flows at high latitudes, whose Taylor cylinders are interrupted by the region with $Q \gtrsim 1$? 2) To what extent is differential rotation in the envelope responsible for the spin-axisymmetry of Saturn's magnetic field?

1. Introduction

In Earth and other terrestrial planets, we are accustomed to a clear delineation between the highly conducting, low viscosity region (the core) in which hydromagnetics are important and an outer very poorly conducting, high viscosity region (the mantle) in which hydromagnetics are unimportant. Nobody seriously advocates important hydromagnetic effects in plate tectonics, oceanography or lower atmosphere motions. In the Sun and many other stars, we are equally accustomed to the idea that hydromagnetic effects can be important in the observable atmosphere (*e.g.*, in sunspots). The giant planets do not easily conform to either of these limits. There is little doubt that Jupiter and Saturn possess highly conducting metallic hydrogen cores (Stevenson 1982), but there is also the likelihood that molecular hydrogen regions only a small distance (few thousand kilometers) below the atmosphere are sufficiently conducting to have significant hydromagnetic effects. The central question is this: To what extent are the observed atmospheric flows (the zonal winds) affected by or even determined by the planetary magnetic field? We offer here only a partial and qualified answer to this question, but an interesting answer nonetheless because it suggests a connection between surficial winds, deep-seated flows, and the planetary dynamo.

The ideas are not entirely new. Hide (1965) suggested that the field of Jupiter may be generated near the surface, and Smoluchowski (1972; 1975) pointed out the likely semiconducting properties of impure or even pure molecular hydrogen at high pressure and temperature. An attempt has even been made to determine the depth of the field-generating region in Jupiter (Hide and Malin 1979), but the data are

Figure 1.1. Schematic view of deep zonal flow in Jupiter and Saturn . The surface flow extends into the interior on concentric Taylor cylinders but is excluded from a magnetically interacting region at depth.

MAGNETIC
CORE
SATURN

MAGNETIC
CORE
JUPITER

insufficient for a convincing result. On the other hand, many workers have assumed that the entire region external to the metallic hydrogen cores of Jupiter and Saturn can be treated as an insulating fluid (*i.e.*, using the Navier-Stokes equation without the Lorentz force). Busse (1976; 1983) proposed that the surficial structure of clouds or winds may be directly matched to the columnar convective cells expected deep within rapidly rotating, adiabatic fluid planets. Ingersoll and coworkers, motivated largely by a perceived difficulty in explaining the winds by thin shell meteorology, have pursued the related idea that the observed winds are the surface expression of zonal flows on cylindrical surfaces (Smith *et al.* 1982; Ingersoll and Pollard 1982; Ingersoll and Miller 1986). This is illustrated in Figure 1.1. The arguments against confining the winds to a thin shell have become less compelling because of the possibly large role of latent heat effects in the fluid motions (Allison and Stone 1983; Conrath and Gierasch 1984). Although no fully quantitative dynamic theory exists, deep-seated zonal flows still remain an attractive hypothesis because these planets have bottomless atmospheres and very stable wind patterns. Ironically, the work reported here provides support for moderately deep-seated flows, yet invalidates the original views of Busse and Ingersoll, who envisaged columnar or cylindrical flows that completely filled the region external to the metallic hydrogen core.

Our basic ideas are these: A deep zonal flow has a differential rotation that leads to a generation of toroidal field $H_\phi \sim R_m H_r$, where H_r is the imposed radial field, $R_m \sim \frac{v\ell}{\lambda}$ is the magnetic Reynolds number, v a characteristic azimuthal flow velocity, ℓ some lengthscale (ill-defined, as yet), and λ is the magnetic diffusivity. Even for a conductivity tens orders of magnitude less than that of copper at room temperature, $\lambda \simeq 10^{12}\,\mathrm{cm^2\,s^{-1}}$, $R_m \sim 1$ for $v \simeq 10^4\,\mathrm{cm\,s^{-1}}1$ (typical of Jupiter and Saturn) and $\ell \simeq 10^8\,\mathrm{cm}$. The toroidal field has an associated poloidal current which, when crossed with the radial field, yields a Lorentz force with an azimuthal compo-

nent $\sim \frac{R_m H_r^2}{4\pi\rho\ell}$ per unit mass (ρ is the fluid density). Since there cannot be a net azimuthal torque on a cylinder of fluid in steady flow (Taylor 1963), this force must be balanced by a "viscous" force $\sim \frac{\nu v}{\ell^2}$, where ν is the kinematic eddy viscosity. It follows that we require $\frac{H_r^2 \ell^2}{4\pi\rho\nu\lambda} \sim 1$. This dimensionless number was first introduced by Chandrasekhar (1965) although for different reasons. As we go down into the planet, the conductivity increases and λ decreases, so this requirement translates into a progressively smaller ℓ. In effect, the zonal flow is forced to have large shears. Our thesis is that this requirement imposed by the Chandrasekhar number implies a rapid drop-off in the zonal flow and thereby limits the width of the equatorial jet in giant planets. To put it another way, if these planets did not have magnetic fields, then the observed equatorial jet would extend to much higher latitudes, corresponding to deeper flows. Some aspects of this model were independently developed (but not quantified) by Drobyshevskiĭ (1979*a*;*b*). Here, we attempt a quantitative model.

Clearly, the biggest uncertainties lie in the diffusivities λ and ν, which could range over many orders of magnitude. The value of λ is computed in Section 2, using semiquantitative theories of liquid semiconductors, published band structure calculations of molecular hydrogen, experimental results for the conductivity of water, and published temperature-density structures of giant planets. The value of ν might seem to be much more uncertain because it is not likely to be the very small intrinsic fluid value ($\sim 10^{-2}\,\mathrm{cm^2\,s^{-1}}$) but is a crude representation of the nonlinear effects of the flow. However, it is bounded above by the requirements of the first and second laws of thermodynamics: the local viscous dissipation $\sim \nu \left(\frac{v}{\ell}\right)^2$ should not greatly exceed the total planetary thermal energy loss per unit mass ($\simeq 10^{-6}\,\mathrm{erg\,g^{-1}\,s^{-1}}$ in Jupiter and Saturn), so $\nu \lesssim 10^3\,\mathrm{cm^2\,s^{-1}}$. In fact, careful scaling arguments (Ingersoll and Pollard 1982) give a value of this order. A Chandrasekhar number of order unity then typically corresponds to the level in the planet at which $\lambda \simeq 10^{10}\,\mathrm{cm^2\,s^{-1}}$. Our thesis

is that the spin-aligned cylinder circumscribing the sphere on which this conductivity is obtained must intercept the planetary surface at the latitude corresponding to the outer extremities of the equatorial zonal jet. In this way, we can reproduce the observed widths of the jets on Jupiter, Saturn, and possibly Uranus. In Section 3, we develop the mathematical theory to support the above heuristic arguments, showing how the Taylor constraint leads to the identification of a Chandrasekhar number. The model is applied to Jupiter and Saturn in Section 4 and to Uranus and Neptune in Section 5. We end in Section 6 with some comments on limitations and possible future work.

2. The Magnetic Diffusivity

In hydromagnetics, it is conventional to characterize the electrical conductivity, σ, in terms of the magnetic diffusivity $\lambda \equiv \frac{c^2}{4\pi\sigma}$, where c is the speed of light and σ is in e.s.u. units (s^{-1}). A resistivity of 1 $\mu\Omega$ cm is equivalent to $\lambda = (\frac{250}{\pi})\,cm^2\,s^{-1}$. Typical values of λ are 10^2–$10^3\,cm^2\,s^{-1}$ (good metals), $10^6\,cm^2\,s^{-1}$ (good electrolytes), and $\sim 10^{12}\,cm^2\,s^{-1}$ for pure or nearly pure germanium at 500 K. Molecular H_2 is effectively an insulator at low pressure (band gap $E_g \simeq 10$ eV), but this gap is believed to diminish progressively as the pressure increases. Although diamond cells (without a hydrogen sample) have now achieved in excess of 4 Mbar (Xu *et al.* 1986; Goettel *et al.* 1986) and quantitative experiments on H_2 at ~ 1.5 Mbar have been reported (Mao *et al.* 1985), there are no data on the band gap, except for the inference that the band gap is still finite at the highest pressures attained. The much discussed and anticipated transition to monatomic (alkali metal) hydrogen, conventionally called "metallic hydrogen," probably occurs at much higher pressures still (perhaps $\sim$ 3–4 Mbar; see Ross 1985; Min *et al.* 1986) but has no bearing on the issues addressed in this paper.

We rely here on theoretical calculations for the band gap in crystalline H_2

(Friedli and Ashcroft 1977; Min *et al.* 1986). These results can be well represented by the empirical formula:

$$E_g = 32.5\rho^{\frac{2}{3}}(x + \epsilon x^2)\,\mathrm{eV}, \qquad \text{where} \qquad x \equiv \left(\frac{\rho_0}{\rho}\right)^{\frac{1}{3}} - 1 \tag{2.1}$$

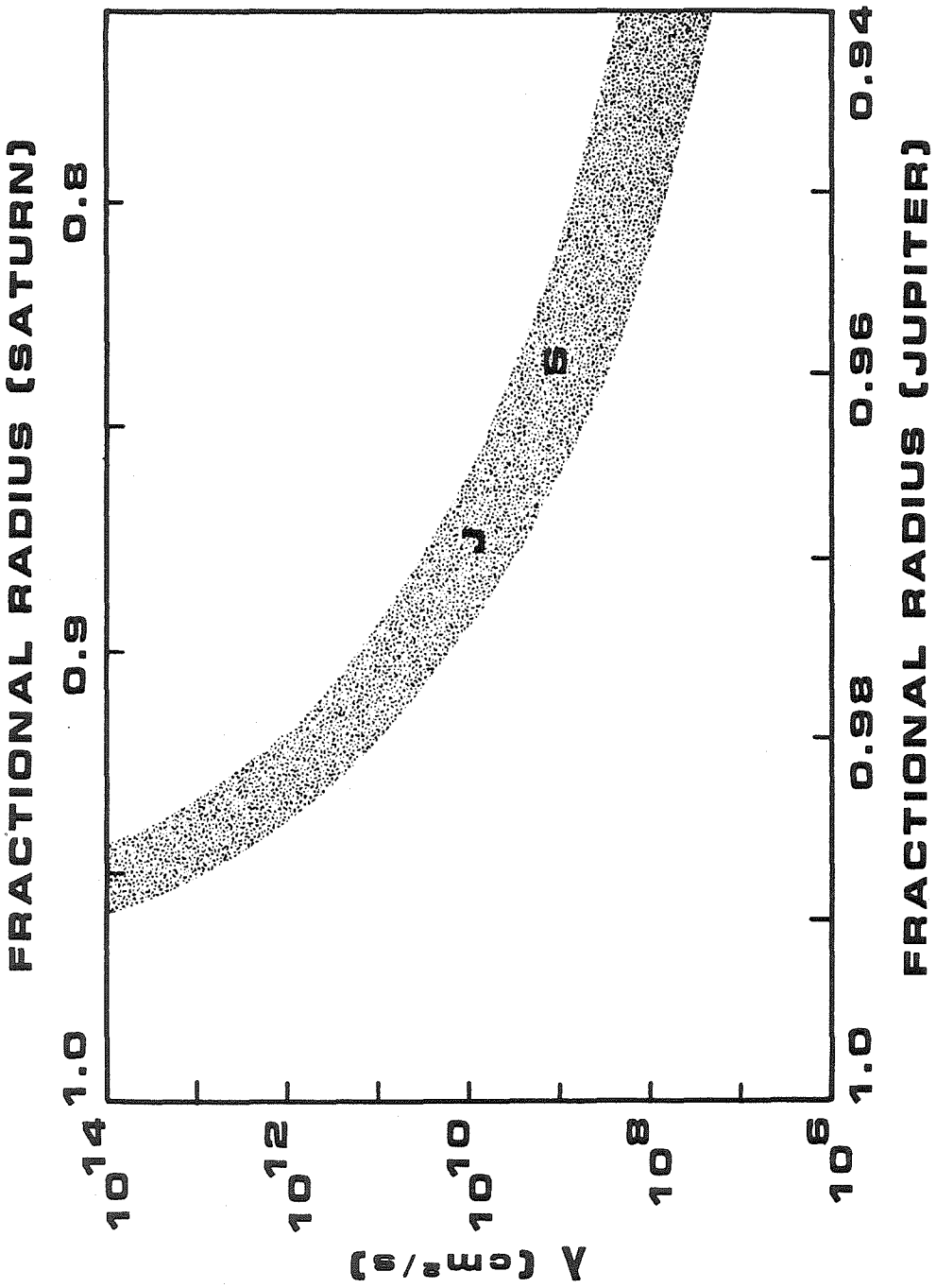
and ρ is the density in $\mathrm{g\,cm^{-3}}$, ρ_0 is the density at band closure ($0.81\,\mathrm{g\,cm^{-3}}$ for Friedli and Ashcroft, $0.90\,\mathrm{g\,cm^{-3}}$ for Min et al.), $\epsilon = 0.3$ for Friedli and Ashcroft and 0.2 for Min et al. Although there are differences between the two results, they are most striking near band closure, a high pressure region that does not interest us. A more serious concern is the use of a crystalline calculation to describe a liquid; evidence for liquid semiconductors suggests that the appropriate band gap or mobility gap is probably slightly smaller. Our calculation may therefore be conservative.

In a semiconductor, the fractional occupancy of current carrying states is proportional to $\exp\left(\frac{-E_g}{2kT}\right)$, where k is Boltzmann's constant and T is the temperature. The factor of two in the denominator of the exponent is an unavoidable consequence of the law of mass action. It follows that, in general,

$$\lambda = \lambda_0 \exp(\frac{E_g}{2kT}), \tag{2.2}$$

where λ_0 can be a function of density and temperature. Smoluchowski (1975) chose to use the standard crystalline semiconductor result (*e.g.*, Ashcroft and Mermin 1976) in which $\lambda_0 \propto T^{3/2}$ with no density dependence. We shall adopt the semiempirical results of the theory of Mott (1971) for liquid semiconductors, according to which $\lambda_0 \simeq 10^5\,\mathrm{cm^2\,s^{-1}}$, roughly independent of temperature. The two approaches disagree by an order of magnitude or less at the densities and temperatures of interest. In the low density limit, λ_0 should eventually approach the prediction of dilute gas theory (Chapman and Cowling 1952), which we estimate to be $\sim 10^4\rho^{1/3}\left(\frac{10^3\,\mathrm{K}}{T}\right)\,\mathrm{cm^2\,s^{-1}}$, where ρ is in $\mathrm{g\,cm^{-3}}$. We do not use this result, but its approximate consistency with our adopted value of $10^5\,\mathrm{cm^2\,s^{-1}}$ indicates no serious extrapolation difficulties.

Figure 2.1. Magnetic diffusivity of the Jovian and Saturnian envelopes . The pressure-induced semiconductivity of pure H_2 was calculated according to equations (2.1) and (2.2); the shaded region indicates the uncertainty due to the input data.


FRACTIONAL RADIUS (SATURN)
1.0
0.9
0.8
FRACTIONAL RADIUS (JUPITER)
1.0
0.98
0.96
0.94
10^{14}
10^{12}
10^{10}
10^{8}
10^{6}
λ (cm²/s)
1
5
7

The temperature and density profiles within Jupiter and Saturn are obtained from published interior models which differ little in the range of interest (Stevenson and Salpeter 1976; Hubbard and Horedt 1983; Hubbard and Stevenson 1984). At a given fraction of the outer radius, Saturn is much colder than Jupiter, mainly because it has a lower gravitational acceleration but partly because it has a lower specific entropy (*i.e.*, colder atmosphere). However, they are both adiabatic planets, so that the resulting functional dependence of λ on fractional planetary radius is the same for the two planets, except for a scale factor.

This is exhibited in Figure 2.1, based on calculations using equations (2.1), (2.2), and the planetary models. The shaded region is an attempt to indicate the combined uncertainties of all the inputs but does not include systematic errors (*e.g.*, the possibility that the mobility gap is systematically smaller than E_g as given by eqn. 2.1).

It is also possible that Figure 2.1 systematically overestimates the true value of λ because of impurities mixed with the hydrogen. However, we doubt that atoms with small ionization energies, such as sodium, are present in a chemically unbound form (as assumed by Smoluchowski 1972). In order of decreasing abundance (approximate mole fractions in brackets), the impurities are expected to be He (0.1), H_2O ($\sim 10^{-3}$), CH_4 ($\sim 10^{-3}$), NH_3 ($\sim 10^{-4}$), Ne ($\sim 10^{-4}$), silicate and iron particles ($\sim 10^{-4}$–10^{-5}). None of these appear likely to overwhelm the conductivity at $T \simeq 3000\,\mathrm{K}$ (where $\lambda \simeq 10^{10}\,\mathrm{cm^2\,s^{-1}}$), the region of most importance for our hydromagnetic effects. Accordingly, we use the results for pure H_2.

Our approach to the crude estimates for Uranus and Neptune is different and more closely tied to the data. The conductivity of the envelopes of these planets is controlled by the "ice" component, specifically H_2O (shockwave experiments suggest that NH_3 has about an order of magnitude lower conductivity under similar

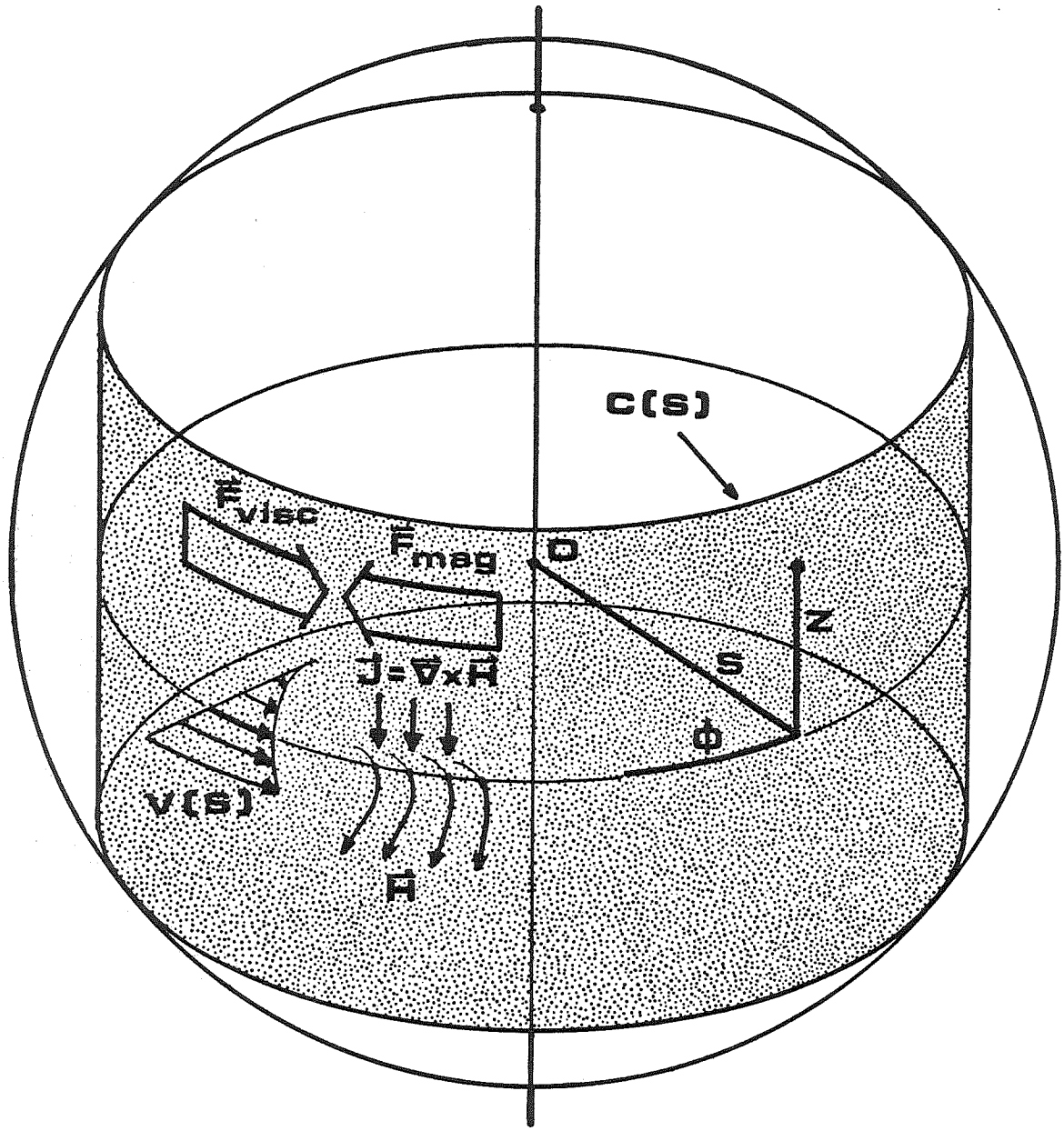
conditions; Ross *et al.* 1981) — regardless of one's assumptions about the internal structure. At $T \lesssim 2000\,\mathrm{K}$, the "gas" layer of three-layer models (Hubbard and MacFarlane 1980) is too cold to be significantly conductive according to equations (2.1) and (2.2), and Lorentz forces first become important in the "ice" layer. On the other hand, unpublished two-layer models by one of us suggest an ice to gas ratio of $\sim 2.5 : 1$ in the envelope, and under these circumstances the icy component will dominate the conductivity at any given depth. Accordingly, we compute the conductivity of H_2O using Holtzapfel's (1969) fit to static and shock wave data at temperatures up to 1000°C and pressures to 100 kbar. We account for the probable suppression of ionization by dissolved hydrogen-helium in the two-layer model crudely but conveniently, by adopting a magnetic diffusivity an order of magnitude greater than that of pure water.

3. The Taylor Constraint

A complete dynamical model of giant planet envelopes that includes hydromagnetic effects is obviously beyond the scope of this paper. We instead present a simple one-dimensional model illustrating the effect of inward-increasing conductivity on a deep-seated pattern of differential rotation.

The main assumptions made are, first, that the mean flow is zonal and depends only on cylindrical radius ($\mathbf{u} = v(s)\hat{e}_\phi$ in the (s, ϕ, z) cylindrical coordinates of Figure 3.1), as guaranteed by the Taylor-Proudman theorem for an inviscid, isentropic, nonconducting medium; second, that the effects of turbulence and convection may be parameterized by a constant eddy viscosity (of either sign as yet); third, that magnetic forces can be expanded about the equatorial plane $z = 0$ (where, for given s, the conductivity is greatest); and, finally, that the magnetic diffusivity varies exponentially with depth: $\lambda = \lambda_0 \exp(\alpha\sqrt{s^2 + z^2})$. (In applying our model to the giant planets we will linearize the theoretical prediction for $\ln \lambda$ about a point near where

Figure 3.1. The Taylor constraint (equation 3.1) is a balance between the integrated viscous and magnetic torques on a cylindrical surface (Taylor column) $\mathcal{C}(s)$ in a differentially rotating planet. The balance is illustrated for a positive eddy viscosity, giving a monotonic zonal velocity $v(s)$. The (s, ϕ, z) cylindrical coordinates are replaced with local (x, y, z) Cartesian coordinates for the actual calculations.


C(S)
$\vec{F}_{visc}$
$\vec{F}_{mag}$
D
Z
S
$\vec{J}=\vec{\nabla}\times\vec{H}$
Φ
V(S)
$\vec{H}$

$v \to 0$.)

For such a system in differential rotation on cylinders, the Taylor constraint (Moffatt 1978) states that the sum of magnetic and viscous torques on a cylindrical surface $\mathcal{C}(s)$ vanishes:

$$\oint_{\mathcal{C}(s)} \left(\frac{(\nabla \times \mathbf{H}) \times \mathbf{H}}{4\pi\rho} \pm \nu \nabla^2 \mathbf{u} \right) \cdot \hat{e}_\phi s^2 \, d\phi \, dz = 0, \tag{3.1}$$

where we write the eddy viscosity as $\pm\nu$ with $\nu > 0$. Local variations of the magnetic torque from its mean value on $\mathcal{C}(s)$ will be balanced by pressure gradients which do not enter our analysis.

The Taylor constraint must be supplemented by the dynamo equation for the evolution of the magnetic field, in steady state:

$$\frac{\partial \mathbf{H}}{\partial t} = -\nabla \times (\lambda \nabla \times \mathbf{H}) + \nabla \times (\mathbf{u} \times \mathbf{H}) = 0. \tag{3.2}$$

We will also use the solenoidal character of $\mathbf{H}$:

$$\nabla \cdot \mathbf{H} = 0. \tag{3.3}$$

Rather than simplifying equation (3.1) by formally expanding $\mathbf{H}$ about $z = 0$, we present here a much simpler derivation in Cartesian coordinates, appropriate to the equatorial plane far from the rotation axis, and with $\frac{\partial}{\partial z} \equiv 0$. The essential features of the problem are more clearly exhibited, and we readily obtain an analytic solution agreeing to lowest order in $\left| \frac{s \, \partial \lambda}{\lambda \, \partial s} \right|^{-1}$ with that obtained by expansion in cylindrical coordinates — provided we measure the velocity in the latter case with respect to solid-body rotation.

In (x, y, z) coordinates with $\frac{\partial}{\partial z} = 0$ the vector equations (3.2) and (3.3) hold, but the equivalent of the Taylor constraint is a force balance on a plane of constant x:

$$\oint \left(\frac{(\nabla \times \mathbf{H}) \times \mathbf{H}}{4\pi\rho} \pm \nu \nabla^2 \mathbf{u} \right) \cdot \hat{e}_y \, dy = 0, \tag{3.4}$$

where the integral in y is over one cycle of the periodic function $\mathbf{H}$. The vertical integral is obviated by the assumption $\frac{\partial}{\partial z} = 0$; in applying the solution we will assume that the viscous force is constant, while the Lorentz term varies on a given cylinder $\mathcal{C}(s)$ in proportion to σ. The result is merely to multiply the first term of equation (3.4) by a weakly varying function $\delta(s) = \sqrt{\frac{\pi s}{2\alpha(R^2 - s^2)}}$, since the integral over exponentially varying s is readily performed.

The Lorentz term in the Taylor constraint may be expanded:

$$\begin{aligned}((\nabla \times \mathbf{H}) \times \mathbf{H})_y &= (\nabla \times \mathbf{H})_z H_x - (\nabla \times \mathbf{H})_x H_z \\ &= \left(\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y}\right) H_x - \frac{\partial H_z}{\partial y} H_z. \end{aligned} \tag{3.5}$$

The term involving H_z is a total differential in y and hence does not contribute to the integral (3.4). The first term may be rewritten by integrating the dynamo equation (3.2) to give

$$\nabla \times \mathbf{H} = \frac{\mathbf{u} \times \mathbf{H}}{\lambda} + \nabla\phi, \tag{3.6}$$

where $\phi(x, y)$ is an arbitrary function of integration (fortunately not appearing in the z-component of the equation, in which we are interested). These results may be used to simplify the Taylor constraint to yield (with $\mathbf{u} = v(x)\hat{e}_y$):

$$\left[\frac{d^2}{dt^2} \mp \frac{\langle H_x \rangle^2 \delta}{4\pi\rho\nu\lambda(x)}\right] v = 0. \tag{3.7}$$

As usual, the upper sign corresponds to a positive eddy viscosity, and the root mean squared value of H_x is $\langle H_x \rangle = \left(\frac{\oint H_x^2\, dy}{\oint dy}\right)^{1/2}$. Writing $H = H(x)e^{iky}$+ complex conjugate, with complex $H(x)$, we have $\langle H_x \rangle = \sqrt{p}|H_x|$ ($p = 1$ if $k = 0$, $p = \frac{1}{2}$ otherwise). In cylindrical coordinates $\mathbf{H} \propto e^{im\phi}$+c.c., so we identify $k = \frac{m}{\bar{s}}$, with $\bar{s}$ a typical cylindrical radius and $m = 0$, 1 the harmonics of greatest interest.

With the above assumptions about $\mathbf{u}$ and $\mathbf{H}$, and $\lambda = \lambda_0 e^{\alpha x}$, the components of the dynamo equation become:

$$\left[\lambda\left(\frac{d^2}{dx^2} - k^2\right) - ikv\right] H_x = 0, \tag{3.8a}$$

$$\left[\lambda\left(\frac{d^2}{dx^2}+\alpha\frac{d}{dx}-k^2\right)\right]H_y=\left[ik\alpha\lambda-\frac{dv}{dx}\right]H_x, \tag{3.8b}$$

$$\left[\lambda\left(\frac{d^2}{dx^2}+\alpha\frac{d}{dx}-k^2\right)-ikv\right]H_z=0. \tag{3.8c}$$

It may occur (*e.g.*, in the case of Saturn and Uranus) that multiple harmonic components $\mathbf{H}_m$ of the magnetic field with differing wavenumbers k_m are important. Each component, of course, satisfies the linear dynamo equation individually, but all must be included in the Taylor constraint. Because the integral in equation (3.4) is over a whole number of periods of each component, cross-terms of the form $(\nabla\times\mathbf{H})_m\times\mathbf{H}_{m'}$ vanish for $m\neq m'$, and we can replace $\langle H_x\rangle^2$ by $\sum_m\langle H_{x,m}\rangle^2$ in equation (3.7).

We nondimensionalize equations (3.7) and (3.8a) in terms of $\xi\equiv\alpha(x-x_0)$, $h_x\equiv\frac{H_x\sqrt{p}}{H_0}$ (where $H_0\equiv\lim_{x\to\infty}\langle H_x\rangle$), and $u=\frac{\alpha v}{\omega}$ (where $\omega\equiv\alpha\lim_{x\to\infty}v$ or $\omega\equiv\lim_{x\to\infty}\frac{dv}{dx}$, whichever is finite). The boundary conditions as $\xi\to\infty$ are thus $h_x=1$ (the phase of h_x is arbitrary) and $u=1$ or $Du=1$ where $D\equiv\frac{\partial}{\partial\xi}$. There are then three dimensionless parameters in the problem: $K=\frac{k}{\alpha}$ ($\ll 1$ by assumption), the Chandrasekhar number $Q=\frac{H_0^2\delta}{4\pi\rho\nu\alpha^2\lambda(x_0)}$ (cf. Chandrasekhar 1965) and the magnetic Reynolds number $R_m=\frac{\omega}{\alpha^2\lambda(x_0)}$. The Chandrasekhar number expresses the importance of magnetic forces in determining v, and because of the variation of λ we can always choose x_0 sufficiently deep that $Q=1$ there. The magnetic Reynolds number, on the other hand, expresses the importance of v in modifying the magnetic field. Once we have chosen x_0, $R_m=\frac{4\pi\rho\nu\omega}{H_0^2\delta}$, which may be of any magnitude. (In the case of multiple field harmonics of comparable strength, we make the obvious generalization of H_0^2 to $\sum_m H_{0,m}^2$ in these formulae.) Note that H_0 and ω are not predicted by the model but must be given as boundary conditions. Prediction of these quantities would require descriptions, respectively, of the regenerative dynamo, and of the nonlinear interaction between convection and the zonal mean flow in the nonmagnetohydrodynamic limit (for which no satisfactory theory currently exists).

The appropriate nondimensionalization of the remaining field components is

$h_y = \frac{H_y\sqrt{p}}{R_m H_0}$ and $h_z = \frac{H_z\sqrt{p}}{\lim\limits_{x\to\infty}\langle H_z\rangle}$, with boundary conditions $h_y = 0$ (no external "toroidal" field) and $h_z = 1$ as $\xi \to \infty$. The full set of nondimensional equations (for $D^2 \gg KD$) is:

$$\left(D^2 \mp e^{-\xi}|h_x|^2\right) u = 0, \tag{3.9a}$$

$$\left(D^2 - iKR_m e^{-\xi} u\right) h_x = 0, \tag{3.9b}$$

$$\left(D^2 + D\right) h_y = \left(-e^{-\xi} Du\right) h_x, \tag{3.9c}$$

$$\left(D^2 + D - iKR_m e^{-\xi}\right) u = 0, \tag{3.9d}$$

In the limit $KR_m = 0$, the above equations decouple and an analytic solution may be obtained. Equations (3.9b,d) reduce to $D^2 h_x = (D^2+D)h_z = 0$ so $h_x = h_z = 1$ (we exclude a solution $h_z \propto e^{-\xi}$ due to "leakage" of currents from the dynamo region at great depth; its amplitude should be small). Equation (3.9a) thus becomes:

$$\left(D^2 \mp e^{-\xi}\right) u = 0, \tag{3.10}$$

or

$$\zeta^2 \frac{d^2u}{d\zeta^2} + \zeta \frac{du}{d\zeta} \mp \zeta^2 u = 0, \qquad \text{where} \qquad \zeta \equiv 2e^{-\xi/2}. \tag{3.11}$$

The solutions are $u = \mathcal{C}_0(\sqrt{\mp 1}\zeta)$ (Abramowitz and Stegun 1965), where $\mathcal{C}_0$ is any Hankel or Bessel function of order zero.

Converting to real argument and applying the boundary conditions (including $u < \infty$ as $\xi \to -\infty$, which excludes the solution proportional to $I_0(z)$ for $+\nu$) we find three possible cases, illustrated in Figure 3.2 (solid curves):

$$u(\xi) = \begin{cases} 2K_0(2e^{-\xi/2}), & \text{for } +\nu,\ Du(\infty) = 1, \quad (3.12a) \\ J_0(2e^{-\xi/2}), & \text{for } -\nu,\quad u(\infty) = 1, \quad (3.12b) \\ cJ_0(2e^{-\xi/2}) - \pi Y_0(2e^{-\xi/2}), & \text{for } -\nu,\ Du(\infty) = 1, \quad (3.12c) \end{cases}$$

where c is an arbitrary constant ($c = 0$ is illustrated). We refer to these solutions generically as $u = c\mathcal{F}_0(2e^{-\xi/2})$. The exclusion of the fourth possibility (which has

Figure 3.2. Dimensionless solutions for the zonal velocity . For $KR_m = 0$ (solid curves) three analytic solutions (equations 3.12) are shown: for positive eddy viscosity one monotonic solution $2K_0(\zeta)$ where $\zeta = 2e^{-\xi/2}$, and for negative viscosity two oscillatory solutions $J_0(\zeta)$ and $-\pi Y_0(\zeta)$. For $KR_m = 10$ (dotted curves) three numerical solutions are shown, similar in character but displaced to the right because of the inward amplification of the magnetic field h_x.

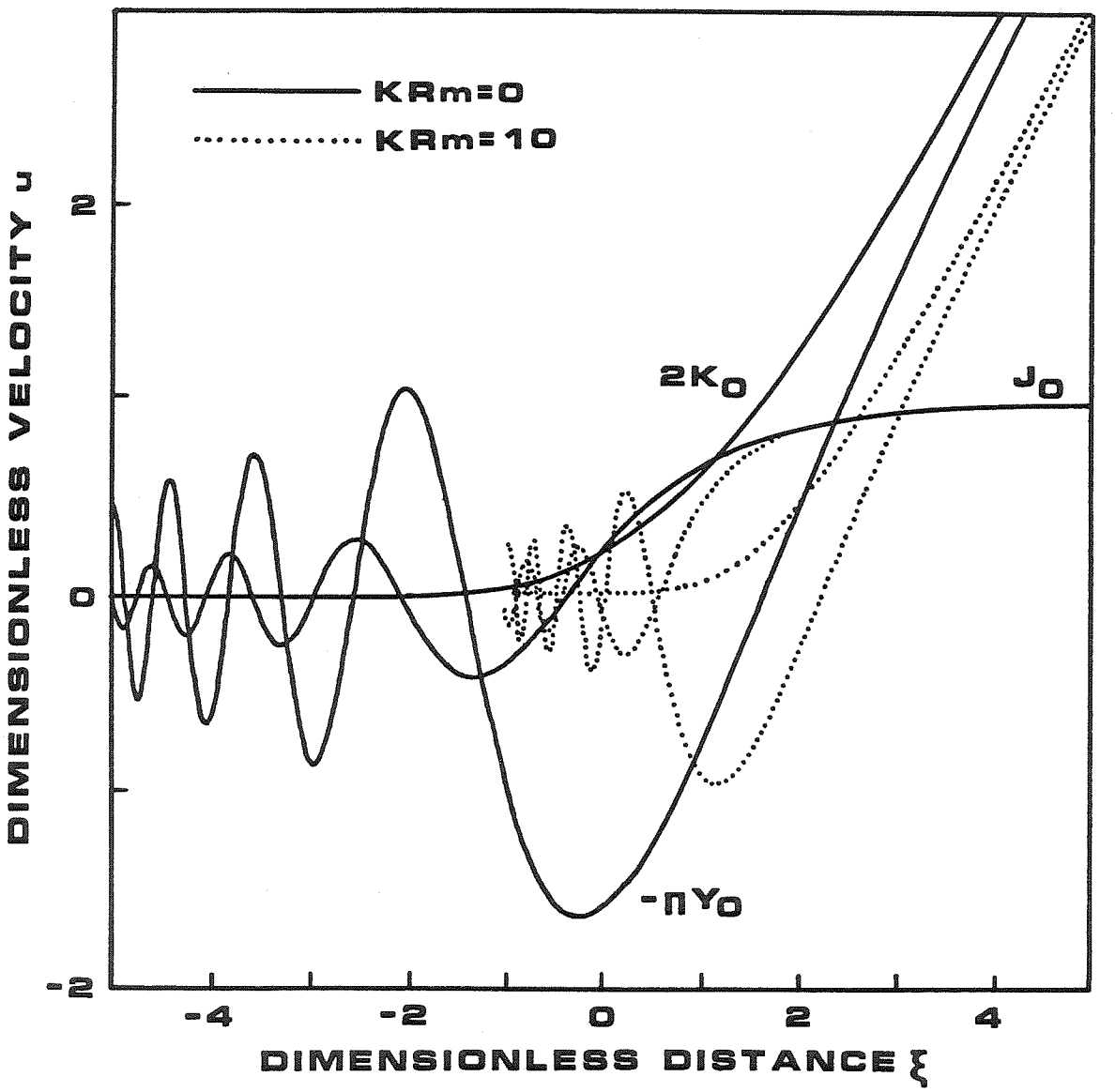
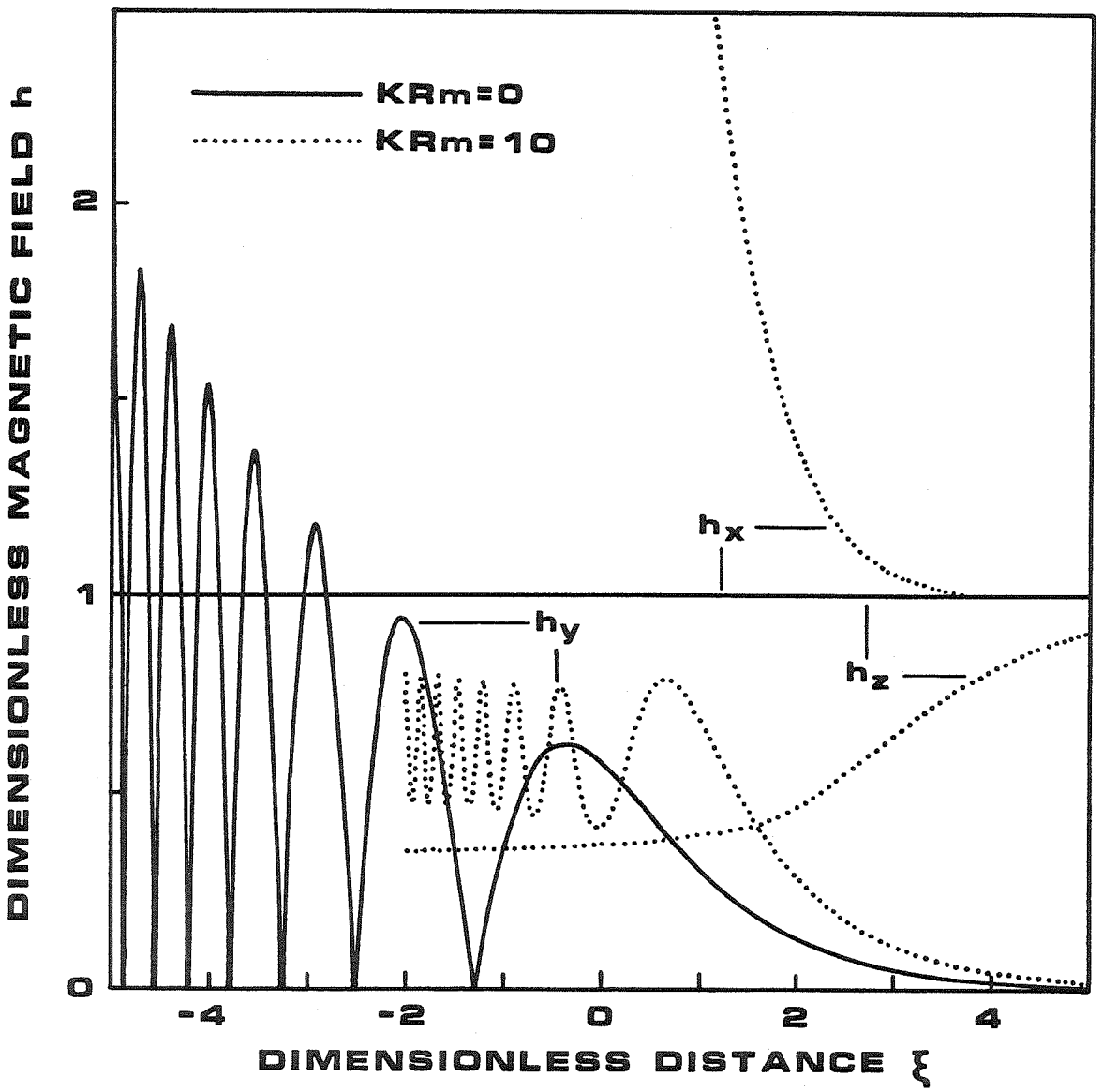

KRm=0
KRm=10
2K0
J0
-ΠY0
DIMENSIONLESS VELOCITY u
DIMENSIONLESS DISTANCE ξ
2
0
-2
-4
-2
0
2
4

Figure 3.3. Dimensionless solutions for the magnetic field , corresponding to the J_0 analytic and J_0-like velocity solutions of Figure 3.2. For $KR_m = 0$ (solid curves) the "radial" field h_x and "vertical" field h_z are unaffected by the motion but a substantial "toroidal" field h_y is induced. For $KR_m = 10$ (dotted curve) h_x is inward-amplified and h_z inward-attentuated. These field components can be thought of as the first term in an expansion about the equatorial plane.


KRm=0
KRm=10
h_x
h_y
h_z
DIMENSIONLESS MAGNETIC FIELD h
DIMENSIONLESS DISTANCE ξ
0
1
2
-4
-2
0
2
4

$Du(\infty) = 0$) means that the condition of zero viscous stress at the exterior cannot be met for a positive eddy viscosity. For $+\nu$, a flux of energy and momentum from the outside is required to sustain the motion against viscous and Ohmic losses. In contrast, $-\nu$ permits "viscous" extraction of convective energy to balance Ohmic losses so that an energy/momentum input is not needed (though it can be accommodated). Indeed, the simplest description of the energy supply needed for the solution (3.12a) is a region (outside the portion of the planet modeled) of negative eddy viscosity. For these reasons we consider the torque-free solution (3.12b) to be the one of greatest relevance and importance.

To complete the solution when $KR_m = 0$, we note that equation (3.9c), $(D^2 + D)h_y = e^{-\xi}Du$, may be integrated twice (again, we exclude an exponentially decaying "leakage" field by choosing the constant $u_0 = 0$ below) to give

$$\begin{aligned} h_y(\xi) &= \int_{\xi}^{\infty} (u(\xi') - u_0)\, d\xi' \\ &= \frac{1}{2}\int_0^{2e^{-\xi/2}} \zeta c\mathcal{F}_0(\zeta)\, d\zeta \\ &= ce^{-\xi/2}\mathcal{F}_1(2e^{-\xi/2}). \end{aligned} \tag{3.13}$$

Figure 3.3 illustrates the three field components for the velocity solution $u = J_0(\zeta)$.

From our solutions, one can calculate the Ohmic dissipation per unit volume $\dot{E}_{Ohmic} = \frac{1}{\sigma}\left\langle \left(\frac{c}{4\pi}\nabla \times \mathbf{H}\right)\right\rangle^2$ and viscous dissipation (or energy *release* $\dot{E}_{visc} = \pm\rho\nu\left(\frac{dv}{dx}\right)^2$, averaged over one period in y. Defining $P = \rho\nu\omega^2$ and making use of $Q = 1$, we find

$$\dot{E}_{Ohmic} = \quad P\left(2e^{-\xi/2}\frac{c}{2}\mathcal{F}_0(2e^{-\xi/2})\right)^2, \tag{3.14a}$$

and

$$\dot{E}_{visc} = \pm P\left(2e^{-\xi/2}\frac{c}{2}\mathcal{F}_1(2e^{-\xi/2})\right)^2, \tag{3.14b}$$

illustrated in Figure 3.4 for the $2K_0$ and J_0 solutions. For $+\nu$, both functions are positive and decay double-exponentially as $\xi \to \infty$; we find $\int_{-\infty}^{\infty} \dot{E}_{Ohmic}\, dx' \simeq 0.9980\,\frac{P}{\alpha}$

and $\int_{-\infty}^{x} \dot{E}_{visc}\, dx' \simeq (\alpha x - 2.1450)\frac{P}{\alpha}$ as $x \to -\infty$. For $-\nu$, however, $\dot{E}_{Ohmic}$ and $\dot{E}_{visc}$ are oscillatory functions of opposite sign, and both diverge like $e^{-\alpha x/2}$. Clearly, the assumption of constant eddy viscosity must break down at some depth where the mean flow attempts to extract more energy from convective eddies than is actually available.

For $KR_m \neq 0$, the coupled equations (3.9) must be solved numerically; but as Figure 3.2 shows, the character of the velocity solutions is not changed: for $+\nu$ there is a single "K_0-like" monotonic solution with $Du(\infty) = 1$; for $-\nu$ there are both "J_0-like" ($u(\infty) = 1$) and "Y_0-like" ($Du(\infty) = 1$) oscillatory solutions. Figure 3.3 illustrates the behavior of the magnetic field for the torque-free "J_0-like" case (the other cases are qualitatively similar). We see that $|h_x|$ is inward-amplified to an asymptote which appears linear in ξ; in fact, as $\xi \to -\infty$, $D|h_x| \simeq \sqrt{KR_m}(0.1\xi - \ln\sqrt{2KR_m})$, so that $|h_x|$ is quadratic in ξ (the asymptotic form of the phase of h_x may also be obtained but we do not discuss it here). Dimensionally, this implies that $\frac{d^2|H_x|}{dx^2}$ is fixed (for given ν, ω, etc.) and the external field value H_0 depends exponentially on $\frac{d|H_x|}{dx}$ at given depth in the interior. This constitutes a boundary condition on the deep-seated region of dynamo action in striking contrast to the conventional $\frac{dH_x}{dx} = 0$ (to the approximation $D^2 \gg KD$) in the absence of a conductivity gradient. A similar result is to be expected for the fully three-dimensional problem.

The "toroidal" field h_y differs qualitatively from that for $KR_m = 0$ only in that (for the case of $-\nu$) $|h_y|$ oscillates between positive bounds as $\xi \to -\infty$ rather than between zero and a diverging upper bound. In particular, the period of oscillation still decreases so rapidly as $\xi \to -\infty$ that $\dot{E}_{Ohmic}$ and $\dot{E}_{visc}$ diverge.

Unlike h_x, h_z is inward-attenuated, reaching a limit of $\frac{1}{\sqrt{KRm}}$ as $\xi \to -\infty$ for large KR_m. This may, however, be a consequence of our assumption $\frac{\partial}{\partial z} = 0$ rather than a reflection of the actual behavior of the field in three dimensions.

Figure 3.4. Dimensionless viscous and Ohmic dissipation (solid and dotted curves, respectively) for two cases with $KR_m = 0$. For positive viscosity (velocity solution $u = 2K_0(\zeta)$ in Figure 3.2) both terms are positive and vanish as $\xi \to -\infty$. For negative viscosity ($u = J_0(\zeta)$) they are of opposite sign and divergent; "viscous" liberation of convective energy replenishes Ohmic losses. Similar behavior occurs for finite KR_m.

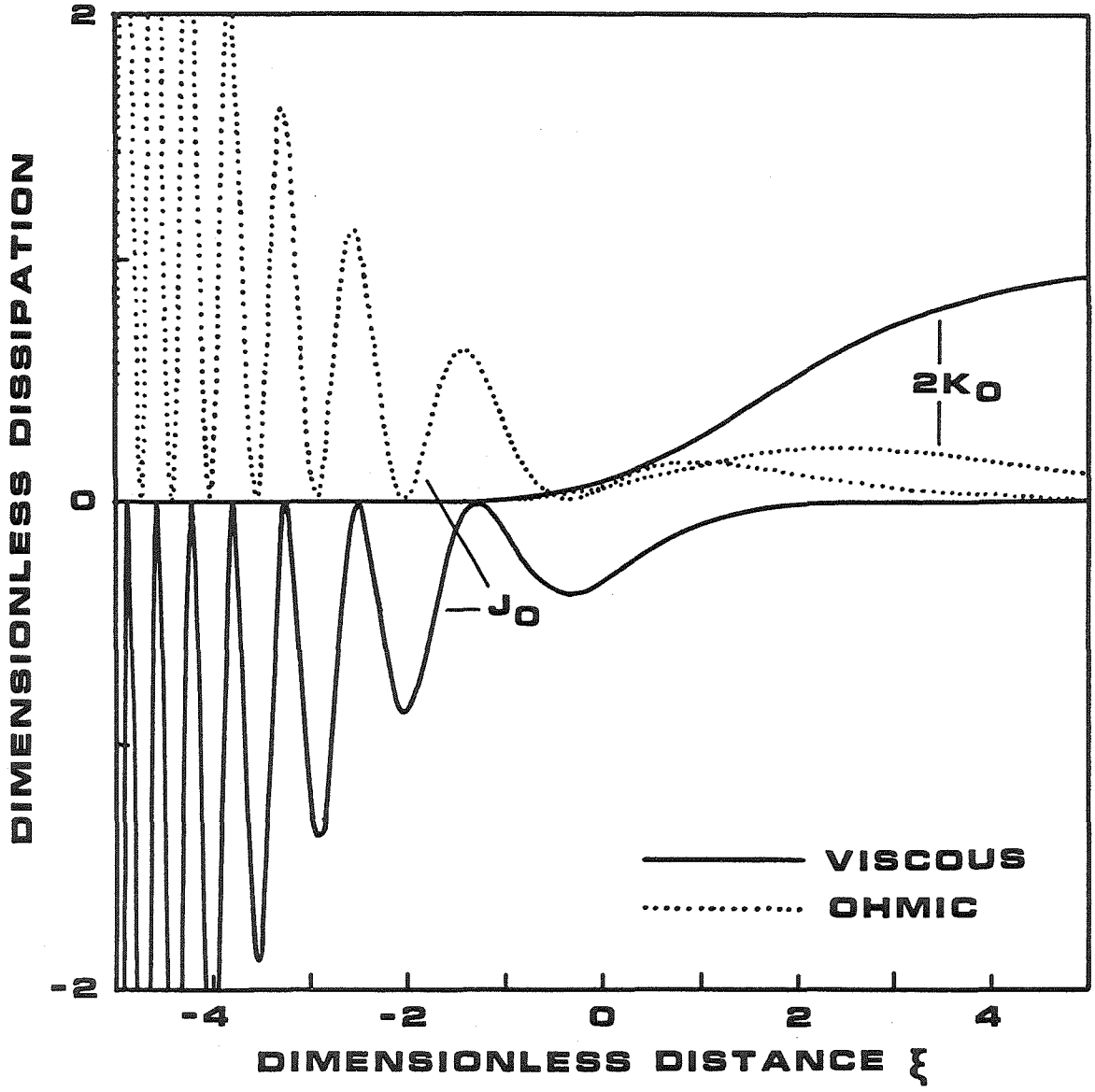

2
0
-2
DIMENSIONLESS DISSIPATION
2K0
-J0
VISCOUS
OHMIC
-4
-2
0
2
4
DIMENSIONLESS DISTANCE ξ

Table I. Parameters for Equatorial Jet Models

Quantity	Units	Jupiter		Saturn		Uranus	
		J_0-like	K_0-like	J_0-like	K_0-like	J_0-like	K_0-like
R_e	km	71398		60330		25440	
ϵ		0.0637		0.102		0.024	
$\theta_{v\,\mathrm{north}}$	°	15.0		37.2		...	
$\theta_{v\,\mathrm{south}}$	°	16.0		40.0		22	
ω	$10^{-4}\,\mathrm{s}^{-1}$	1.5		0.71		0.66	
θ_v	°	15.6	15.7	37.2	36.4	22	[a]
$\Delta\theta_v$ [b]	°	1.0		1.7		1.8	
S_0/R_e		0.9687		0.8282	0.8247	0.9354	0.9350
$H_{0\,m=0}$ [c]	G	0.057		0.0531	0.0510	0.111	
$H_{0\,m=1}$	G	0.568	0.569	0.0048	0.0046	0.175	
ρ [c]	$\mathrm{g\,cm}^{-3}$	0.0787	0.0794	0.141	0.140	0.481	
ν	$\mathrm{cm}^2\,\mathrm{s}^{-1}$	2500		2500		494	507
λ [d e]	$10^9\,\mathrm{cm}^2\,\mathrm{s}^{-1}$	28.8	29.2	3.03	2.60	7.41	7.25
α^{-1}	km	296	298	1400	1320	383	
K^{-1} [c]		233	231	35.5	38.1	61.8	
R_m [d]		4.62	4.66	457	474	13.1	13.4
KR_m [d]		0.0198	0.0201	12.9	12.4	0.213	0.216

[a] If not shown, value is the same as for the J_0-like model.
[b] Variation of θ_v corresponding to a change in λ of $\sqrt{10}$.
[c] Evaluated at $v = 0$ location where model (H_0, α, K constant) is fitted to planet.
[d] Extrapolated to $Q = 1$ with constant α.
[e] Values for Jupiter, Saturn are at the upper limit of uncertainty (cf. Figure 2.1).

4. Application to Jupiter and Saturn

To apply the dimensional form of the model described in the last section to the equatorial jet of the giant planets, we use the eddy viscosity ν as a parameter to fit the observed jet width, then show that the required viscosity is consistent with the upper limit derived by Ingersoll and Pollard (1982). Of the remaining quantities, obtained from experiment or theory and listed in Table I, a few deserve comment here.

Zonal velocity profiles based on Voyager imagery (Smith *et al.* 1979; 1982) were used to determine the jet width θ_v and shear amplitude ω. Both for the observations and the J_0-like models, $\theta_v$was defined as the lowest latitude at which $v = 0$ (or an

average of the values in the two hemispheres if they differ). For the K_0-like models, in which $v \rightarrow 0$ only asymptotically, θ_v was defined by constructing a tangent to the dimensionless velocity $u(\xi)$ at $u = 0.5$ and finding its zero crossing ξ_v, then mapping this location onto the surface of the planet in the usual way. With this definition $\theta_v(\nu)$ is very nearly the same for both types of model when ν is small. The distance from $Q = 1$ to the zero crossing (or extrapolated zero crossing) of v is a function of KR_m and can be substantial: up to several times α^{-1}. Thus, even to calculate a simple width parameter for the jet we must solve the differential equations (3.9) in addition to finding the location s_0 at which $Q = 1$. The shear amplitude was taken as $\omega = \left.\frac{dv}{ds}\right|_{\theta_v}$, a good approximation since the dimensionless shear $Du \simeq 1$ near $u = 0$.

The magnetic field strengths in Table I are rms radial fields in the equatorial plane at the locations where $v = 0$, computed from the multipole expansions of Smith *et al.* (1976, model P11 3I2E) and Connerney *et al.* (1984, models Z3, P11A as corrected in note in proof) according to the formulae $H_{0,m=0} = \frac{3}{2}\left(\frac{R_e}{s}\right)^4 g_2^0$, $H_{0,m=1} = \frac{1}{\sqrt{2}}\left(\frac{R_e}{s}\right)^3 \sqrt{(g_1^1)^2 + (h_1^1)^2)}$. The $m = 1$ value for Saturn is in fact a crude upper limit based on a dipole tilt of 1°; the nonaxisymmetry of Saturn's external magnetic field now appears to lie below the threshold of detectability. For both planets, the larger field component is uncertain to $\sim \pm 10\%$ and the smaller to $\sim \pm 100\%$. In applying our models, we can ignore the Jovian $m = 0$ field, which contributes $\lesssim 1\%$ of the total Lorentz force, whereas for Saturn both field components may be dynamically important, since the nonaxisymmetric term is inward-amplified.

The magnetic diffusivity λ was discussed at length above. The density ρ as a function of depth was estimated using a constant gravitational acceleration and a polytropic equation of state for cosmic-composition gas (Stevenson 1982) in the equation of hydrostatic equilibrium.

Figure 4.1 shows the dependence of jet width on eddy viscosity obtained with

Figure 4.1. Width of the equatorial jet versus eddy viscosity for J_0-like (upper of each pair of curves) and K_0-like models, with effect of ± 5 db variation of λ (cf. Fig. 2.1) indicated by shading. Vertical bars represent Voyager observations for two hemispheres, plotted at the upper limit $\nu = 2500\,\mathrm{cm^2\,s^{-1}}$ based on available convective energy. Circles represent the models of Table I and Figures 4.2–4.5.

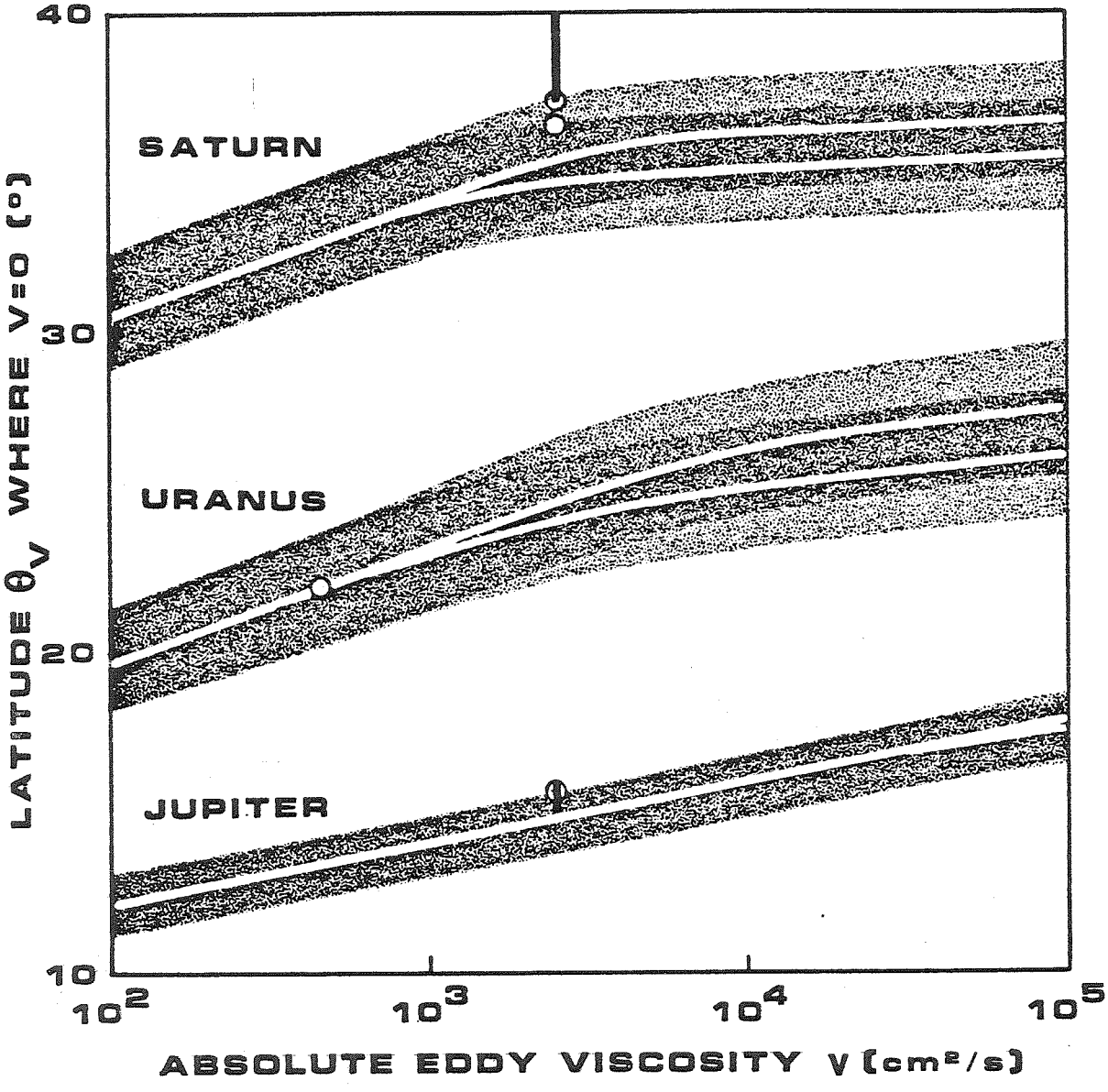

40
30
20
10
SATURN
URANUS
JUPITER
LATITUDE θ_V WHERE V=0 [°]
10^2
10^3
10^4
10^5
ABSOLUTE EDDY VISCOSITY V [cm²/s]

Figure 4.2. Equatorial jet models for Jupiter . Best-fit models with negative (solid curve) and positive (dashed curve) viscosity are plotted along with Voyager data. Attention should be restricted to the regions near the zeros of velocity; the model does not describe the behavior of "broken" Taylor columns at high latitudes or the complex nonmagnetic effects responsible for the fine structure of the jet.

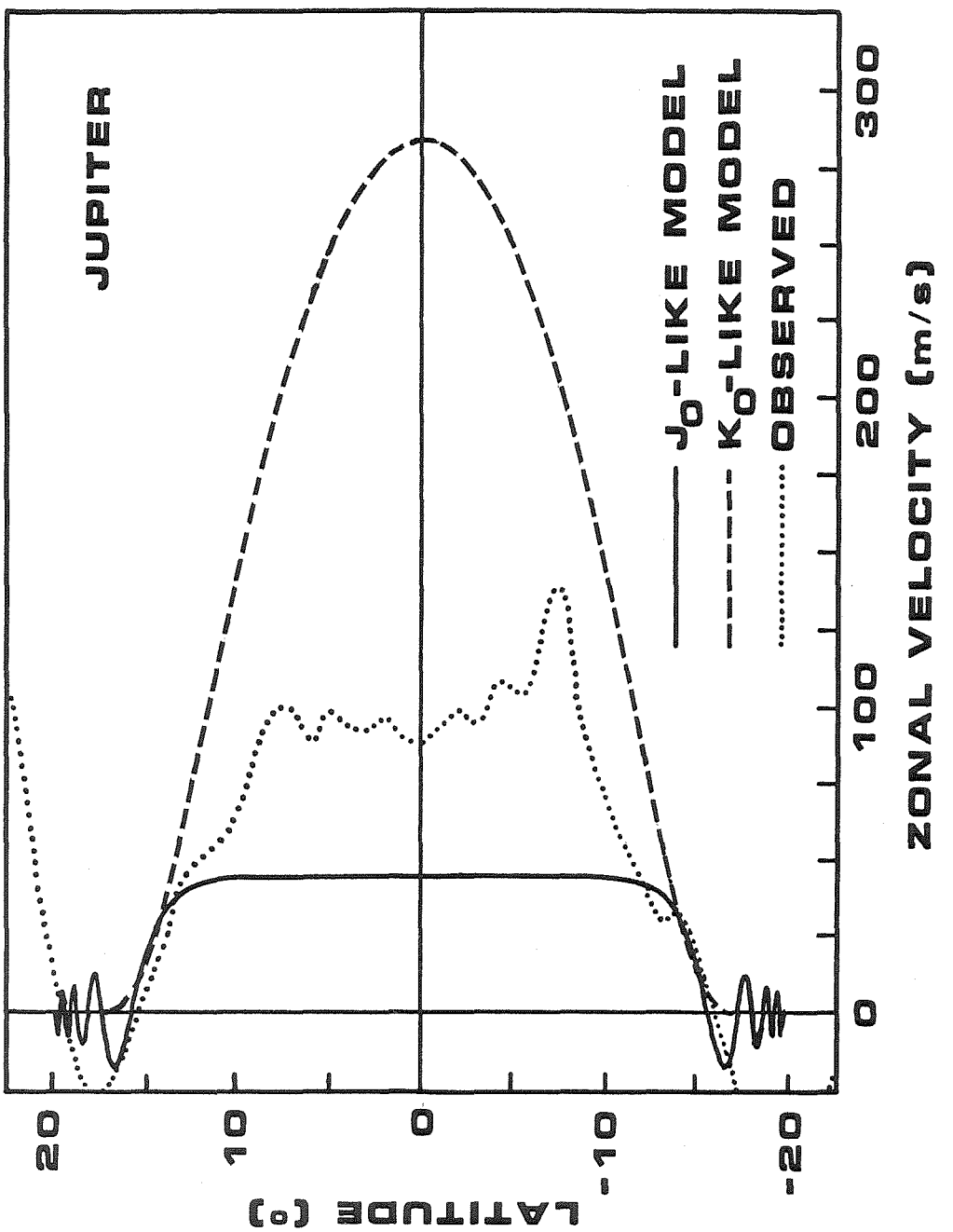

JUPITER
J_0-LIKE MODEL
K_0-LIKE MODEL
OBSERVED
ZONAL VELOCITY (m/s)
0
100
200
300
LATITUDE (°)
20
10
0
-10
-20

Figure 4.3. Equatorial jet models for Saturn . See previous figure caption.

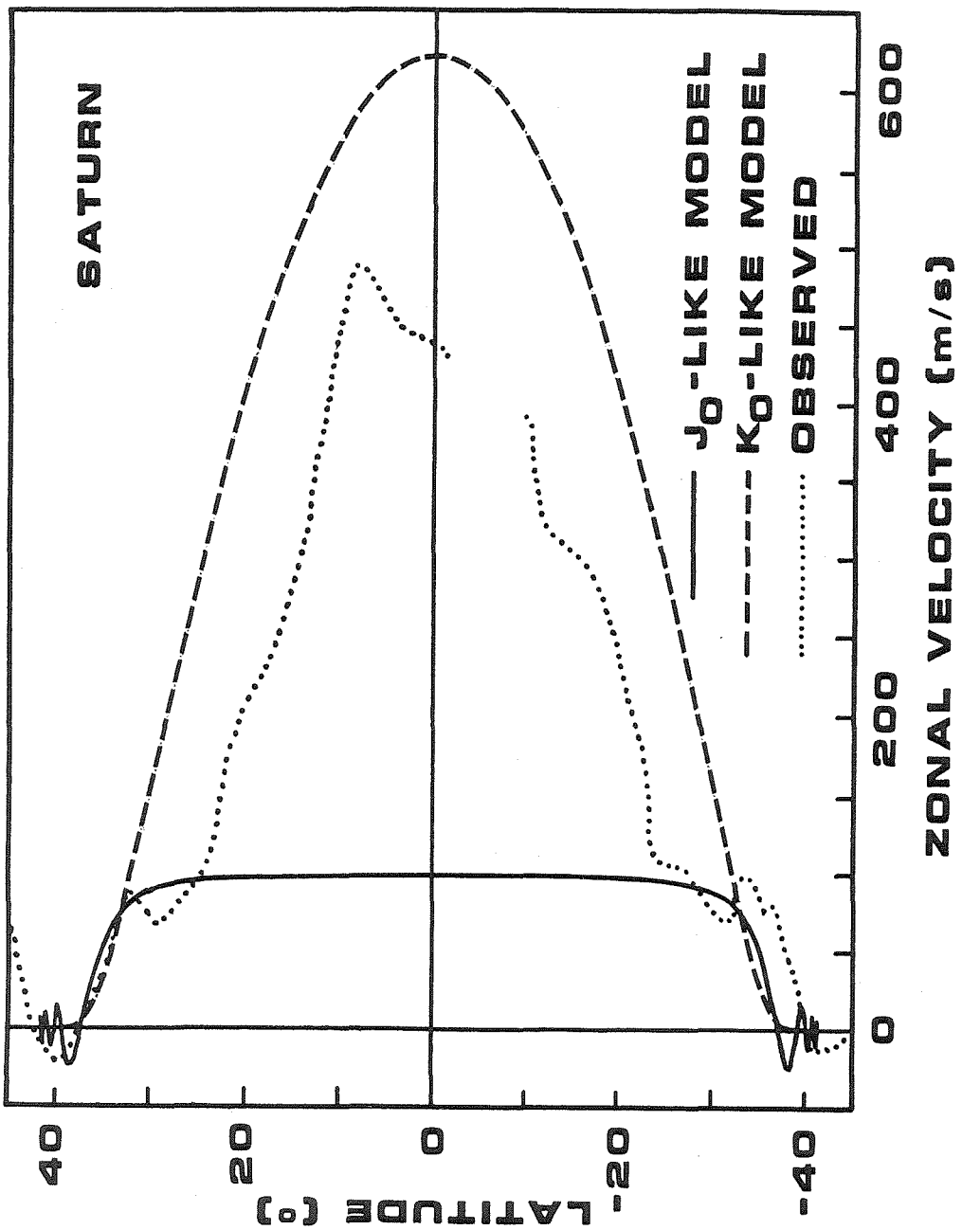

SATURN
J_0-LIKE MODEL
K_0-LIKE MODEL
OBSERVED
ZONAL VELOCITY (m/s)
0
200
400
600
40
20
0
-20
-40
LATITUDE (°)

Figure 4.4. Magnetic field model for Jupiter . Radial (H_x) and toroidal (H_y) fields in the equatorial plane corresponding to the J_0-like velocity solution in Figure 4.2 are shown. H_x is near-constant because of the small magnetic Reynolds number (geometric attenuation has not been included).

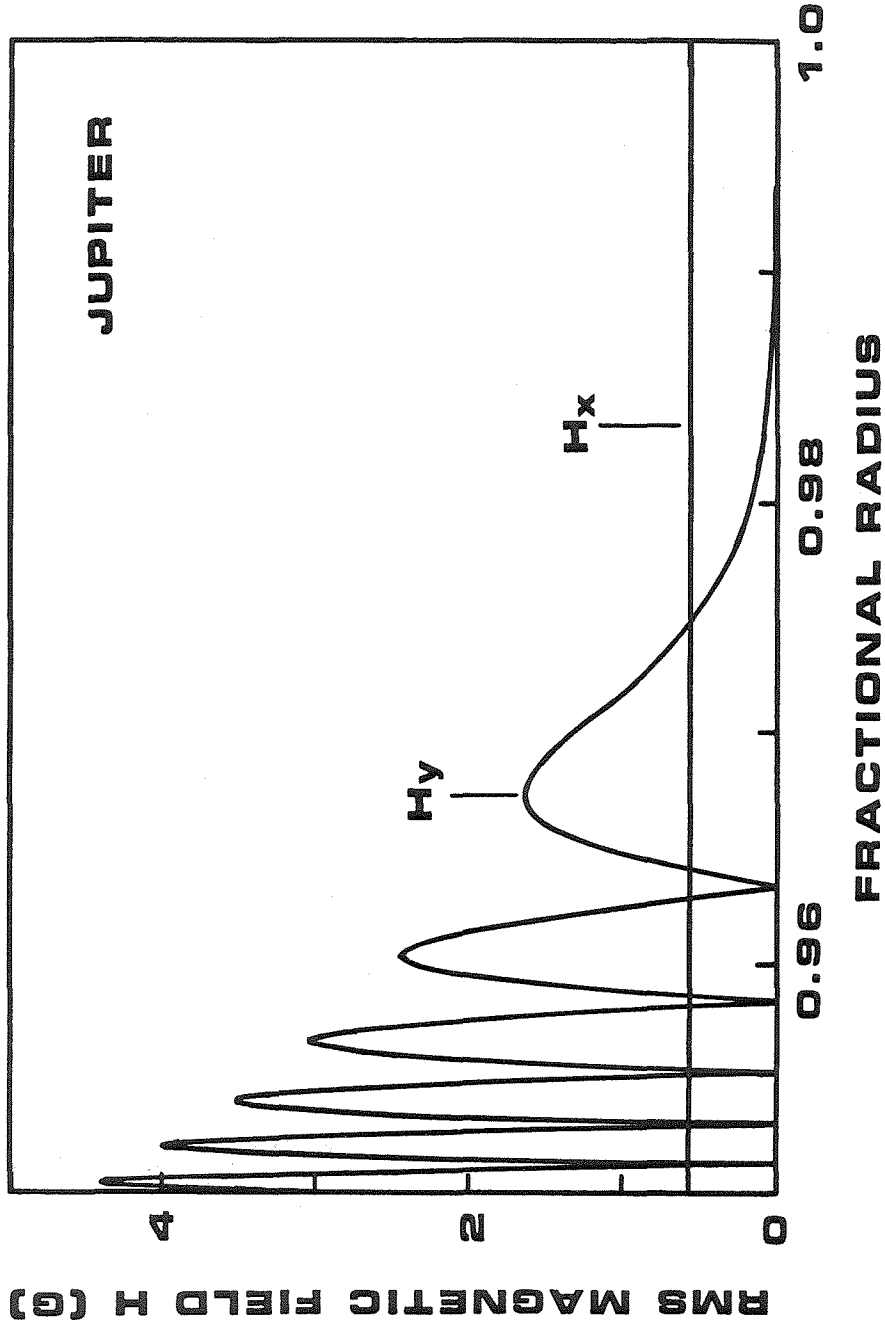
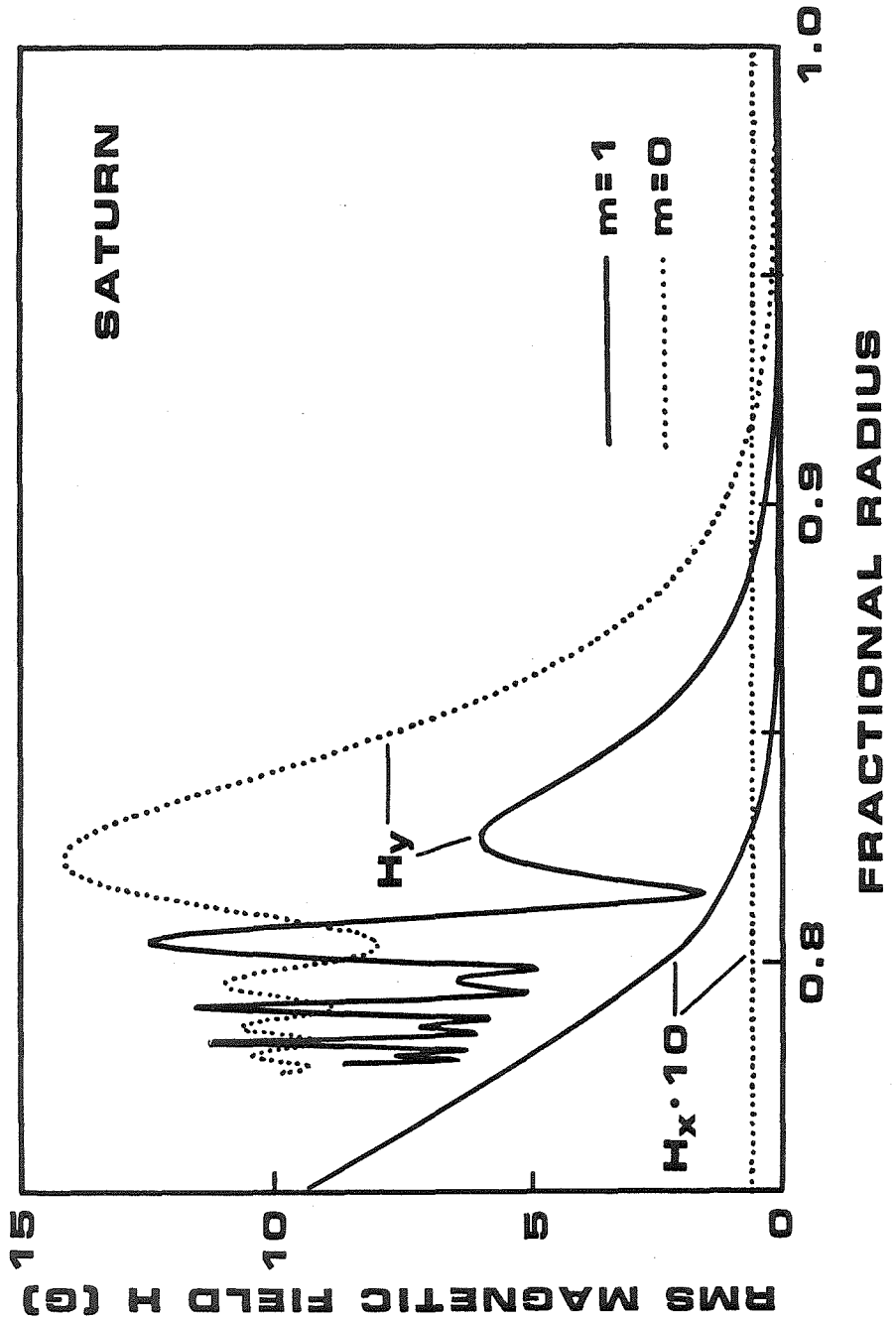

JUPITER
Hx
Hy
0.96
0.98
1.0
FRACTIONAL RADIUS
0
2
4
RMS MAGNETIC FIELD H (G)

Figure 4.5. Magnetic field model for Saturn . See previous figure caption. The non-axisymmetric radial field (solid curve) is small in the exterior but strongly inward-amplified; the axisymmetric (dotted curve) field is constant, neglecting geometric attenuation. Each radial component induces a toroidal component (note different scales) of the same azimuthal wavenumber m.


SATURN
m=1
m=0
Hy
Hx·10
RMS MAGNETIC FIELD H (G)
15
10
5
0
0.8
0.9
1.0
FRACTIONAL RADIUS

these assumptions, and in particular the best-fit models whose parameters appear in Table I. We see that with the nominal magnetic diffusivity, $\overline{\theta}_v = 15\overset{\circ}{.}5$ is obtained for Jupiter with ν only slightly in excess of the Ingersoll and Pollard (1982) limit based on the convective energy available for dissipation. For Saturn, $\theta_v = 38\overset{\circ}{.}6$ cannot quite be attained (essentially because of the decrease of α with depth), but for reasonable ν, θ_v is only a few degrees smaller. An increase of λ by $10^{1/2}$ to the upper limit of our estimated uncertainty (indicated by shading) allows us to fit the equatorial jets of both planets with $\nu \simeq 2500\,\mathrm{cm^2\,s^{-1}}$.

Figures 4.2 and 4.3 illustrate J_0-like and K_0-like models of the zonal velocity $v(\theta)$ computed with $\nu = 2500\,\mathrm{cm^2\,s^{-1}}$ and λ increased to the limit of uncertainty, superimposed on the Voyager data (Smith *et al.* 1979, 1982). Attention should be restricted to $|\theta| \simeq \theta_v$ since the models describe neither the possibility of "broken" Taylor columns at higher latitudes nor the complex interaction between convection and differential rotation that leads to the fine structure at lower latitudes. With these restrictions we see that the data do not discriminate between the two classes of model (*e.g.*, the oscillatory behavior of the J_0-like solution does not account for the existence of jets at $|\theta| > \theta_v$). The presence of a"shoulder" in the observed velocity profile (at $v \simeq 40\,\mathrm{m\,s^{-1}}$ for Jupiter and $v \simeq 80\,\mathrm{m\,s^{-1}}$ for Saturn) similar to that in the J_0-like model is, however, intriguing. The strongest conclusion that can be drawn is that the physical mechanism responsible for the shoulder operates on a lengthscale comparable to α^{-1} in each planet (α differs more than fourfold in the two cases), since when we adjust the model shear ω to the data, the shoulder velocity is reproduced as well.

For the sake of completeness, we show in Figures 4.4 and 4.5 the model radial and azimuthal magnetic field components in the equatorial plane. An oscillatory toroidal field of substantial strength is indicated for both planets. The Jovian mag-

netic Reynolds number is small ($KR_m \simeq 0.02$), so that the radial field is almost constant (geometric attenuation is neglected). In contrast, the much weaker field on Saturn results in $KR_m \simeq 12$. The axisymmetric radial field is unaffected, but the $m = 1$ field is substantially enhanced in the interior, at least in the equatorial plane. We will return in the concluding section of this paper to the intriguing question of whether a similar shielding effect operating at higher latitudes as well may be responsible for the high degree of axisymmetry of the externally measured field.

To restate the important points of this section, the model of Section 3 reproduces the widths of the equatorial jets of Jupiter and Saturn for choices of the eddy viscosity ν that are similar in magnitude and roughly in keeping with the constraint imposed by available convective energy (especially if the assumed magnetic diffusivity is increased within the limits of uncertainty). A large magnetic Reynolds number KR_m, and hence modification of the poloidal magnetic field, is indicated for Saturn but not for Jupiter.

5. Uranus and Neptune

Much less is known about these planets than about Jupiter and Saturn, but, as indicated in Section 2, the presence of abundant water in their envelopes may lead to substantial electrical conductivity. We show that Voyager 2 observations at Uranus are consistent with our equatorial jet model for a plausibly small choice of eddy viscosity, provided one assumes a two-layer internal structure, and make the qualitative prediction of a wide equatorial jet on Neptune.

No equatorial jet has been directly observed on Uranus. Nonetheless, based on the success of our model in reproducing the width of the Jovian and Saturnian jets, even though it does not describe the "broken" Taylor columns that must exist at $|\theta| > \theta_v$, we make the case that parameters extrapolated from slightly *higher* latitudes may be used to characterize a possible Uranian equatorial jet. The Voyager 2 images

(Smith *et al.* 1986) reveal seven cloud features at latitudes from $-25°$ to $-70°$, with rotation periods well described by the relation $\tau - \tau_0 = 0.081(\theta + 22°) + 0.000455(\theta + 22°)^2$ hours, where $\tau_0 = 17.24$ hours is the rotation period of the magnetic field. Three of these clouds lie between $-25°$ and $-27°$, where possible thermal wind corrections to their velocity are small (Hanel *et al.* 1986), lending confidence to the extrapolation to $\theta_v = 22°$, with $\omega = -6.6 \times 10^{-5}\,\mathrm{s}^{-1}$ (*i.e.*, a retrograde equatorial jet).

The Uranian magnetic field is unusual. We estimate significant H_0 for both $m = 0$ and 1 based on the best fit offset, tilted dipole (OTD) field model of Ness *et al.* (1986). The OTD parameters were first converted to planetocentric multipole moments according to the formulae of Smith *et al.* (1976, p. 799), then expressed as H_0 as in Section 4.

We assume a two-layer model of the interior of Uranus. The "gas" layer of three-layer models is too cold to be conductive (cf. Section 2), so that the minimum hydromagnetically determined jet width occurs at $Q \geq 1$ at the top of the "ice" layer; for published models (Hubbard and MacFarlane 1980) this leads to $\theta_v \simeq 45°$, inconsistent with the Voyager data. We therefore model only the two-layer case in detail, converting the pressure-temperature dependence of H_2O conductivity (Holtzapfel 1969) to depth dependence via the best-fit polytropic equation of state of Hubbard (1984) plus an approximate adiabatic temperature distribution, both in reasonable agreement with detailed models. As noted in Section 2, we adopt a ten-fold suppression of the H_2O conductivity by dissolved hydrogen-helium.

With these assumptions, we obtain the $\theta_v(\nu)$ curves shown in Figure 4.1. The inferred width of 22° is obtained with an eddy viscosity of $\sim 500\,\mathrm{cm}^2\,\mathrm{s}^{-1}$, substantially less than the value used for Jupiter and Saturn. This is in keeping with the limit on ν imposed by the availability of convective energy to be dissipated, since a rough upper limit on the internal heat flux of Uranus is an order of magnitude less than

that of Saturn (Hubbard 1984). The magnetic Reynolds number obtained, $KR_m \simeq 0.2$, is small, so that no axisymmetrization of the external field by the differentially rotating envelope is to be expected. This difference from Saturn results from the large inclination and offset of the Uranian dipole, which greatly increase $\sum_m H_{0,m}^2$, despite the comparable total magnetic moments of the two planets. Note that our estimates of ν and KR_m for Uranus are in some sense upper limits, since the Voyager observations only constrain the width of the equatorial jet to be $\lesssim 22°$.

Even less can be said about Neptune than about Uranus, but several factors point to a relatively wide Neptunian jet. First, Neptune is roughly similar to Uranus in temperature and in density (hence, presumably composition), leading us to expect a similar conductivity structure in the two planets. Second, the measurably greater internal heat flux of Neptune (Hubbard 1984) leads to a correspondingly higher limit on the eddy viscosity. Finally, though the total moments of the Uranian and Neptunian magnetic fields might naively be expected to be similar (Hill and Michel 1975), the relevant magnetic field measure H_0 could be much less if Neptune does not share Uranus' high magnetic inclination and offset, which may be a consequence of its large obliquity (Stevenson, in preparation). The latter two effects both work to locate the $Q = 1$ surface deeper in the planet and hence to broaden the predicted equatorial jet.

6. Discussion

The one-dimensional hydromagnetic model presented in the previous sections may be considered a success within the limits of its intended applicability: for plausible values of the diffusivities λ and ν it reproduces the observed widths of the equatorial jets of three very different planets. As such it lends credence both to the hypothesis of (fairly) deep zonal flows and to the asserted importance of hydromagnetic effects outside the cores of the giant planets. Should one wish, however, to look beyond this single-parameter description to the details of the zonal flow, a number of problems

arise that need clarification. Most are attributable (directly or indirectly) to the fact that the model is in essence an expansion about the equatorial plane on an assumed perfectly rigid Taylor column. Subsidiary problems arise from condensing all the physics of turbulent flow into the single parameter ν.

An example of the latter class of problems is the failure of our model to duplicate the structure of the zonal velocity profile at the lowest latitudes. Inasmuch as the flow there is purely hydrodynamic, rather than hydromagnetic, the problem is far outside the scope of this paper and we will say no more on the subject. The breakdown of the model at high latitudes, on the other hand, is both crucial and instructive. Starting with ideally coherent Taylor cylinders, we predict that at high latitudes the zonal wind will either vanish or oscillate on an ever-decreasing lengthscale, far shorter than that which is observed. (The latter case, for negative eddy viscosity, also leads to diverging dissipation. This flaw can be removed by making ν a decreasing function of the shear $\frac{dv}{ds}$, such that $\nu \cdot \left(\frac{dv}{ds}\right)^2$ is bounded, but the problem of too rapid oscillation remains.) We conclude from this that Taylor cylinders reaching the surface at high latitude must be "broken," either trivially in the shallow atmosphere or at depth. In addition, in order for our model to work as well as it does at predicting the jet width θ_v, this breaking must first occur at or slightly above that latitude. Our experience with the present model suggests that the flow is hydrodynamic and independent of axial coordinate z outside a roughly spherical surface on which $Q = 1$, and hydromagnetic, z-dependent, and possibly much slower inside that surface. The two-dimensional (if we assume $\mathbf{H} \propto e^{im\phi}$+c.c.) problem of matching the inner flow to the outer remains as a challenging unsolved problem. A boundary layer analysis (*i.e.*, neglecting horizontal derivatives) indicates that differential rotation on lengthscale $\gg \alpha^{-1} \simeq 10^3\,\mathrm{km}$ could penetrate to the metallic core. Unfortunately, the observed jets on Jupiter and Saturn have widths comparable to α^{-1}, making their

analysis more difficult.

Knowledge of the dynamics of interrupted Taylor columns is also needed to answer the questions: What is the appropriate measure H_0 of the nonaxisymmetric field to use in the equatorial jet model when KR_m is large? As a corollary, can the near spin-axisymmetry of Saturn's magnetic field be attributed to the zonal flow? For Jupiter and Uranus, this question does not arise. It is self-consistent to assume that the nonaxisymmetric field in the equatorial plane is unattenuated (hence given by the appropriate multipole component of the external field) since this leads to a flow model with $KR_m \lesssim 1$ at the $Q = 1$ level. The situation for Saturn is more puzzling, since the radial field in the equatorial plane is strongly outward-attenuated by the magnetic skin effect. Two limiting possibilities suggest themselves, with H_0 ranging from zero to the external multipole value. We describe these extreme cases, without being able to choose between them or their intermediates. (Fortunately, in Saturn $H_{0,m=0}$ is large enough that the predicted value of $\theta_v(\nu)$ is affected only slightly by this uncertainty.) First, if the flow at high latitudes does not share the axisymmetrizing property of the equatorial jet, the observed Y_1^1 (tilted dipole) component of the external field connects (via high latitudes) to the interior. In the equatorial plane H_s is then small in the interior and smaller still outside; we should assume $H_{0,m=1} \simeq 0$ in our model.

It is possible, on the other hand, that the magnetic field in the deep interior of Saturn is less axisymmetric than the external field would suggest. In the absence of an understanding of the two-dimensional hydromagnetic problem, the following conceptual model is instructive. Project the Voyager zonal wind profile along spin-axis concentric cylinders to the depth where $Q = 1$ and calculate KR_m as a function of latitude on this surface. Now idealize the attenuation process as follows: for some (scalar) quantity ϕ with Y_1^1 angular dependence inside $Q = 1$, let its value outside $Q = 1$ be undiminished where KR_m exceeds some threshold KR_m^*, but be attenuated to

zero locally if $KR_m > KR_m^*$. Expanding the exterior function in spherical harmonics then yields an overall attenuation factor for the Y_1^1 component of ϕ from interior to exterior. Numerically, one obtains tenfold attenuation for $KR_m^* \simeq 1.6$, a reasonable threshold based on our experience in the equatorial plane. This would seem to say that deep inside Saturn (but outside the dynamo region) the tilt of the dipole field could be $O(10°)$, comparable to that of Jupiter (and Earth). Though satisfying to the extent that the intrinsic axisymmetry of the Saturnian dynamo need not be exceptional, this result is puzzling. Why would Jupiter and Saturn, with apparently similar fields below $Q = 1$, experience such different amounts of axisymmetrization? A partial answer is that (based on our model) the shielding effect of the zonal flow leads to large field *gradients* as well as large fields in the interior. If the preceding arguments are correct, then the Saturnian dynamo is unexceptional in its dipole tilt — but exceptional in the richness of its higher multipole spectrum (corresponding to strong radial gradients of $\mathbf{H}$). One wonders what has been gained.

It should be clear by now that the present work calls out for an investigation of the hydromagnetic flow of a planetary envelope in two dimensions. The encouraging results of our simple model, meanwhile, will be subject to further testing and refinement by the Voyager 2 encounter with Neptune in 1989 (which will, we hope, increase our collection of equatorial jets by 33%) and by direct observation of band-gap closure in hydrogen at high pressure.

Acknowledgements

This research was supported by NASA grant NAGW–185.

References

ABRAMOWITZ, M., AND I. A. STEGUN (1965). *Handbook of Mathematical Functions*, Dover Publications, NY, p. 362.

ALLISON, M., AND P. H. STONE (1983). Saturn Meteorology: A Diagnostic Assessment of Thin Layer Configurations for the Zonal Flow. *Icarus.* **54**, 296–308.

ASHCROFT, N. W., AND N. D. MERMIN (1976). *Solid State Physics*, Holt, Rinehart, NY, Ch. 28.

BUSSE, F. H. (1976). A Simple Model of Convection in the Jovian Atmosphere. *Icarus.* **29**, 255–260.

BUSSE, F. H. (1983). A Model of Mean Zonal Flows in the Major Planets. *Geophys. Astrophys. Fluid Dyn.* **23**, 153–74.

CHANDRASEKHAR, S. (1965). *Hydrodynamic and Hydromagnetic Stability*, Dover Publications, NY, p. 7.

CHAPMAN, S., AND T. G. COWLING (1952). *The Mathematical Theory of Non-uniform Gases*, Cambridge Univ. Press, Cambridge, p. 321.

CONNERNEY, E. P., ET AL. (1984). Magnetic Field Models. In *Saturn* (T. Gehrels and M. Matthews, Eds.), Univ. of Arizona Press, Tucson, pp. 354–377.

CONRATH, B. J., AND P. J. GIERASCH (1984). Global Variation of the para Hydrogen Fraction in Jupiter's Atmosphere and Implications for the Dynamics of the Outer Planets. *Icarus.* **57**, 184–204.

DROBYSHEVSKIĬ, E. M. (1979*a*). Differential Rotation of the Atmospheres of Jupiter and Saturn. *Sov. Astron.* **23**, 334–340.

DROBYSHEVSKIĬ, E. M. (1979*b*). On the Equatorial Flows on Uranus and Neptune. *Sov. Astron.* **23**, 598–604.

FRIEDLI, C., AND N. W. ASHCROFT (1977). Combined Representation Method for use in Band Structure Calculations: Application to Highly Comperssed Hydrogen. *Phys. Rev.* **16B**, 662–672.

GOETTEL, K. A., W. C. MOSS, R. REICHLIN, AND S. MARTIN (1986). Progress in Diamond Cell Experiments: 460 GPa on the Ruby Fluorescence Pressure Scale. *EOS Trans. Amer. Geophys. Un.* **67**, 565.

HANEL, R., ET AL. (1986). Infrared Observations of the Uranian System. *Science.* **223**, 70.

HIDE, R. (1965). On the Dynamics of Jupiter's Interior and the Origin of his Magnetic Field. In *Magnetism and the Cosmos* (W. R. Hindmarsh, F. J. Lowes, P. H. Roberts, and S. K. Runcorn, Eds.), Oliver and Boyd, Edinburgh, pp. 378–395.

HIDE, R. AND S. R. C. MALIN (1979). The Size of Jupiter's Electrically Conducting Fluid Core. *Nature.* **280**, 42–43.

HILL, T. W., AND F. C. MICHEL (1975). Planetary Magnetospheres. *Rev. Geophys. Space Phys.* **13**, 967–974.

HOLTZAPFEL, W. B. (1969). Effect of Pressure and Temperature on the Conductivity and Ionic Dissociation of Water up to 100 kbar and 1000°C. *J. Chem. Phys.* **50**, 4424–4428.

HUBBARD, W. B., AND G. P. HOREDT (1983). Computation of Jupiter Interior Models from Gravitational Inversion Theory. *Icarus.* **54**, 456–465.

HUBBARD, W. B., AND J. J. MACFARLANE (1980). Structure and Evolution of Uranus and Neptune. *J. Geophys. Res.* **85**, 225–234.

HUBBARD, W. B., AND D. J. STEVENSON (1984). Interior Structure of Saturn. In *Saturn* (T. Gehrels and M. S. Matthews, Eds.), Univ. of Arizona Press, Tucson, pp. 47–87.

HUBBARD, W. B. (1984) Interior Structure of Uranus. In *Uranus and Neptune* (J. T. Bergstralh, Ed.), Washington: NASA CP 2330, pp. 291–325.

INGERSOLL, A. P., AND R. L. MILLER (1986). Motion in the Interior and Atmo-

sphere of Juptiter and Saturn: 2. Barotropic Instability and Normal Modes of an Adiabatic Planet. *Icarus.* **65**, 370–382.

INGERSOLL, A. P., AND D. POLLARD (1982). Motion in the Interiors and Atmospheres of Jupiter and Saturn: Scale Analysis, Anelastic Equations, Barotropic Stability Criterion. *Icarus.* **52**, 62.

MAO, H. K., P. M. BELL, AND R. J. HEMLEY (1985). Ultrahigh Pressure: Optical Observation and Raman Measurements of Hydrogen and Deuterium to 1.47 Mbar. *Phys. Rev. Lett.* **55**, 99–102.

MIN, B. I., H. J. F. JANSEN, AND A. J. FREEMAN (1986). Pressure-Induced Electronic and Structural Phase Transitions in Solid Hydrogen. *Phys. Rev.* **33B**, 6383–6390.

MOFFATT, H. K. (1978). *Magnetic Field Generation in Electrically Conducting Fluids*, Cambridge Univ. Press, Cambridge, p. 298 ff.

MOTT, N. (1971). Conduction in Non-Crystalline Systems. VI Liquid Semiconductors. *Phil. Mag.* **24**, 1–18.

NESS, N. F., ET AL. (1986). Magnetic Fields at Uranus. *Science.* **233**, 85–89.

ROSS, M. (1985). Matter Under Extreme Conditions of Temperature and Pressure. *Rep. Prog. Phys.* **48**, 1–52.

ROSS, M., H. C. GRABOSKE, JR., AND W. J. NELLIS (1981). Equation of State Experiments and Theory Relevant to Planetary Modeling. *Phil. Trans. Roy. Soc. Lond.* **A303**, 303–313.

SMITH, B. A., ET AL. (1979). The Galilean Satellites and Jupiter: Voyager 2 Imaging Science Results. *Science.* **206**, 927–950.

SMITH, B. A., ET AL. (1982). A New Look at the Saturn System: The Voyager 2 Images. *Science.* **215**, 504–537.

SMITH, B. A., ET AL. (1986). Voyager 2 in the Uranian System: Imaging Science

Results. *Science.* **233**, 43–64.

SMITH, E. J., ET AL. (1976). Jupiter's Magnetic Field and Magnetosphere. In *Jupiter* (T. Gehrels, Ed.), Univ. of Arizona Press, Tucson pp. 788–829.

SMOLUCHOWSKI, R. (1972). Electrical Conductivity of Condensed Molecular Hydrogen in the Planets. *Phys. Earth Planet. Inter.* **6**, 48–50.

SMOLUCHOWSKI, R. (1975). Jupiter's Molecular Hydrogen Layer and the Magnetic Field. *Astrophys. J. Lett.* **200**, L119–L121.

STEVENSON, D. J. (1982). Interiors of the Giant Planets. *Ann. Rev. Earth Planet. Sci.* **10**, 257–295.

STEVENSON, D. J., AND E. E. SALPETER (1976). Interior Models of Jupiter. In *Jupiter* (T. Gehrels, Ed.) Univ. Arizona Press, Tucson, pp. 85–112.

TAYLOR, J. B. (1963). The Magneto-Hydrodynamics of a Rotating Fluid and the Earth's Dynamo Problem. *Proc. Roy. Soc.* **A274**, 274–293.

XU, J. A., H. K. MAO, AND P. M. BELL (1986) Ruby and Diamond Fluorescence: Observations at 0.21 to 0.55 Terapascal. *Science.* **232**, 1404–1406.

PAPER III

A Fast Finite-Element Algorithm for Two-Dimensional Photoclinometry

Measure your mind's height by the shade it casts.

— Robert Browning, *Paracelsus*

A Fast Finite-Element Algorithm for Two-Dimensional Photoclinometry

R. L. KIRK

Division of Geological and Planetary Sciences

California Institute of Technology

Pasadena, California 91125

Abstract

It is shown that the problem of two-dimensional photoclinometry (PC) — the reconstruction of a surface $z(x, y)$ from a brightness image $B(x, y)$ — may be formulated in a natural way in terms of finite elements. The resulting system of equations is underdetermined as a consequence of the lack of boundary conditions for z, but a unique solution may be chosen by minimizing a function S expressing the "roughness" of the surface. An efficient PC algorithm based on this formulation is presented, requiring ~ 10.66 (four-byte) memory locations and $\sim 10^4$ floating multiplications/additions per pixel, and incorporating: 1) Minimization of the roughness by the penalty method, which yields the smallest set of equations. 2) Iterative solution of the nonlinear equations by Newton's method. 3) Solution of the linearized equations by an inner iterative cycle of successive over-relaxation, which takes advantage of the extreme sparseness of the system. 4) Multigridding, in which the solutions to the smaller problems obtained by reducing the resolution are used recursively to greatly speed convergence at the higher resolutions, and 5) A rapid noniterative initial estimate of z obtained by exploiting the special symmetry of the equations obtained in the first linearization.

The algorithm is extensively demonstrated on 200 by 200 pixel synthetic "images" generated from digital topographic data for northern Utah over a range of phase angles. Rms error in the solution is ~ 22 m, out of ~ 660 m total relief. The error is dominated by "stripes" with the same azimuth as the light source, resulting from use of the roughness criterion in lieu of boundary conditions; the rms error along profiles parallel to the stripes is only $\sim$ 2–8 m, depending on the phase angle. Satisfactory solutions are obtained even in the presence of quantization error, noise, and moderate blur in the image.

Applications of the PC algorithm to both remote sensing and photomacrography are sketched; a photoclinometric map of a low-relief Precambrian era fossil is

presented as an example of the latter. Prospects for dealing with photometrically inhomogeneous surfaces, and an extension of the method to the analysis of side-looking radar data ("radarclinometry") are also discussed.

1. Introduction

Photoclinometry (PC) is "shape from shading" in the broadest sense: the recovery of geometric information about a surface from photometric data. It is of potential utility whenever photogrammetric methods (*i.e.*, stereo) cannot be used to obtain depth information. Such cases include remote sensing by flyby spacecraft, which image their target only once, by orbiting spacecraft whose orbital geometry constrains them always to view a particular region from the same angle, and even by fixed Earth-based telescopes in the case of the Moon. Surfaces with bland, gentle slopes pose problems for stereometry because of the difficulty of identifying corresponding points in the two images. Finally, stereo methods are difficult or impossible to use on very small regions, because of the restriction imposed on oblique viewing by depth of field. Of these (potential) applications, that to planetary remote sensing has the longest history. PC was first used to estimate the slopes of mare ridges on the Moon over 35 years ago (van Diggelen 1951) and has received sporadic but recurring interest ever since (Bonner 1960; Watson 1968; Bonner and Schmall 1973; Davis and McEwen 1984; Wilson *et al.* 1985), with applications to the Moon (Dale 1962; Wilhelms 1963; McCauley 1965; Rindfleisch 1965; 1966; Lucchitta and Gambell 1969; Rowan *et al.* 1971; Tyler *et al.* 1971), Mercury (Hapke *et al.* 1975; Mouginis-Mark and Wilson 1981), Mars (Davis *et al.* 1982; Davis and Soderblom 1982; 1984; Howard *et al.* 1982; McEwen 1985), icy satellites (Squyres 1981; Passey and Shoemaker 1982), and even Io (Moore *et al.* 1985). Surprisingly, only one author in the remote sensing community (Wildey 1975) has attempted the solution of the full two-dimensional problem of reconstructing a surface $z(x, y)$ from a single image. His algorithm is

general in its applicability and exhibits some of the ideas presented here, but is mathematically somewhat cumbersome and not efficient enough to be practical (the image on which it was demonstrated comprised 25 by 35 pixels). Influenced by the peculiar photometric properties of the lunar surface (and perhaps at least initially by the lack of computational power), other workers have concentrated on the estimation of one component of surface slope at a point and integration of this slope component along a line to yield a profile of the surface.

The two-dimensional approach to photoclinometry presented here, in contrast, originated in the context of small-scale topography. It was initially envisioned as a tool for the enhancement of photomacrographs of fossils from the Ediacaran period, roughly 670–550 Mybp. The Ediacaran fauna are a unique and enigmatic group of soft-bodied organisms, commonly preserved as very low-relief fossils on the bedding surfaces of fine sand- and siltstones (Glaessner 1961). In part because of this, their relationship to later metazoan life is controversial (Lewin 1984). It would therefore be of great interest to be able to apply image-enhancement techniques to the fossils (Kirschvink *et al.* 1982). The most obvious such techniques are matched filtration (to suppress the high spatial-frequency "grain" of the rock relative to the features of the fossil) and "stacking" of multiple fossils (to average the grain away). Reflection indicates that these techniques should properly be applied to the *topography*, rather than to the image[1]. In the case of matched filtration, the topographic signature of the grain has the desirable property of stationarity (statisitical uniformity over the field) whereas its signature in the shaded image it does not. The grains on large-scale slopes away from the light have a greater brightness contrast than those on slopes toward the light. In the case of stacking, adding *images* would be unlikely

[1] Note, however, that the optimal finished product is likely to be a pseudoimage computed from the "enhanced" topography. Greyscale representations of the altitude map turn out to be difficult to interpret visually.

to produce a meaningful result, unless one could ensure that the surface properties and illumination were the same for each. For stacking topography we require only (!) that the morphology of the specimens be close to identical and that they be properly aligned. Algorithms were therefore developed first for "photometric stereo" (analysis using the redundant information in two images with different illumination geometries) and ultimately for true two-dimensional PC from a single image. An example of an Ediacaran fossil image will be presented in Section 3.

It is instructive to consider the problems inherent in the zero- or one-dimensional approach to PC before passing on to the two-dimensional problem, which will be seen in the general case to be conceptually simpler (though of course more taxing computationally in proportion to the greater quantity of data to be dealt with).

Under given illumination and viewing geometry, the photometric function of a surface expresses the dependence of reflected intensity ("brightness") on the orientation of the surface. This orientation must be specified by two quantities, whether they are taken as two components of the unit normal vector to the surface, gradients in two specified directions, strike and dip, or even more arcane combinations such as the coordinates (in some map projection) of the point on a unit sphere that would have the same orientation. Measurement of the brightness at a point provides only a single constraint on these quantities, and in general neither can be determined individually; the zero-dimensional PC problem is severely underdetermined. For the lunar surface, however, it has been shown (Hapke 1963; 1966) that the brightness is independent of the gradient perpendicular to the plane containing the light source, surface point, and observer (the phase plane). The observed brightness may thus be inverted to yield the slope in the phase plane, and these slopes integrated to give a profile of the surface-phase plane intersection (Watson 1968; Bonner and Schmall 1973; Mouginis-Mark and Wilson, 1981). Because the transverse slope is unknown,

these profiles cannot in general be assembled into a full representation of the surface.

Bodies other than the Moon (except perhaps for Mercury) do not share its convenient photometric property[2], and the success of one-dimensional PC on them depends on additional information. Profiles may be made only if the strike is known *a priori* by symmetry, as on the diameter of a crater (Passey and Shoemaker 1982; Davis and Soderblom 1984), or if it may be estimated by inspection (Howard *et al.* 1982). These methods of course require close supervision by a human operator whose built in "photoclinometry software" enables him or her to look at the image and determine the strike, the computer then taking responsibility for the dip. Moreover, they are necessarily tied to the content of the specific image to which they are being applied.

Wildey (1975) was the first and so far the only author to realize the advantages to be gained by taking a two-dimensional approach to the PC problem from the start. In this approach, the computer is provided at the start with the single most important fact that enables the human to interpret the image: the knowledge that *this is a picture of a continuous surface.* The "shape" we seek to extract from shading is now a single quantity, the surface altitude $z(x,y)$, rather than two independent gradients, and the problem of calculating it from the observed brightness $B(x,y)$ is properly determined, at least in the case of an infinitely large image. (Alternatively, we can view the problem as that of determining the gradients $f = \frac{\partial z}{\partial x}$ and $g = \frac{\partial z}{\partial y}$ as before, but with the additional requirement that they form a total differential: $\frac{\partial f}{\partial y} = \frac{\partial g}{\partial x}$.) Since the information is specified in terms of the surface derivatives, however, we should not be surprised that information about the boundary conditions is needed

[2] Note, however, that the phase plane must be a plane of symmetry for the brightness, slopes to either side having the same effect. Thus for given in-plane slope, b is an extremum for zero slope normal to the phase plane. It follows that if the typical slopes in the transverse direction are small, the brightness will be independent of this slope component to first order, for any photometric function.

to complete the problem properly. Provided one uses an imaging system with a sufficiently wide total field of view, the needed boundary information is available at the edges of objects and sometimes at cusps on the surface, and indeed there exists a modest amount of literature in the discipline of computer vision (*e.g.*, Ikeuchi and Horn 1981; Brown 1984) on the application of this information (the description of surface orientation in terms of position on the unit sphere, mentioned above, is useful in this context because unlike the gradients, these coordinates are nonsingular at the limb). Unfortunately, in remote sensing (and also in photomacrography) one is frequently interested in images that do not contain the limb, and hence for which boundary information is not available. In this situation one must substitute an *ad hoc* constraint which one hopes is "harmless," bearing in mind that the problem is at any rate *almost* fully determined. As we shall see, this hope is at least partially fulfilled. In the following sections I describe the PC algorithm, demonstrate its properties on a case where the actual topography is known, and finally discuss problems that could impede its application (with sketches of some possible solutions), and generalizations of the method.

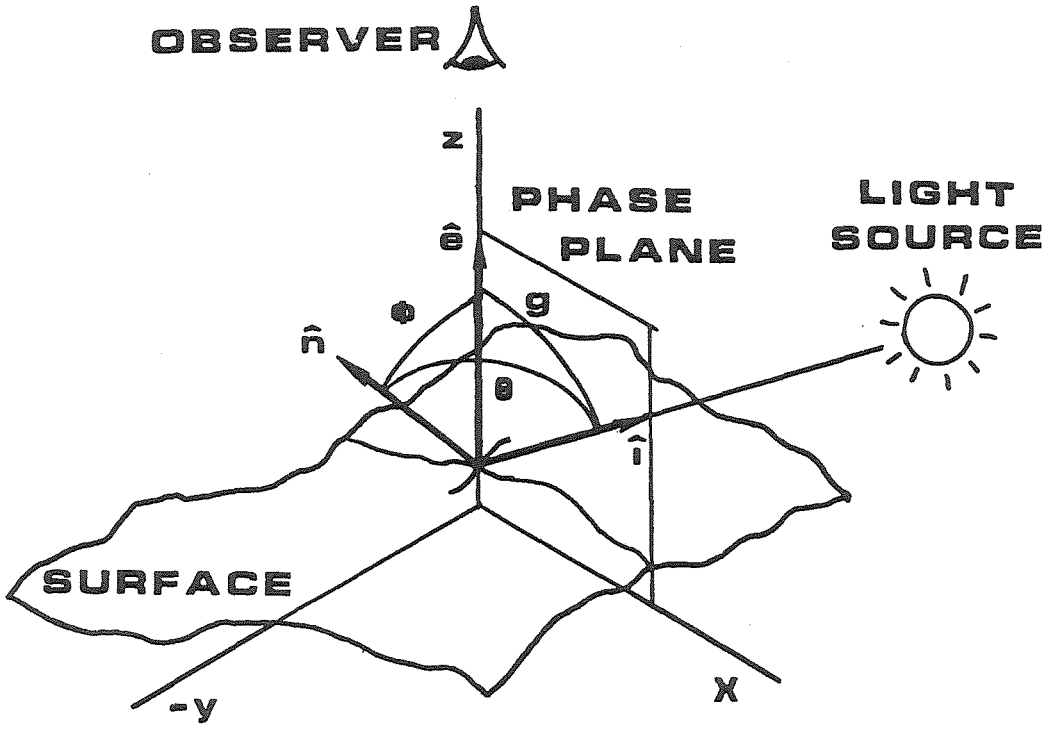
2. The Photoclinometry Algorithm

In practice, of course, one obtains an finite array of discrete brightness measurements (pixel values), each averaged over the instantaneous field of view or "footprint" of the imaging system. If the pattern of sensitivity of the system is described by the kernel $K(x,y)$, we have for an M by N image MN integral equations of the form:

$$B^e = \iint K(x-\overline{x}^e, y-\overline{y}^e)\, b(\frac{\partial z(x,y)}{\partial x}, \frac{\partial z(x,y)}{\partial y})\, dx\, dy, \qquad e = 1,2,\ldots,MN, \quad (2.1)$$

These are the fundamental equations of two-dimensional photoclinometry. Here and below it is convenient to choose coordinates with the observer looking down along the z axis, rotated so the source of illumination lies in the (x,z) plane, and with the

Figure 2.1. Coordinate system for photoclinometry . Cartesian coordinates are chosen with the observer looking down the z axis and the light source in the (x, z) plane. Unit vector $\hat{\imath}$ points toward the light source, $\hat{e}$ toward the observer, and $\hat{n}$, normal to the surface at the point (x, y, z). Then the photometric function b expresses the brightness in terms of the *incidence angle* $\theta = \hat{n} \cdot \hat{\imath}$, the *emission angle* $\phi = \hat{n} \cdot \hat{e}$, and the *phase angle* $g = \hat{\imath} \cdot \hat{e}$. The task of two-dimensional photoclinometry is to find the surface $z(x, y)$ whose normal vector is consistent with the observed brightness at every point.


OBSERVER
z
PHASE
PLANE
LIGHT
SOURCE
ê
g
n̂
θ
î
SURFACE
-y
X

interpixel spacing chosen as the unit of distance. In these equations $b(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y})$ is the photometric function written in terms of the surface gradients at the point (x, y), and B^e is the brightness measurement of pixel e, which is centered on $(\overline{x}^e, \overline{y}^e)$. The PC algorithm places no restrictions on the functional form of $b(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y})$[3].

To obtain a solution to the equations (2.1) we must discretize the altitude function $z(x, y)$, in such a way that we can calculate the gradients in terms of the discrete z values. At the very *least* (if the footprint K is very narrow, say) we need the derivatives at the centroid $(\overline{x}^e, \overline{y}^e)$ of each pixel. Whether we formulate the problem in terms of finite elements or finite differences, we will need an array of $M+1$ by $N+1$ to calculate derivatives (differences) at M by N points. The system of equations will be underdetermined, though by a relative amount that becomes small as M and N become large. The origin of the indeterminacy (the need for boundary conditions) is manifest in the way the altitude array "sticks out" beyond the edges of the image.

Lacking boundary conditions, we must further constrain the problem in some *ad hoc* way. In accordance with Occam's razor, the most natural constraint one can envision is a requirement that the surface be no more "rough" than is required by the data:

$$\delta S = 0, \tag{2.2}$$

for some roughness function S, subject to (2.1). (The δ notation indicates that S is to be made stationary, *i.e.*, its derivatives with respect to the adjustable parameters are to be set to zero.) This is most emphatically *not* to suggest that natural surfaces obey some kind of minimum-roughness criterion; in fact, they often have fractal properties (Mandelbrot 1982). We merely seek not to add any roughness of nonphysical origin

[3] For simplicity, in this paper I treat the function b as independent of position. Generalization to a photometric function that varies in a *known* way, *e.g.*, because of variation of the phase angle over the field of view, is straightforward. I discuss the case of a surface with nonuniform reflectivity below and argue that in some cases it may be possible to create an "albedo-corrected" image using multispectral data. Photoclinometry could then be performed on this image under the assumption of spatially invariant b.

to the topography in our solution.

The precise definition of "roughness" turns out to be unimportant. Three possibilities are:

$$S = \begin{cases} \iint (z - Z)^2 \, dx \, dy, & \text{the rms altitude,} \quad (2.3a) \\ \iint \sqrt{1 + \left(\dfrac{\partial(z-Z)}{\partial x}\right)^2 + \left(\dfrac{\partial(z-Z)}{\partial y}\right)^2} \, dx \, dy, & \text{the area, or} \quad (2.3b) \\ \iint -(c(z-Z) + e^{-1}) \ln(c(z-Z) + e^{-1}) \, dx \, dy, & \text{the entropy.} \quad (2.3c) \end{cases}$$

Here $Z(x, y)$ is a reference surface such as the mean plane[4]. The first criterion has the advantage of being simplest; the second is only slightly more difficult to evaluate, is perhaps more elegant, and was used by Wildey (1975). In Appendix A I give the explicit form of the roughness criteria in terms of finite elements.

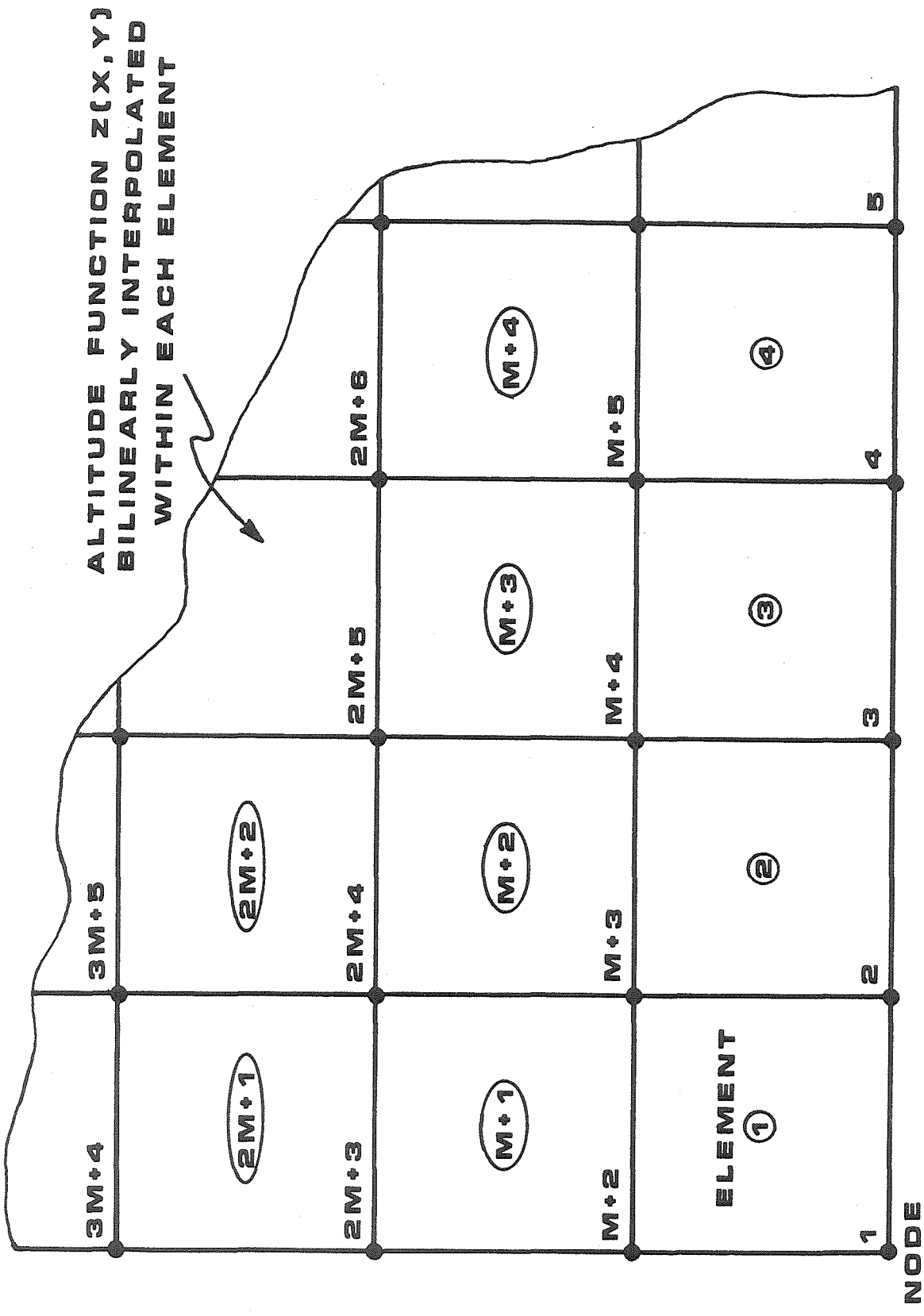
2.1 Finite Elements

The PC problem as formulated above is one of the constrained minimization of an integral functional. It is therefore ideally suited to solution by the method of finite elements (FE), which deals with systems of simultaneous integral equations (Zinkiewicz 1977; Stasa 1985). Since FE is best known as a method of solving differential equations, a brief outline of its operation is instructive. It may be divided into the following conceptual steps:

1) *Cast the problem in an integral form.* For differential equations one either sets various integrals over the residual to zero (the method of weighted residuals) or attempts to find an equivalent variational form of the problem. In the latter case the set of simultaneous equations is obtained by setting the derivatives of the functional to zero. The PC problem is already in integral form; indeed,

[4] Strictly speaking, the integration should be over $dX \, dY$, where the coordinates are chosen so Z is parallel to the (X, Y) plane. The given expressions are approximations accurate to $O(\tan(\Theta_1)\tan(\Theta_2))$, where Θ_1 is a typical angle between the normal to Z and the z axis, and Θ_2 is a typical slope of the surface with respect to Z.

Figure 2.2. Finite element mesh for photoclinometry . The M by N image is divided into square elements as shown, each identified with a pixel, and numbered sequentially from 1 to MN. Each element has four nodes, one at each corner; adjacent elements share nodes. The altitude in the interior of each element is obtained by bilinear interpolation between the values at the four nodes. Nodes are numbered sequentially from 1 to $(M+1)(N+1)$.


ALTITUDE FUNCTION Z(X,Y)
BILINEARLY INTERPOLATED
WITHIN EACH ELEMENT
3M+4
3M+5
2M+1
2M+2
2M+3
2M+4
2M+5
2M+6
M+1
M+2
M+3
M+4
M+5
ELEMENT
1
2
3
4
1
2
3
4
5
NODE

of the functional to zero. The PC problem is already in integral form; indeed, Wildey (1975) used the calculus of variations to convert it to a differential equation, which he then solved using finite differences!

2) *Divide the domain of solution into elements, i.e.*, choose a set of small regions of simple shape that cover the domain exactly once. For the PC problem, we choose an array of square elements, each centered on (and identified with) a pixel of the image.

3) *Specify an interpolation scheme.* In each element choose a set of *nodes* at which the discrete values of the desired quantities will be specified, and a method of interpolating the quantities between these nodes that satisfies the differentiability requirements of the problem. Nodes on the boundary between elements are shared and must have the same values in each (this assures that the different interpolations in neighboring elements match at the boundary). For the PC problem, we use four-node square elements with a node at each corner (cf. Figure 2.2). The altitude is then bilinearly interpolated within each element.

4) *Evaluate the integrals in terms of the nodal values.* The integral over each element may be performed separately, and the results "assembled." For linear problems and simple interpolation schemes, the element integrals may be performed explicitly; otherwise, it is necessary to choose a numerical scheme (*e.g.*, Gauss quadrature) to convert the integral to a sum.

5) *Solve the resulting system of algebraic equations.* This may be far from trivial, if the problem is nonlinear or very large (PC is both).

6) *Obtain the desired output quantities.* The nodal values may suffice; otherwise values at points inside the elements may be obtained using the interpolation scheme specified previously. For the PC problem it is convenient to interpolate

to get an M by N array of altitudes evaluated at the pixel centers.

Once the mesh of elements has been chosen, steps 4) through 6) may be carried out entirely by computer.

Although an arbitrary "footprint" in the PC problem may be handled by numerical quadrature, in this paper it is convenient to specialize to the case where $K(x,y)$ is narrow, and only one element contributes to each pixel. Then the further choice of one-point Gauss quadrature leads to a great simplification of the integrals (2.1). Requiring that K be normalized so its integral is unity, we can write

$$\iint K(x-\overline{x}^e, y-\overline{y}^e)\, b(\frac{\partial z(x,y)}{\partial x}, \frac{\partial z(x,y)}{\partial y})\, dx\, dy \simeq b(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y})\Big|_{\overline{x}^e, \overline{y}^e}. \tag{2.4}$$

The right-hand side depends only on the nodal altitudes in the element e. It is convenient to adopt vector notation: let $\{z\}$ be a column vector of length $(M+1)(N+1)$ containing the nodal altitudes, $\{B\}$ be a vector of length MN containing the observed brightnesses, and $\{b\}$ the corresponding estimates based on $\{z\}$. This single-index numbering scheme may be maintained in parallel with the more familiar two-index scheme. Conceptually, this is a matter of *lexicographic ordering* (Andrews and Hunt 1977, p. 40). In a `FORTRAN` implementation of the algorithm we use the `EQUIVALENCE` statement to identify a one-dimensional vector with a two-dimensional array. If we number the nodes and elements in the order shown in Figure 2.2, and express the paired indices in the order (column, row) appropriate to a discretized (x,y) coordinate system, then the mesh is said to be *row scanned.* The vector will consist of the rows of the array, one after another. In vector notation the PC problem may be formulated as follows:

$$\left\{\frac{\partial S}{\partial \{z\}}\right\} = 0, \qquad \text{subject to} \qquad \{b(\{z\})\} = \{B\}. \tag{2.5}$$

The potentially confusing notation $\left\{\frac{\partial S}{\partial \{z\}}\right\}$ simply means the column vector whose ith entry is $\frac{\partial S}{\partial z_i}$.

2.2 Penalty Method Minimization

The way in which the constraints are imposed on the extremization problem (2.5) is important to the efficiency of the resulting algorithm. An exact solution to the constrained minimization could be obtained by the introduction of MN Lagrange multipliers λ^e, one for each element:

$$\delta\Big(S + \langle\lambda\rangle(\{B\} - \{b\})\Big) = 0, \tag{2.6}$$

where $\langle\lambda\rangle = \{\lambda\}^T$ is a row vector and the minimization is to be performed by varying both the nodal z_i values and the λ^e. Not only does this method nearly double the number of equations and unknowns, but it involves a mix of element quantities and nodal quantities, making efficient ordering of the equations difficult.

Instead, we use an approximate method of doing the minimization known as the penalty method (Albert 1972):

$$\delta\Big(S + \frac{\alpha}{2}(\langle B\rangle - \langle b\rangle)(\{B\} - \{b\})\Big) = 0. \tag{2.7}$$

This equation superficially resembles (2.6) but there is a crucial difference: here α is an arbitrary large ($O(10^4)$) constant known as the *penalty number*, and the minimization is done only with respect to the z_i. Dividing through by α, and making the extremization explicit, we obtain the (nonlinear) matrix equation:

$$\frac{1}{\alpha}\left\{\frac{\partial S}{\partial\{z\}}\right\} - \left[\frac{\partial\langle b\rangle}{\partial\{z\}}\right](\{B\} - \{b\}) = 0. \tag{2.8}$$

Here, $\left[\frac{\partial\langle b\rangle}{\partial\{z\}}\right]$ is an $(M+1)(N+1)$ by MN matrix with (i,e)th entry $\frac{\partial b^e}{\partial z_i}$. We begin to see how the penalty method works: we take the MN brightness constraints $\{B\} - \{b\} = 0$ and combine them linearly according to the matrix $\left[\frac{\partial\langle b\rangle}{\partial\{z\}}\right]$ to get $(M+1)(N+1)$ equations. Conveniently, this particular choice of combinations will, upon linearization, yield a symmetric matrix equation. The system is singular, so we add in a *small* amount $\frac{1}{\alpha}$ of the $(M+1)(N+1)$ equations implied by the roughness

criterion. A tradeoff is available: as α is increased, the brightness constraints are more precisely satisfied but the system moves toward singularity, whereas when α is decreased, the set of equations becomes better conditioned but accuracy is lost.

An additional desirable property of the penalty method is the fact that we are minimizing the norm of $\{B\} - \{b\}$, rather than directly setting it to zero. Thus, if the brightness constraints are inconsistent (due to the presence of noise, or to albedo variations in violation of the assumption of uniformity, say), the method will fail gracefully. A smooth solution will be chosen that reproduces the observed brightnesses as nearly as possible.

2.3 Newton-Raphson Iteration

We now have the PC problem in a concise form: "invert equation (2.8)." Unfortunately, although this equation involves a matrix multiplication, both the matrix and vectors depend nonlinearly on $\{z\}$, so the solution is more than a matter of linear algebra. We must iterate for the solution using the Newton-Raphson method (Newton's point-slope root finding algorithm generalized to a system of equations). The $k+1$st approximation to $\{z\}$ is obtained by solving:

$$[K^k]\{\Delta z^k\} = \{E^k\}, \tag{2.9a}$$

for $\{\Delta z^k\}$ and forming

$$\{z^{k+1}\} = \{z^k\} + \{\Delta z^k\}, \tag{2.9b}$$

where

$$[K^k] \equiv \frac{1}{\alpha}\left[\frac{\partial^2 S}{\partial\{z\}\partial\langle z\rangle}\right] + \left[\frac{\partial\langle b\rangle}{\partial\{z\}}\right]\left[\frac{\partial\{b\}}{\partial\langle z\rangle}\right] \tag{2.9c}$$

is the *Hessian matrix*, and

$$\{E^k\} \equiv -\frac{1}{\alpha}\left\{\frac{\partial S}{\partial\{z\}}\right\} + \left[\frac{\partial\langle b\rangle}{\partial\{z\}}\right](\{B\} - \{b\}) \tag{2.9d}$$

is the gradient of the function being minimized, both evaluated at $\{z\} = \{z^k\}$. A second derivative term $\sum_e \left[\frac{\partial^2 b^e}{\partial\{z\}\partial\langle z\rangle}\right](B^e - b^e)$ has been omitted from $[K]$ (Press et al.

1986, p. 523). When $\{B\} - \{b\} \simeq 0$, this term merely contributes a small amount of noise to the equations, decreasing their stability. Clearly, use of the modified Hessian matrix does not affect the property $\{\Delta z\} \rightarrow 0$ as $\{E\} \rightarrow 0$ that assures that we obtain the correct solution. I will discuss the problem of obtaining an initial estimate $\{z^0\}$ below.

2.4 Successive Over-Relaxation

Efficient solution of (2.9a) requires that we exploit to the fullest degree possible the special properties of the Hessian matrix $[K]$: though large, it is symmetric, banded, and highly sparse in an orderly way. The (i,j) component connects node i to node j in terms of the derivatives of S and b, each of which may be divided into a sum over contributions from the various elements. Thus, $K_{ij} \neq 0$ only if there exist one or more elements containing both nodes i and j. In the numbering system of Figure 2.2, node i connects to itself and to its eight nearest neighbors $i-M-2$, $i-M-1$, $i-m$, $i-1$, $i+1$, $i+m$, $i+M+1$, and $i+M+2$. (For nodes at the edge of the mesh, some of these points may not exist.)

Even taking advantage of the fact that $[K]$ is symmetric with half-bandwidth $M+3$, a direct solution of equation (2.9a) will require $\sim M^3N$ multiplications (operation counts will be given only to leading order) for matrix factorization, and $\sim 2M^2N$ more for forward reduction/back substitution *at each step of the Newton-Raphson method.* Storage will be dominated by the $\sim M^2N$ locations needed for the matrix because the sparsity is destroyed. With M, N $O(10^2)$, the problem is obviously too large to be tractable.

There exist a number of methods for solving sparse systems of equations which leave the zero elements of the matrix unfilled-in. These methods also have in common the property of being *iterative*, that is, of requiring an indeterminate number of repeated steps to attain a result with a specified accuracy. For such a method to

be practical in terms of computation as well as storage, the number of iterations actually required must not be too large. One extremely powerful such technique, the incomplete Cholesky-conjugate gradient method, or ICCG (Meijerink and van der Vorst 1977; Kershaw 1978) was applied to the PC problem without complete success. In this method, the usual Cholesky decomposition form of Gaussian elimination is carried out on the matrix $[K]$, except that whenever this algorithm would make a previously zero element nonzero, the zero element is left unchanged. The result is an approximate factorization $[L][U] \simeq [K]$ in which $[L]$ and $[U]$ have the same sparsity as $[K]$ ($[U] = [L]^T$ for a symmetric matrix). This factorization does not lead *directly* to a solution, but the facts that $([L][U])^{-1}$ is trivially computed and that $([L][U])^{-1}[K]$ is nearly the identity matrix can be exploited to greatly increase the efficiency of the iterative conjugate method due to Hestenes and Stiefel (1952). The ICCG method works very well for many problems, *including* the initial linearization of the PC equations about the plane $z = 0$. Its weakness is that pivoting cannot be incorporated into the incomplete factorization process. There exist *ad hoc* fixes which allow one to proceed despite bad pivots, but if such pivots are too numerous the subsequent conjugate gradient iteration will not succeed. This turns out to be the case for the subsequent iterations of the PC equations. Bad pivots are inevitably encountered, and the iteration process diverges, adding increasing amounts of high frequency "checkerboard" noise to the solution.

The method of successive over-relaxation (SOR) was found to be much better suited to the PC problem. In contrast to factorization, SOR is based on an *additive* decomposition $[K] = [L] + [D] + [U]$ where $[D]$ is a diagonal matrix, $[L]$ is lower-triangular with zeroes on the diagonal, and $[U]$ is upper-triangular (Ortega 1970; Press *et al.* 1986, pp. 652–659). Since the elements of the matrices in the decomposition are the same as those of $[K]$, no additional memory locations are required. We start

with $\{\Delta z^{k,0}\} = 0$ and iterate, solving

$$\Big([L^k] + [D^k]\Big)\{\Delta\Delta z^{k,l}\} = \{E^k\} - [K^k]\{\Delta z^{k,l}\}, \tag{2.10a}$$

and forming

$$\{\Delta z^{k,l+1}\} = \{\Delta z^{k,l}\} + \omega\{\Delta\Delta z^{k,l}\}. \tag{2.10b}$$

The quantity ω is known as the *relaxation parameter*; strictly speaking, we are over-relaxing only for $\omega > 1$. When $\omega = 1$ the method is known as Gauss-Seidel iteration. The PC algorithm utilizes *Chebysheff acceleration* (Press *et al.* 1986, p. 658), in which $\omega = 1$ immediately after each linearization, gradually increasing to its (empirically determined) optimum value. At the beginning of each Newton-Raphson step $\sim 5MN$ multiplications are needed to form $[K]\{\Delta z\}$ for the right-hand side. Then another $\sim 5MN$ multiplies are needed to do the forward reduction to get each increment $\{\Delta\Delta z\}$. The method will thus be faster than factorization if less than $\sim M^2/5$ iterations are needed (though it will always use less memory).

Unfortunately, it is possible to show that $O(M^2)$ iterations are needed to achieve convergence of the Newton-Raphson method. We can make this result intuitive by considering how the forward reduction process used to solve (2.10a) works. We sweep through the mesh, considering each node i in turn. The value that satisfies the ith equation is chosen, and the appropriate multiples of it are subtracted from all succeeding equations. Each $\Delta\Delta z_i^l$ thus takes into account the recent changes at all previous nodes, but (since the above-diagonal terms have been set to zero) not the possible effects due to increments at nodes yet to be considered. There is information about these "downstream" nodes in the right-hand side, but since the matrix $[K]$ connects only neighboring nodes, $\Delta\Delta z_i^l$ knows about $\{\Delta\Delta z^{l-1}\}$ only at its nearest neighbors, $\{\Delta\Delta z^{l-2}\}$ two elements away, and so on. Information about the solution propagates *diffusively*, and it is thus not surprising that $O(M^2)$ steps are required for global convergence on a mesh of width M.

The successive over-relaxation method attempts to achieve convergence in fewer steps by exaggerating the correction at each step by a factor ω. This over-correction can be thought of as the addition of a kind of "inertia" to the system, allowing the "wavelike" propagation of information. Hence it is not surprising that one can show that, for simple problems (*e.g.*, Poisson's equation) with the correct choice of ω, convergence takes only $O(M)$ iterations. This is a substantial improvement over the Gauss-Seidel method, but still impractically slow for reasonable sized images. Furthermore, the optimal value of ω for a complicated problem such as photoclinometry can be estimated only by trial and error.

2.5 Multigridding

A powerful method of accelerating convergence known as multigridding (Brandt 1977) was incorporated in the PC algorithm presented here. The success of multigridding is based on the fact that (as shown above) SOR (or Gauss-Seidel) iteration rapidly eliminates local errors in the solution but is slow to correct errors involving nodes many elements apart. Complimentary corrections may thus be obtained by solving the equivalent problem on a coarser mesh. If the mesh spacing is doubled, not only does information propagate twice as far per SOR iteration, but each iteration requires only one quarter as much computation; the long-wavelength components of the solution are thus obtained very efficiently. Error will be introduced by interpolating the coarse-mesh correction onto the fine mesh, but this will be local and hence rapidly eliminated by iteration at the higher resolution. Of course, the method may be applied recursively, with a quarter-resolution mesh used to provide corrections to the half-resolution mesh, and so on. If necessary, the solution on the coarsest mesh may be obtained by noniterative means. In any event, the total work can be shown to be $O(MN)$, *i.e.*, the number of iterations per element needed at all resolutions does not depend on M or N.

To implement the multigrid algorithm we need to know *how* to pass from one resolution to another, and *when* to do so. Brandt (1977) considers problems of the form $LU = F$ (plus boundary conditions), where L is a possibly nonlinear differential operator acting on U, and we want to model this system discretely as $Lu = f$. He concludes "Full efficiency of the multigrid algorithm is obtained for stopping parameters that do not depend on the geometry and the mesh size, and which may change over a wide range, provided the correct *forms* of the stopping criteria are used and some basic rules of interpolation are observed." Those criteria and rules may be summarized as follows:

1) The coarse mesh should always have half the resolution of the fine mesh. This is near optimal, and the standardization is worthwhile.
2) Decrease resolution when convergence is "slow," *i.e.*, when the norm of the residual is more than some fraction η of its value at the previous iteration. For simple problems an optimal value of η may be derived, but any value $\eta \lesssim 0.9$ was found to be acceptable. On the coarsest mesh, of course, iteration is cheap and this criterion may be ignored.
3) Decrease resolution by injecting the solution u, and the *error* in the right-hand side $f - Lu$ rather than f itself. This assures that the correction to u on the coarse grid is an approximation to what is needed on the fine grid. By injection is meant interpolation or simply copying of the nodal values if the nodes of the coarse mesh are a subset of the nodes of the finer mesh.
4) Increase resolution when the residual on the coarse mesh is less than some fraction δ of the previous residual on the fine mesh (or is limited by the attainable accuracy. If this latter obtains at the highest resolution we are done.) Again, the optimal δ may be estimated for simple problems but it is not critical; $0.001 \leq \delta \leq 0.5$ was found to slow the algorithm by less than 20% from

the optimum.

5) Increase resolution by interpolating the *changes* to u made on the coarse mesh and adding them to the fine mesh solution. This assures that the short-wavelength part of the solution obtained previously is not thrown away. The order of interpolation should be equal to or greater than the order of the differential equation.

Note that the regridding process effectively includes a linearization of the problem (the value of $\{B\}-\{b\}$ injected from the higher resolution is used at all subsequent iterations on the coarser mesh). It thus does not in general suffice to pass through each mesh size once; the algorithm "wanders" up and down in resolution until the desired accuracy is achieved.

These rules were applied somewhat loosely to the PC problem. As shown in Figure 2.3, the coarse mesh was laid out with each element occupying the space of four elements of the next finer mesh. On decreasing resolution the z values at the corner of the big element were simply copied (injected) from the corresponding nodes on the fine mesh. Corrections to the altitudes on the fine mesh were obtained by bilinear interpolation of the changes made during iteration on the coarse mesh[5]. The Euclidean norm of the residual vector $\{E\}$ (eq. 2.9d) was used in the stopping criteria, subject to being overruled by the user's judgement.

Injection of the brightness information is less straightforward. The constrained minimization equation (2.8) is not in the form $Lu = f$, but the underdetermined brightness problem $\{b(\{z\})\} = \{B\}$ on which it is based *is*. To choose $\{B\}$ on the coarse mesh so that $\{E\}^{coarse}$ equals the injection of $\{E\}^{fine}$, we would have to solve a large system of equations $\left[\frac{\partial\{b\}}{\partial\{z\}}\right]^{coarse} \{B\}^{coarse} = \cdots$, which is exactly the kind of

[5] Actually, the altitudes were halved on injection, and doubled on interpolation, so that on each mesh z was measured in units of the node spacing. The same functional form of b could then be used at all resolutions.

Figure 2.3. Coarse and fine meshes for multigridding . The fine mesh of Figure 2.2 is shown dashed with the new, half-resolution mesh superimposed. The old nodal values are "injected," *i.e.*, those occurring at nodes of the new mesh are simply copied to it. The new brightness is chosen so that the error in its estimate is injected. After iteration at low resolution, any changes in the nodal altitudes will be interpolated so that they can be applied at all nodes of the fine mesh.

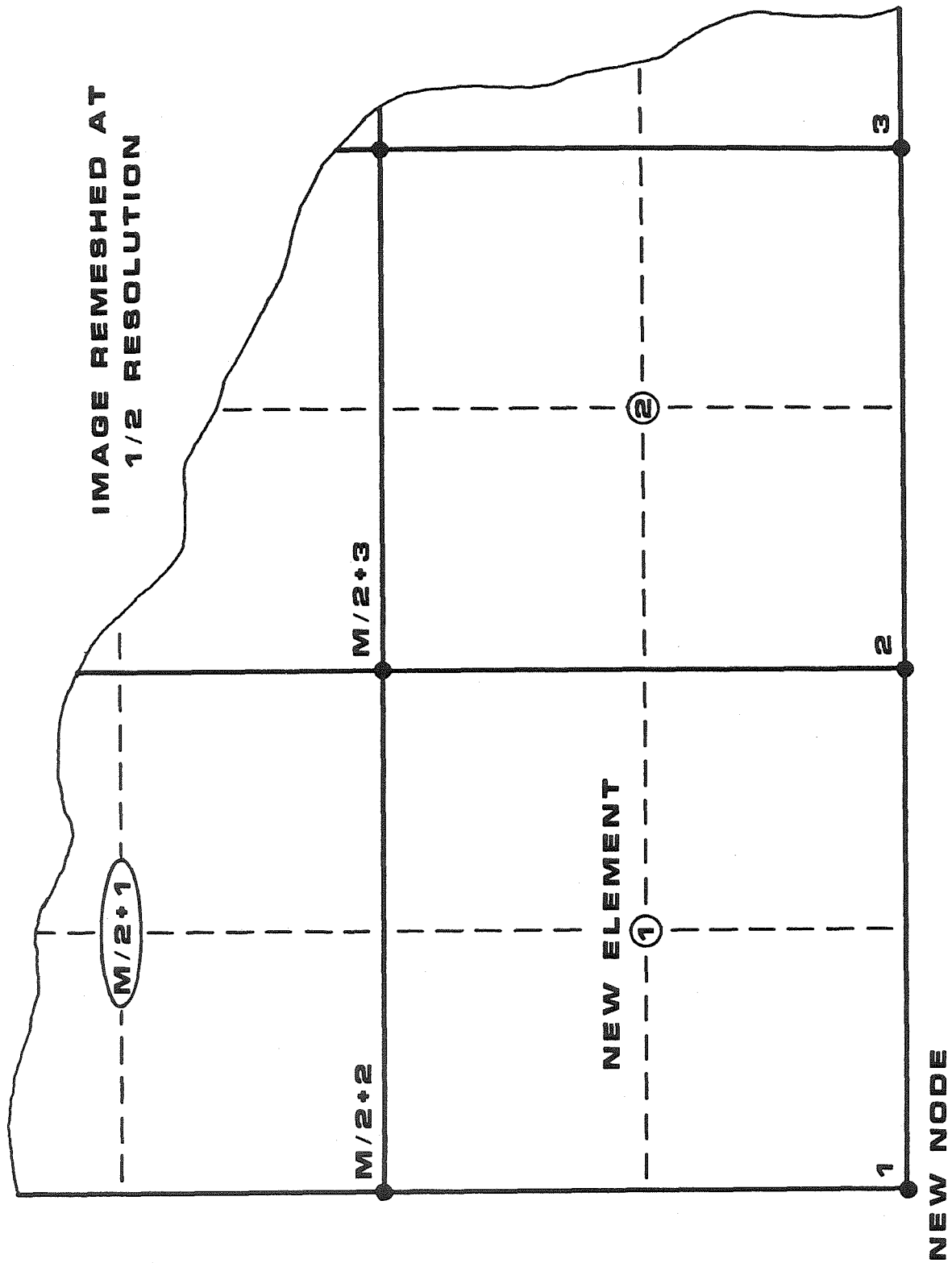


IMAGE REMESHED AT
1/2 RESOLUTION
M/2+1
M/2+2
M/2+3
NEW ELEMENT
1
2
3
NEW NODE

thing we are attempting to avoid doing. We therefore take the simpler approach of choosing $\{B\}^{coarse}$ so that $(\{B\} - \{b\})^{coarse}$ is the injection of $(\{B\} - \{b\})^{fine}$. The consequences of this approximation will become apparent in Section 3.

2.6 The Initial Estimate

So far nothing has been said about the altitude estimate $\{z^0\}$ with which to begin the iteration cycle (2.10). The limiting behavior of the Newton-Raphson method can depend on the initial guess in a remarkably complex way (Peitgen and Richter 1986), but experience shows that a reasonable solution to the PC problem can often be obtained starting with $\{z^0\} = \{Z\}$, where $\{Z\}$ is the mean plane approximating the surface of the object in the field of view. If the surface is approximately perpendicular to the line of sight, however, the remarkable properties of the PC equations linearized about $\{z\} \equiv 0$ make possible a noniterative solution for $\{z^0\}$ which is an excellent approximation to the true $\{z\}$. Not only is the number of subsequent iterations required greatly reduced, but the iteration process is more stable in the presence of noise.

Consider a Maclaurin series expansion for the brightnesses:

$$\{B\} = \{b(\{z\})\} \simeq b_0\{1\} + \left[\frac{\partial\{b\}}{\partial\langle z\rangle}\right]\Bigg|_0 \{z\} + \cdots, \tag{2.11}$$

where b_0 is the brightness of a level surface, $\{1\}$ a vector of MN ones, and $[A] \equiv \left[\frac{\partial\{b\}}{\partial\langle z\rangle}\right]\Big|_0$ the $(M+1)(N+1)$ by MN matrix of partial derivatives evaluated for a flat surface. The first two terms of (2.11), taken as an equation for the initial approximation to the altitudes, form an underdetermined system. It is a standard result of linear algebra, however (Pratt 1978), that the solution to this equation with minimum norm $|\{z\}|$ is the pseudoinverse (PI), $\{\hat{z}\}$, which obeys

$$\{\hat{z}\} = [A]^T\{\beta\}, \tag{2.12a}$$

where

$$[A][A]^T\{\beta\} = \Big(\{B\} - b_0\{1\}\Big). \tag{2.12b}$$

Solution for $\{\hat{z}\}$ turns out to be simple, and $\{\hat{z}\}$ is the desired $\{z^0\}$, provided we take the rms altitude (2.3a) as our roughness function, with $\{Z\} = 0$. This formulation is not useful for subsequent iterations because 1) the special properties of the matrix are lost, and 2) we would obtain a minimum-norm increment to $\{z\}$, rather than an increment to the minimum-norm solution for $\{z\}$.

To see why (2.12) is easy to solve, we must look at the structure of $[A]$, and for this it is convenient to introduce a *local* numbering of the nodes in a given element (cf. Appendix A). Returning to the definition of $[A]$ above, we see that its (e,i) entry $\frac{\partial b^e}{\partial z_i}$ relates node i to element e, and is zero if i is not one of the four nodes in e. Furthermore, since the linearization was done about a uniform flat surface, the nonzero values can depend only on the local index (1 = upper right ... 4 = lower left), and not on e. Now, we have chosen the (x,z) plane to be the phase plane, so $\frac{\partial b}{\partial(\partial z/\partial y)}$ evaluated for a level surface must vanish; b is symmetric in $\frac{\partial z}{\partial y}$. The final simplification comes from the expression for the in-phase-plane slope at the center of the element: $\frac{\partial z}{\partial x} = \frac{1}{2}(-z_1^e + z_2^e + z_3^e - z_4^e)$, using the local numbering and expressing z in units of pixel widths. Thus we obtain $\frac{\partial b}{\partial z_1^e} = -\frac{\partial b}{\partial z_2^e} = -\frac{\partial b}{\partial z_3^e} = \frac{\partial b}{\partial z_4^e} = -\frac{1}{2}\left.\frac{\partial b}{\partial(\partial z/\partial x)}\right|_0 \equiv b'$.

Rewriting the previous equations in the form:

$$\{\hat{z}\} = [\mathcal{A}]^T\{\mathcal{B}\}, \tag{2.13a}$$

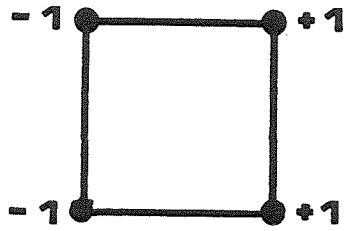
where

$$[\mathcal{A}][\mathcal{A}]^T\{\mathcal{B}\} = \frac{\{B\} - b_0\{1\}}{b'}, \tag{2.13b}$$

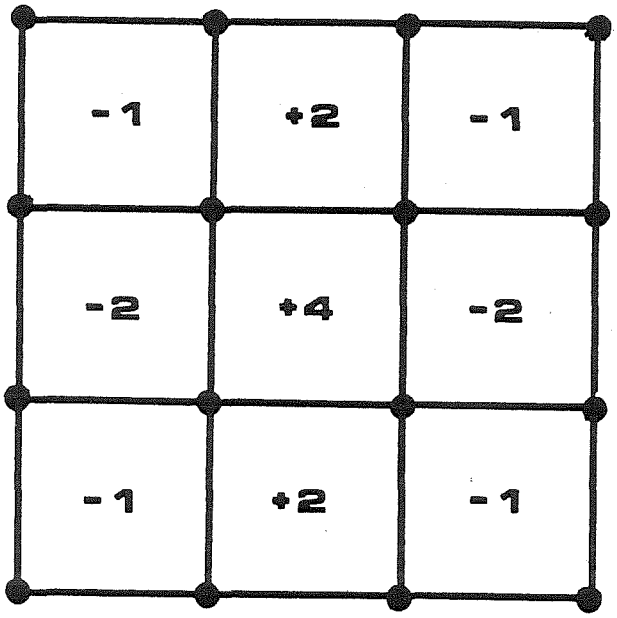
we obtain matrices $[\mathcal{A}] = \frac{1}{b'}[A]$ and $[\mathcal{A}][\mathcal{A}]^T$ that are independent of the photometric function chosen as well as of position e within the image. Figure 2.4a shows the weights by which $[\mathcal{A}]$ relates an element to its four nodes; since all are ± 1, the calculation of (2.13a) requires no storage and no multiplications, only additions and

Figure 2.4. Matrices used in the initial estimate of z by SSIPSF-PI. (a) The entries of $[\mathcal{A}]$ relate an element e (shown) to the nodes at its corners by ± 1 as shown. (b) The entries of $[\mathcal{A}][\mathcal{A}]^T$ relate an element e (in center) to itself and its neighbor elements by a sum over entries of $[\mathcal{A}]$ for nodes held in common.

a



b



subtractions. The weights by which $[\mathcal{A}][\mathcal{A}]^T$ relates an element to its neighbors (by a sum over the entries of $[\mathcal{A}]$ for nodes they have in common) appear in Figure 2.4b. Inspection shows that the matrix is *separable* into operations on neighboring rows and operations on columns. Mathematically, it is expressible as an open product

$$[\mathcal{A}][\mathcal{A}]^T = [R] \otimes [C], \tag{2.14a}$$

where

$$[R] = \begin{bmatrix} 2 & -1 & & & & 0 \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & & & \\ 0 & & & & -1 & 2 \end{bmatrix} \tag{2.14b}$$

is a tridiagonal Toeplitz matrix operating on each row, and

$$[C] = \begin{bmatrix} 2 & +1 & & & & 0 \\ +1 & 2 & +1 & & & \\ & +1 & 2 & +1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & & & \\ 0 & & & & +1 & 2 \end{bmatrix} \tag{2.14c}$$

is another such matrix operating on columns. It follows that the *inverse* of the matrix may also be so decomposed (Pratt 1978):

$$\left([\mathcal{A}][\mathcal{A}]^T\right)^{-1} = [R]^{-1} \otimes [C]^{-1}. \tag{2.15}$$

These tridiagonal matrices are readily factored in advance, and the solution to (2.13b) may be obtained by doing forward reduction and back substitution first on each row, then on each column. Approximately $2M$ multiplications per row and $2N$ per column, or a total of ~ 4 per pixel, are required.

In the jargon of image processing, the matrix $[\mathcal{A}][\mathcal{A}]^T$ expresses a separable, spatially invariant point-spread-function (SSIPSF) operating on the brightness image (Andrews and Hunt 1977, p. 70). I therefore refer to the pseudoinverse solution by the abbreviation SSIPSF-PI.

3. Demonstration of the Algorithm

In testing the PC algorithm, it is desirable to work with a dataset for which the true topography is known, so that the error in the result may be calculated. At the same time, because we are testing an *ad hoc* roughness criterion, it is important that the data have the roughness properties of a real geologic surface. Testing of the algorithm on images with a wide range of illumination geometries, photometric functions, signal to noise ratio, etc. is also desirable. To fulfill all of these objectives, a series of pseudoimages were generated from actual digitized topography.

Figure 3.1 shows the region chosen for study: a 60 by 60 km square area in the Wasatch mountains in northern Utah. A portion of the National Digital Topographic Dataset covering the northern third of Utah was obtained from L. A. Soderblom at the U. S. Geological Survey. The raw data are in the form of elevations, rounded to the nearest foot (but largely binned at intervals of $\sim$ 170 feet), on a mesh of spacing 150 m. To smooth out the artificial "cliffs" at the edges of the 170-foot plateaus, the data were convolved six times with a Laplacian filter. Alternating data points in the study area were split between two meshes of 300 m spacing: one, 201 by 201, to serve as nodal values from which to compute the brightnesses of an array of 200 by 200 elements, and the second, 200 by 200, to serve as the reference to which to compare the PC elevations in the center of each element. (An rms error of $\sim$ 1.2 m was found when the larger dataset was interpolated to the centers of the elements *without* doing PC and compared with the smaller set.) Finally, the elevations were rescaled in terms of the width of the elements and the mean altitude was subtracted, since PC is incapable of determining absolute elevation. Figure 3.2a shows the topographic dataset for the study area in greyscale form, with brightness corresponding to elevation.

A series of pseudoimages were created from the resulting nodal elevations by calculating the orientation at the center of each element by bilinear interpolation and

Figure 3.1. Study area for PC algorithm . Digital topographic data for a 60 km square area (stippled) in northern Utah were used to generate pseudoimages for inversion by the PC algorithm.

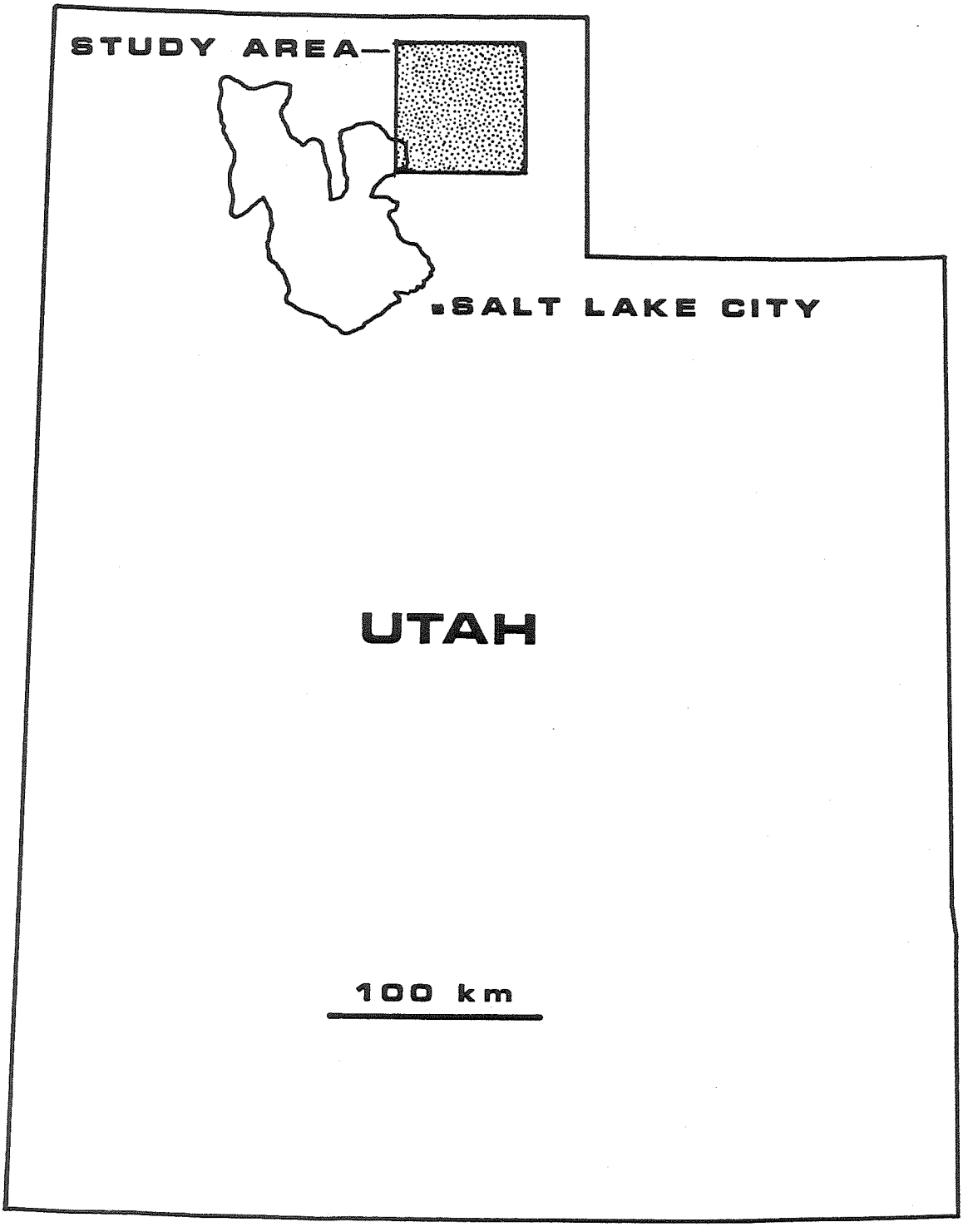
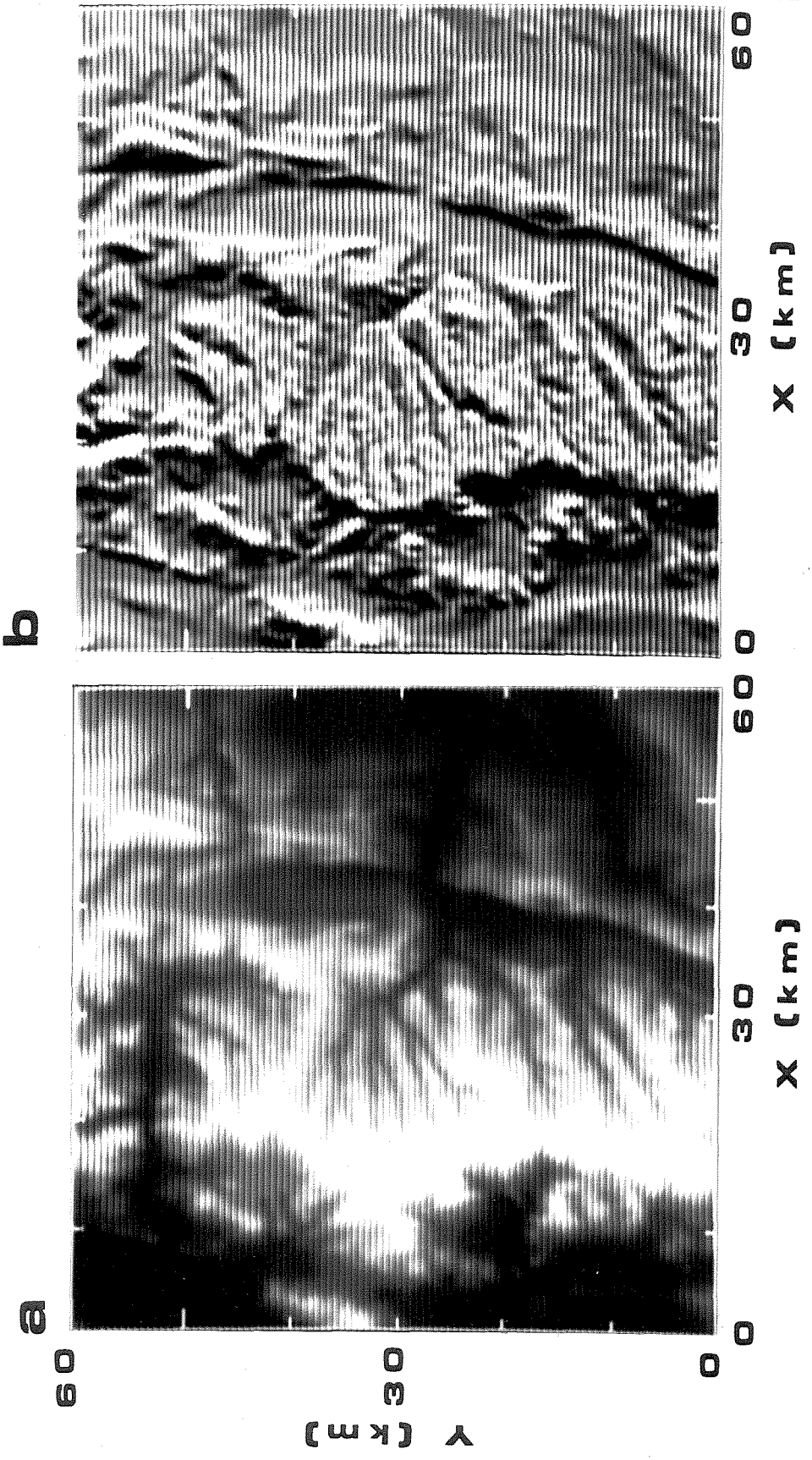

STUDY AREA
SALT LAKE CITY
UTAH
100 km

Figure 3.2. Study area topography and pseudoimage . (a) Greyscale representation of digital topography for the study area (Figure 3.1). Darkest tone corresponds to an elevation of 1800 m, brightest to 2400 m. (b) Pseudoimage created from data in (a), asssuming a uniform Lambert scattering surface, illuminated from the right with a phase angle $g = 45°$. Image is contrast-enhanced; actual contrast is $\sim 4.4\%$.


b
0
30
60
X (km)
a
0
30
60
X (km)
0
30
60
Y (km)

assuming a Minnaert photometric function:

$$b = F(g)(\cos\theta)^{k(g)}(\cos\phi)^{k(g)-1} \tag{3.1a}$$

$$= F(g)(\hat{n}\cdot\hat{\imath})^{k(g)}(\hat{n}\cdot\hat{e})^{k(g)-1}. \tag{3.1b}$$

Refer to Figure 2.1 for the definitions of the incidence angle θ, emission angle ϕ, phase angle g, and the various unit vectors. This form of the photometric function was chosen because of its simplicity and because of its applicability to bright, icy planetary surfaces such as the polar regions of Mars and the satellites of the outer planets (Veverka 1973) and to bright coatings such as colloidal MgO that can be generated in the laboratory. I give results here only for $k = 1$, known as Lambert scattering; the behavior of the photometric function for $k \neq 1$ is not very different qualitatively. Appendix A gives the details of how (3.1) was computed in terms of the nodal elevations — both in making the pseudoimage and during the PC algorithm.

Preliminary experimentation indicated that a relaxation parameter $\omega \sim 1.5$ was close to optimal (at least in the early stages of iteration), and that it was satisfactory to do SOR iteration until $\frac{|\Delta\Delta z|}{|\Delta z|} \lesssim 0.1$. (For the first linearization at a given resolution convergence was faster and a cutoff of 0.2 was used.) The multigrid stopping parameters $\eta \simeq 0.8$ and $\delta \simeq 0.3$ were also used throughout. Other parameters were varied about a "nominal" case using the area roughness criterion, with a penalty number $\alpha = 10^4$ on an image of a Lambert surface ($k = 1$) at a phase angle of $g = 45°$, with no quantization, noise, or blur. Figure 3.2b shows the shaded image for this nominal case. Note that it has been contrast stretched; the rms dispersion in brightness as a fraction of the mean is only $\sim 4.4\%$.

I will display the output of the PC algorithm for the nominal case in a variety of forms. The four panels of Figure 3.3 are greyscale representations of the initial SSIPSF-PI altitude estimate, the error in this estimate, the final estimate after 96 iterations, and the error in the latter. The same mapping between elevation and

brightness as in Figure 3.2a has been used in all four panels. The most obvious property of the residuals is the organization of most of the the error into "stripes" parallel to the phase plane. It is also apparent that the SSIPSF-PI estimate is lower (darker) in the west than it ought to be; this is a consequence of using a linear approximation to the photometric function. The final estimate shows much less of this bias. Close examination reveals a smaller component of error correlated with the curvature of the topography, *i.e.*, tracing out the ridges and valleys.

We can get a better idea of the magnitude of the residual from the perspective plots, Figure 3.4. The exact topography of the study area is shown in 3.4a, while parts b and c show respectively the error in the SSIPSF-PI estimate and in the final estimate. The view is from the southeast at an elevation of 45°, and the vertical exaggeration is 25 : 1 in all cases. The final error is modest in comparison with the scale of the topography, and the error within any given row is so small as to be essentially invisible. These assertions are quantified by the histograms of the distribution of altitudes in Figure 3.5. The first three panels of this figure correspond to the three parts of Figure 3.4, while the fourth shows the distribution of residuals *within rows* of the final estimate (note that the scales in each panel are different). The rms error in the final estimate is 22.3 m (4.4 m within rows), compared with a total range of elevation of over 600 m. A final look at the residuals is provided in Figure 3.6. Estimates of the one-dimensional power spectral density of the ensemble of rows (3.6a) and of columns (3.6b) are presented. Once again, the three curves in each panel correspond from top to bottom to the topography (Figure 3.4a), the error in the SSIPSF-PI estimate (3.4b), and the final error (3.4c). The increase in signal to noise ratio is dramatic for the rows, especially at intermediate frequencies. The error in the column direction is affected much less by iteration except at frequencies $\gtrsim 0.5\,\mathrm{km}^{-1}$. This is, however, where the SNR was initially lowest (< 1 above $\sim 0.8\,\mathrm{km}^{-1}$).

Figure 3.3a, b. SSIPSF-PI estimate of topography and residual . Correspondence between elevation and brightness in all panels is the same as in Figure 3.2a. (a) Greyscale representation of SSIPSF-PI estimate of topography. (b) Difference between SSIPSF-PI estimate and true topography. Note striping, low elevations (dark) on left side of region.

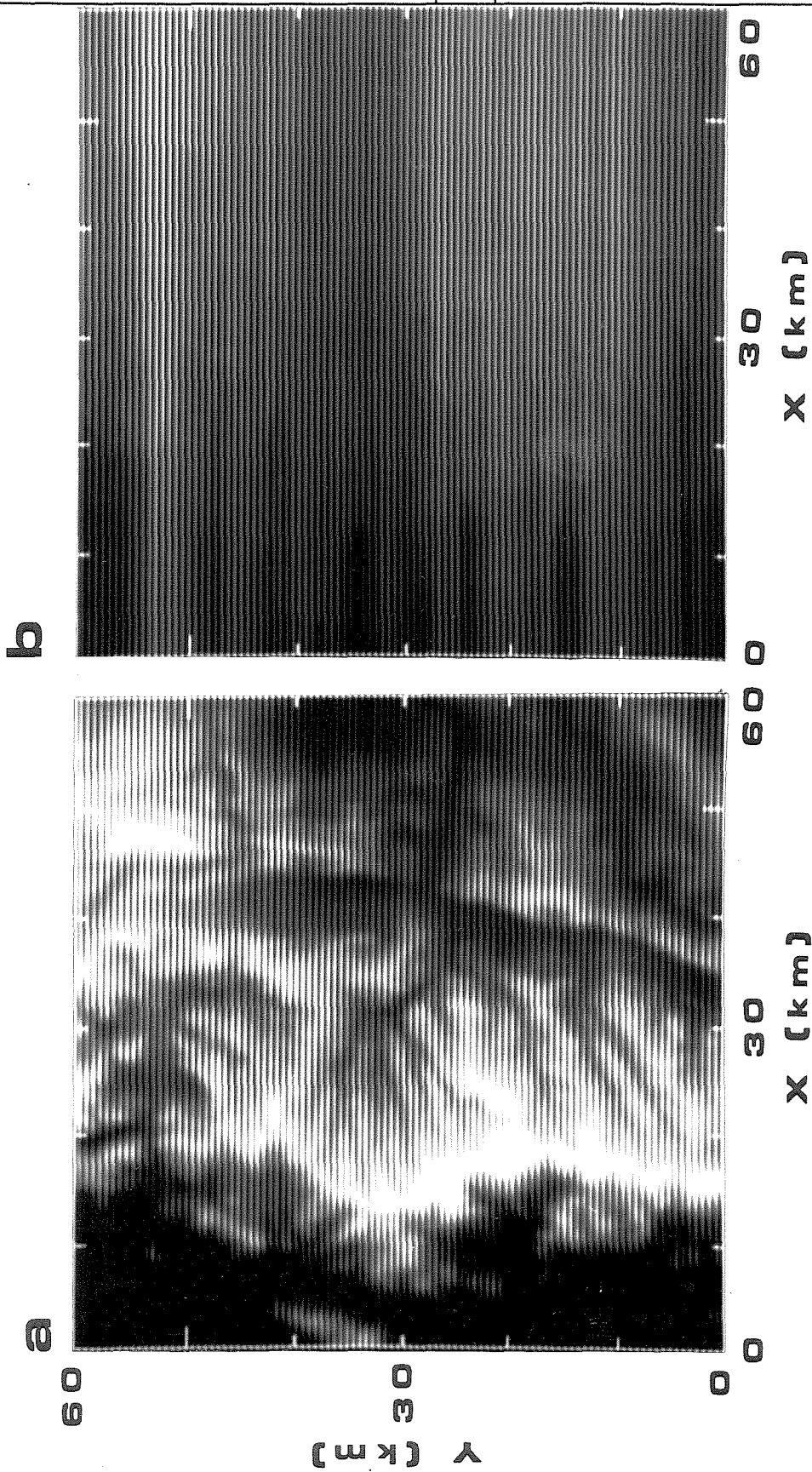

b
0
30
60
X (km)
a
0
30
60
X (km)
0
30
60
Y (km)

Figure 3.3c, d. Final PC estimate of topography and residual . (c) Estimated topography after 96 iterations with multigridding. (d) Difference between final estimate and true topography. Striping is reduced but still present; nonstripelike errors are greatly reduced.

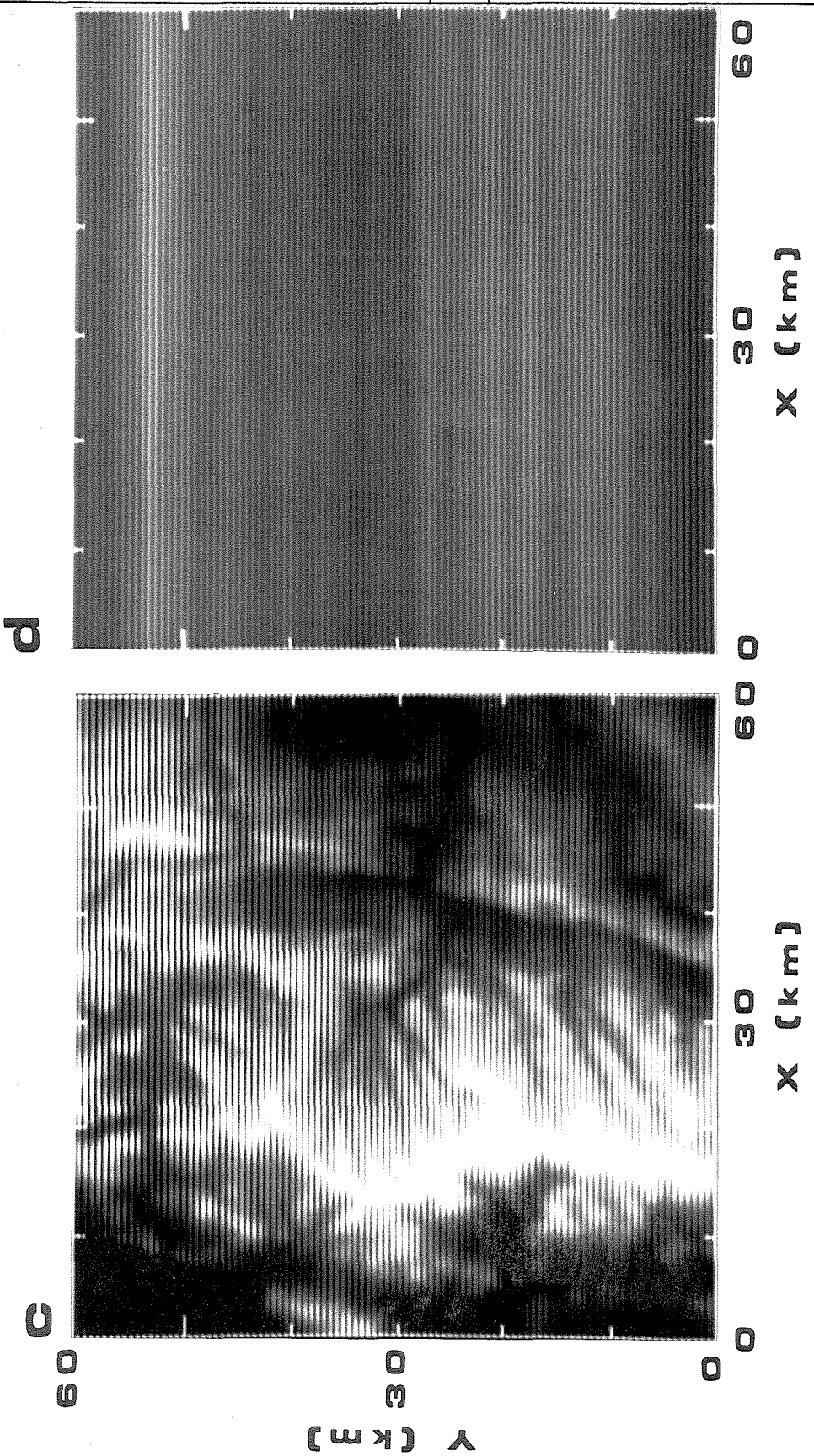

d
0
30
60
X (km)
c
0
30
60
X (km)
0
30
60
Y (km)

Figure 3.4. Perspective plots of topography and residuals . All plots are viewed from the southeast (lower left) corner at an elevation of 45° and have a vertical exaggeration of 25 : 1. (a) Exact topography (compare Figure 3.2a). (b) Error in SSIPSF-PI estimate (Figure 3.3b). (c) Error in final estimate (Figure 3.3d).

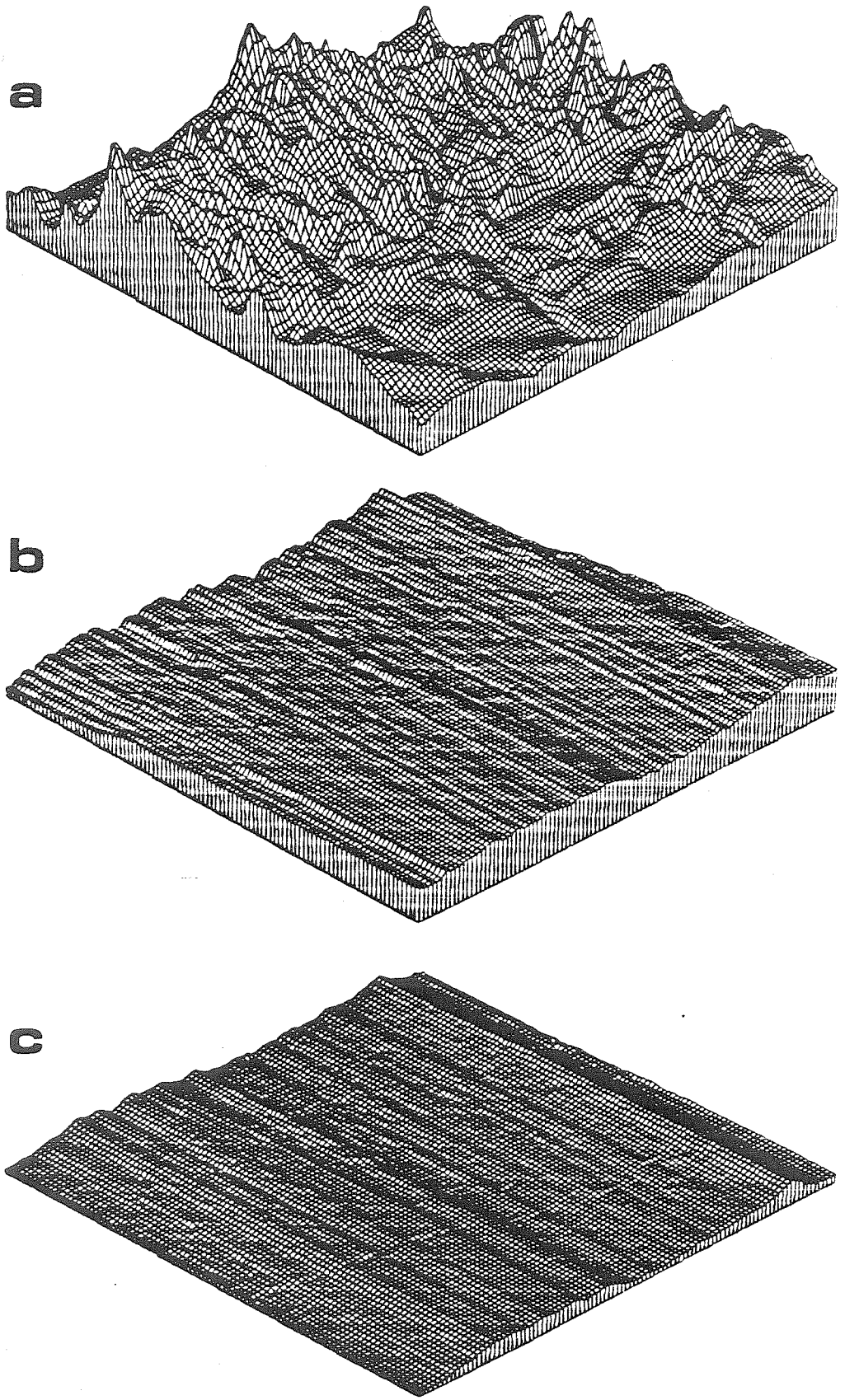

a
b
c

Figure 3.5. Histograms of topography and residuals . Note differing horizontal and vertical scales. (a) Distribution of elevations in exact topography. (b) Distribution of residuals to SSIPSF-PI estimate of topography. (c) Distribution of residuals to final estimate of topography. (d) Distribution of residuals within individual rows of the final estimate, with the mean of each row corrected.

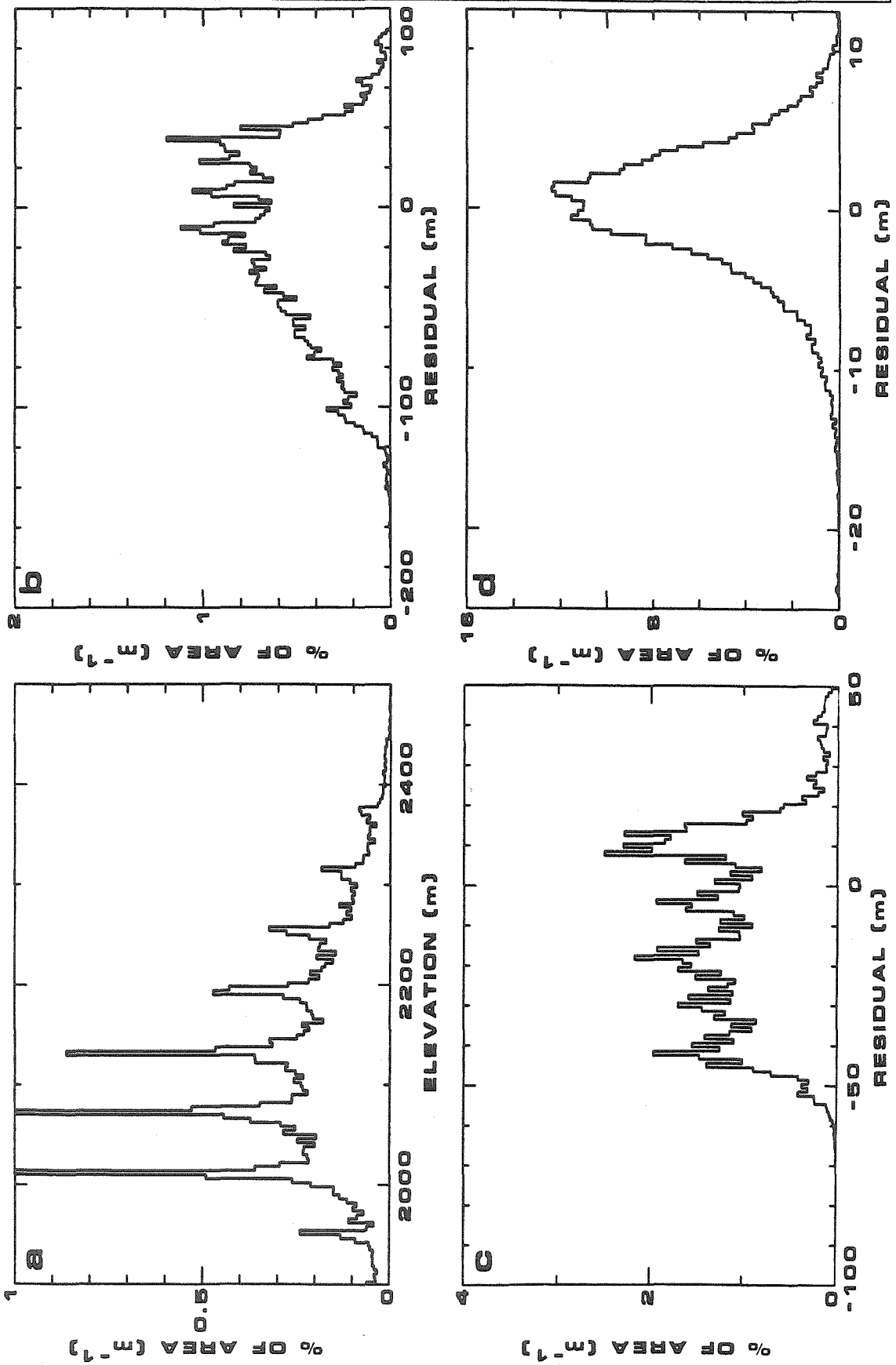

a
% OF AREA (m⁻¹)
0
0.5
1
2000
2200
2400
ELEVATION (m)
b
% OF AREA (m⁻¹)
0
1
2
-200
-100
0
100
RESIDUAL (m)
c
% OF AREA (m⁻¹)
0
2
4
-100
-50
0
50
RESIDUAL (m)
d
% OF AREA (m⁻¹)
0
8
16
-20
-10
0
10
RESIDUAL (m)

Figure 3.6. Power spectra of topography and residuals . One-dimensional power spectral densities estimated using Hanning-windowed periodograms. (a) Spectra of ensemble of rows: from top to bottom, exact topography, residual to SSIPSF-PI estimate, residual to final estimate. (b) As in part (a), but for ensemble of columns. Signal to noise ratio is lower because of row striping.

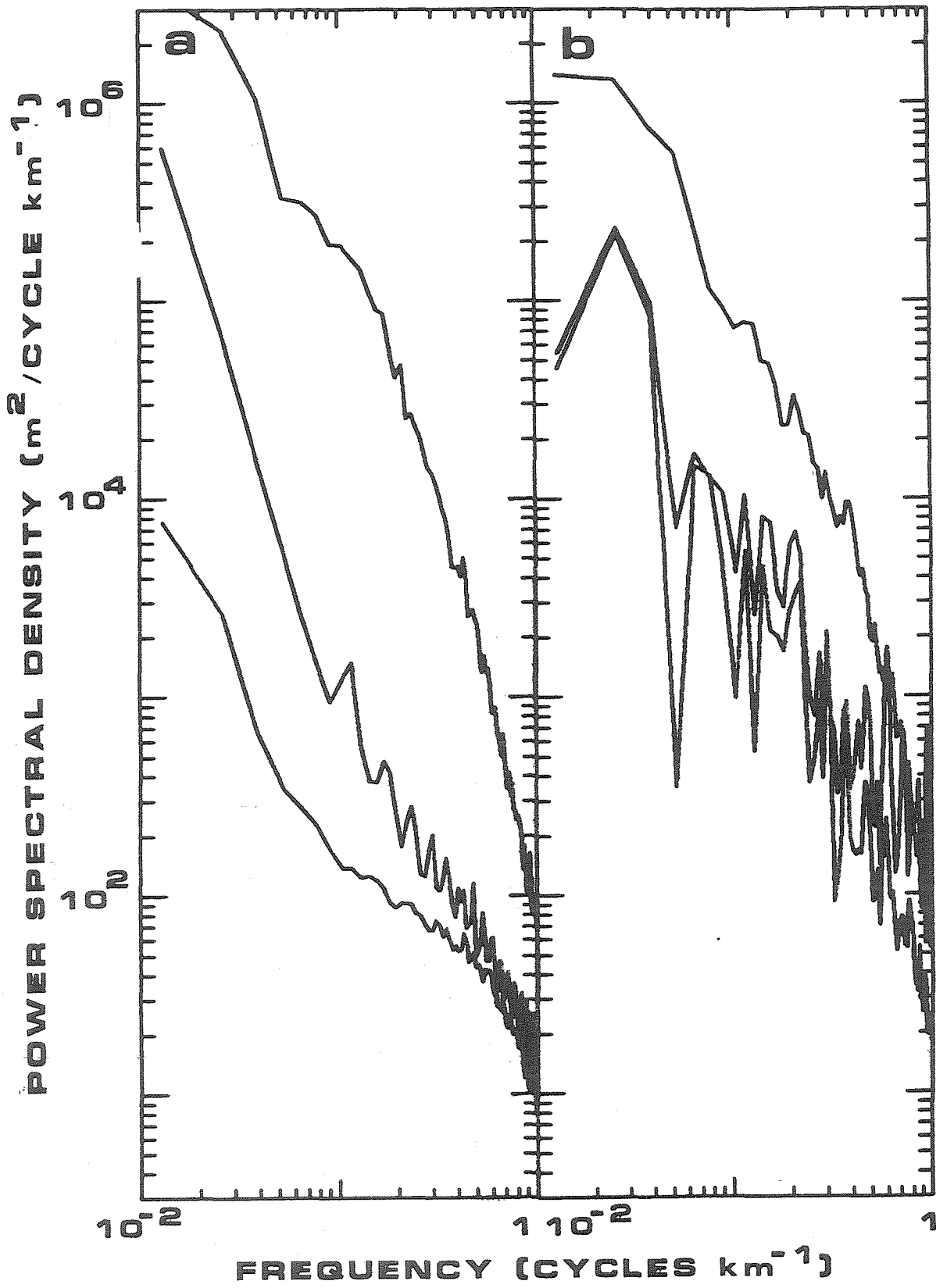

a
b
POWER SPECTRAL DENSITY (m2/CYCLE km-1)
10^6
10^4
10^2
10^-2
1 10^-2
1
FREQUENCY (CYCLES km-1)

Figure 3.7a. Error in PC equations versus iteration number . Nominal case with Lambert surface, $g = 45°$, $\alpha = 10^4$, area penalty. Solid curve is norm of gradient vector $\{E\}$ (cf. equation 2.9d). Size of dots indicates resolution of mesh (largest dots for full resolution). Dotted curve shows norm of brightness estimate error $\{B\} - \{b\}$.

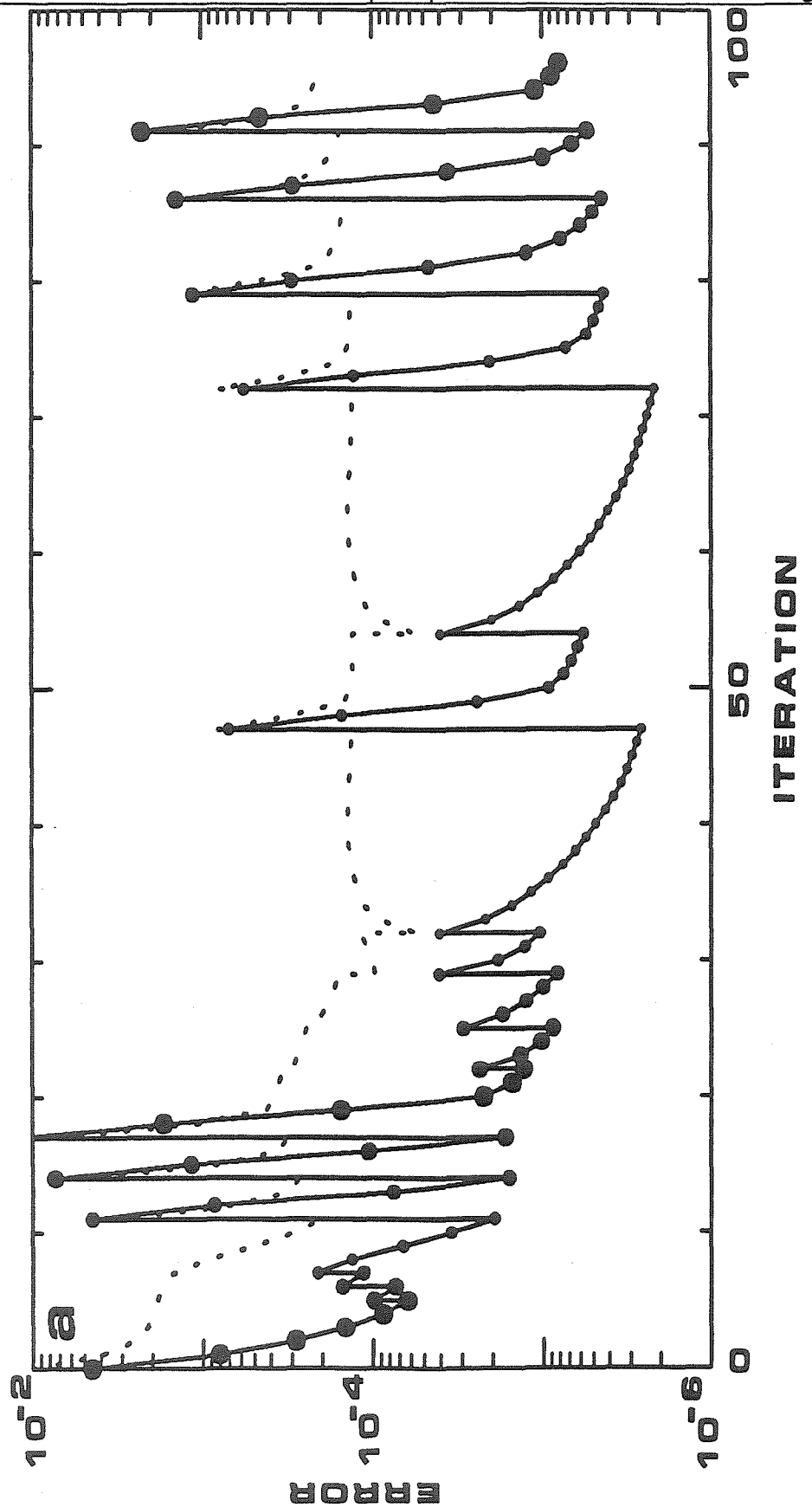
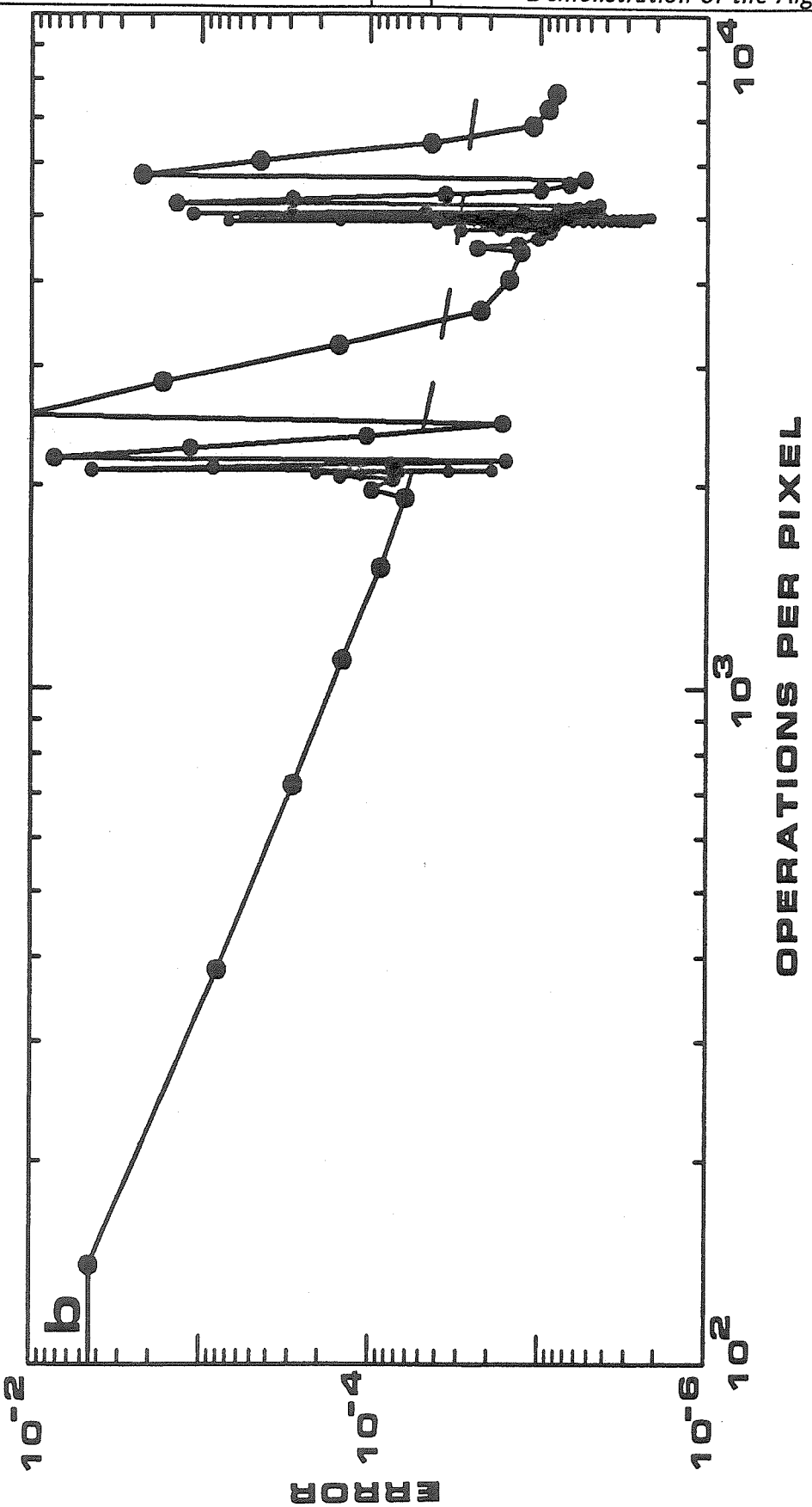

a
ERROR
10^{-2}
10^{-4}
10^{-6}
ITERATION
0
50
100

Figure 3.7b. Error in PC Equations versus computational effort . Solid curve is the same as in part (a), but plotted against the number of operations (additions and multiplications) per pixel. Dashed curve shows the result of 20 iterations without remeshing.


b
10^{-2}
10^{-4}
10^{-6}
ERROR
10^{2}
10^{3}
10^{4}
OPERATIONS PER PIXEL

Before we compare the performance of the algorithm for various other parameter choices, it is of interest to watch it at work on the nominal case. The solid curve in Figure 3.7a shows the norm of the gradient (cf. equation 2.9d), the quantity we are actually driving toward zero. The dotted curve shows the rms value of $\{B\} - \{b\}$ (normalized so $b_0 = 1$), which is what we would *like* to have vanish. Iteration on the different-size meshes is indicated by the size of the dots (largest for full resolution, smallest for one-sixteenth resolution). We clearly see the "diminishing returns" in decrease of the error per iteration that drives us to use the multigrid technique. Substantial error is introduced at each change of resolution, but it is local and is rapidly smoothed away. At first, the value at which the error levels off on succeeding meshes decreases rapidly, but when we compare iterations 54–72 to 33–48 it is clear that no further progress is being made. The formal stopping criteria were therefore abandoned at this point and the resolution was gradually increased to the full. The reason for this limit to the decrease in the error is our incorrect method of choosing $\{B\}$ when decreasing the resolution. We choose $\{B\}$ so that $\{B\} - \{b\}$ is injected, rather than the true right-hand side $\{E\}$. We are thus not *quite* solving the "corresponding" problem at all resolutions. Eventually, the error introduced by this approximation becomes larger than the improvement available from the coarser mesh, and multigridding is no longer helpful.

Figure 3.7a is potentially misleading, to the extent that the iterations on the coarse meshes are given equal prominence to those on the finer. Figure 3.7b shows the rms gradient as a function of the number of operations performed per pixel. (In contrast to the multiplication-counts for the solution of linear equations given above, the quantity plotted here includes all additions, multiplications, and higher operations done in preparing as well as solving the linearized PC equations at their correct relative importance in terms of time. It is thus an estimate of the actual computational

effort expended.) Clearly, very little time is spent working on the coarser meshes. For purposes of comparison, the dashed curve indicates the reduction in the rms gradient to be obtained for the same amount of effort *without* multigridding. The final rms gradient is about thrice as great (and the long-wavelength components are particularly large); the rms altitude residual is 35.7 m, 27.8 m within rows. It is also worth noting that the residual is reduced to nearly its "final" value after the first remeshing sequence: after $\sim$ 4500 operations per pixel it is only 5% larger than when iteration was terminated after $\sim$ 7300 operations per pixel (55% larger within rows).

The behavior of the PC algorithm for a variety of other cases is summarized in Table I in terms of the rms altitude residuals. Qualitatively, the residuals and the behavior under iteration are similar to the nominal case. I briefly summarize the results in the table. (The "final" errors given are somewhat subjective, since judgement must be exercised as to when iteration is no longer fruitful, but the amount of computational effort expended in each case was similar.) First and most generally, the error is always dominated by "stripes" along the row (phase plane) direction, and the minimum amplitude to which the stripes can be reduced by iteration is nearly invariant for a given choice of roughness criterion. The area criterion is nonetheless to be preferred, as it leads to slightly less severe striping, and is computationally more efficient than the entropy criterion as well. Evidently, the striping is an unavoidable characteristic of the "smoothest" solution compared to that which obeys the proper boundary conditions. It is *not* a consequence of slower convergence of the algorithm for between-rows variations than for variations within each row, as was reported in an earlier abstract (Kirk 1984). This conclusion is supported by tests of the algorithm on images generated from multiples of the Utah topography (with the residuals of course calculated with respect to the scaled elevations). The total rms error (*i.e.*, the striping) scales linearly with the amplitude of the topography; it is intrinsic to the

Table I. Performance of the Photoclinometry Algorithm

g (°)	α	Rougness Criterion	—RMS Error inAltitude Estimate (m)— —SSIPSF – PI[a]– Total	Rows[b]	——Final—— Total	Rows[b]
45	10^4	Area	43.7	37.5	22.3	4.44
45[c]	10^4	Area	43.7	37.5	35.7	27.8
45[d]	10^4	Area	14.8	9.59	11.3	1.50
45	10^4	Area	43.7	37.5	22.3	4.44
45[e]	10^4	Area	150.	143.	47.0	17.0
45	10^4	RMS z	43.7	37.5	34.2	26.8
45	10^4	Area	43.7	37.5	22.3	4.44
45	10^4	Entropy[f]	43.7	37.5	33.8	26.2
45	10^3	Area	43.7	37.5	22.6	7.65
45	10^4	Area	43.7	37.5	22.3	4.44
45	10^5	Area	43.7	37.5	22.9	4.09
30	10^4	Area	69.2	65.5	22.5	7.90
45	10^4	Area	43.7	37.5	22.3	4.44
45	10^4	Area	31.0	21.4	22.4	3.15
45	10^4	Area	24.4	9.74	22.4	2.30
45	10^4	Area	43.7	37.5	22.3	4.44
45[g]	10^4	Area	43.6	37.4	22.8	4.49
45[h]	10^4	Area	45.5	40.0	24.1	8.80
45[i]	10^4	Area	43.7	37.5	22.4	6.53
45[j]	10^4	Area	43.8	37.7	25.3	12.0

[a] Initial estimate for photometric function linearized about $z = 0$.
[b] Mean elevation of each row corrected to indicate errors in individual row profiles.
[c] Same computational effort, without multigridding.
[d] Amplitude of topography reduced to half its normal value before image generation.
[e] Amplitude of topography increased to twice its normal value.
[f] Entropy coefficient $c = 1 \times 10^3\,\mathrm{m}^{-1}$. (cf. equation 2.3c).
[g] Quantized at 1/128 of mean brightness.
[h] Quantized and Gaussian noise with $\sigma = 1/128$ of mean brightness added.
[i] Quantized and coherent noise with amplitude 1/128 of mean brightness, $\lambda = 3\,\mathrm{km}$ added.
[j] Image blurred by convolution with 3 by 3 boxcar.

solution toward which iteration is converging.

The error within rows is largely a consequence of imperfect convergence to the solution of the nonlinear PC equations. It should therefore vanish (relative to the total topography) in the limit of small amplitude and decreasing departure from linearity. As the table shows, the within-rows residual indeed scales nearly as the

square of the topography. It is also understandably more sensitive to the parameters of the algorithm than is the striping. Increasing α to 10^5 does not decrease the row residual substantially, but decreasing it to 10^3 increases it. The value $\alpha = 10^4$ was thus adopted for the remaining tests. Perhaps the most interesting quantity to vary is the phase angle (Figure 3.8). As usual, this has no effect on the final rms error, but both the error in the SSIPSF-PI estimate and the within-rows final error decrease with increasing g. This is in all probability due to the increasing image contrast: when the strength of brightness variation increases and the rms value of $\{B\} - \{b\}$ is held roughly constant, the error in $\{z\}$ will decrease. The rms variation in brightness across the shaded image (normalized to $b_0 \equiv 1$) is also shown in Figure 3.8, and follows the same $\cot g$ trend as the altitude residuals, supporting this explanation. The algorithm was also tried for $g = 15°$, but no solution was obtained because of unavoidable divergence of the SOR iteration. Divergence was detected to by examining both $\frac{|\Delta\Delta z^{k+1}|}{|\Delta\Delta z^k|}$ (the rate of decrease of the rms increment) and the ratio of the maximum nodal $\Delta\Delta z^e$ to the rms value. The latter was of use in detecting some cases in which subsequent examination showed that the solution diverged only in a localized region of the image.

Next, the sensitivity of the PC altitude estimate to noise, in particular quantization error, is examined. From Figure 3.8 we see that the rms width of the brightness distribution at $g = 45°$ is only $\sim 4.4\%$ of the mean brightness. The typical accuracy to which the PC solution leads to the observed brightnesses is $\sim 0.02\%$. Quantization is thus a potentially serious concern, since an eight-bit digital image exposed so that b_0 corresponds to half scale cannot resolve brightness differences $\lesssim 0.8\%$ of the mean. Nonetheless, as the table indicates, such an image *can* be successfully inverted and the rms residual is no larger than in the unquantized case! The only difficulty encountered is that the SSIPSF-PI solution contains noise at extremely high spatial frequencies,

Figure 3.8. Residual to topography versus phase angle . Triangles: rms residual to SSIPSF-PI estimate. Squares: residual to final estimate, dominated by stripes dictated by smoothness criterion. Circles: residual to final estimate, within rows, dependent on phase angle through image contrast. Open circles: inverse measure of image contrast, b_0/σ_B, where b_0 =brightness of mean plane, σ_B =standard deviation of brightness distribution. Dotted line is the fit $\sigma_B/b_0 = 0.0015 + 0.0437 \tan g$.

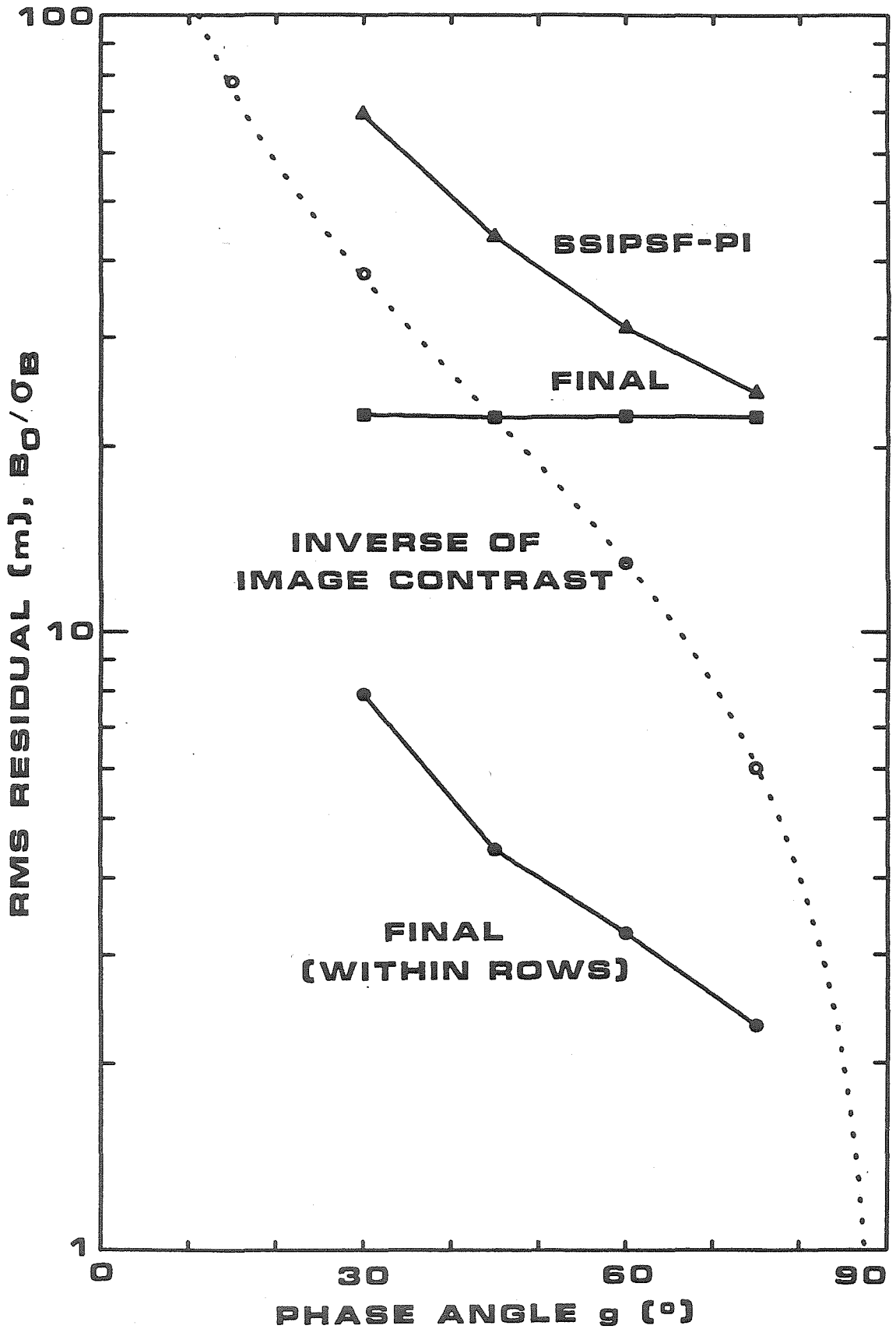

100
10
1
0
30
60
90
RMS RESIDUAL (m), B_0/σ_B
PHASE ANGLE g (°)
SSIPSF-PI
FINAL
INVERSE OF
IMAGE CONTRAST
FINAL
(WITHIN ROWS)

Figure 3.9a, b. *Spriggina* image and greyscale topography . Fossil *Spriggina* from the Ediacaran Pound Quartzite, South Australia. Approximate scale is indicated.. (a) Portion of a photomacrograph of an MgO coated latex replica of the fossil, obtained with the PFUEI charge-coupled device camera. (b) Greyscale representation of PC reconstructed topography from the image in (a).

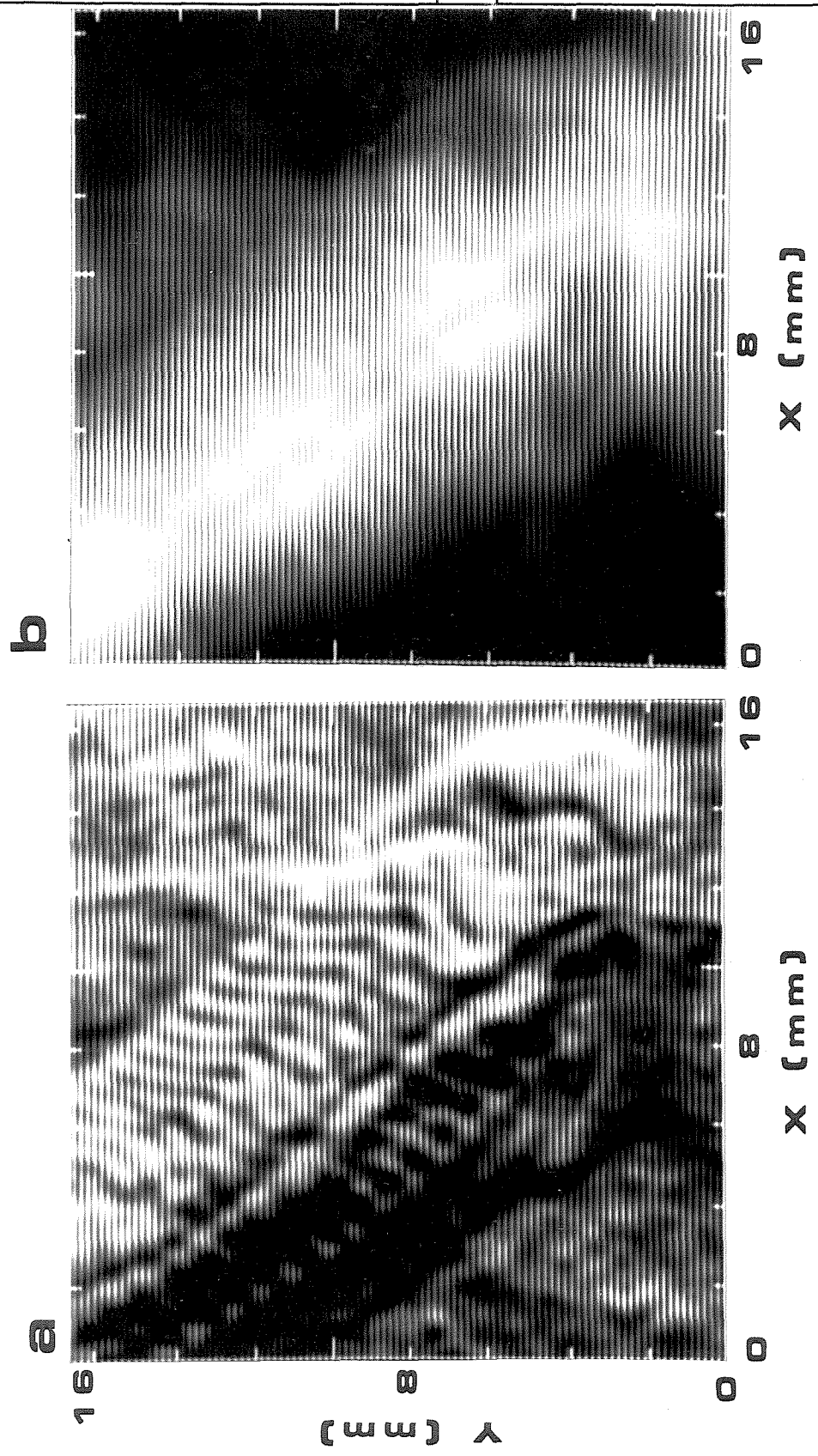
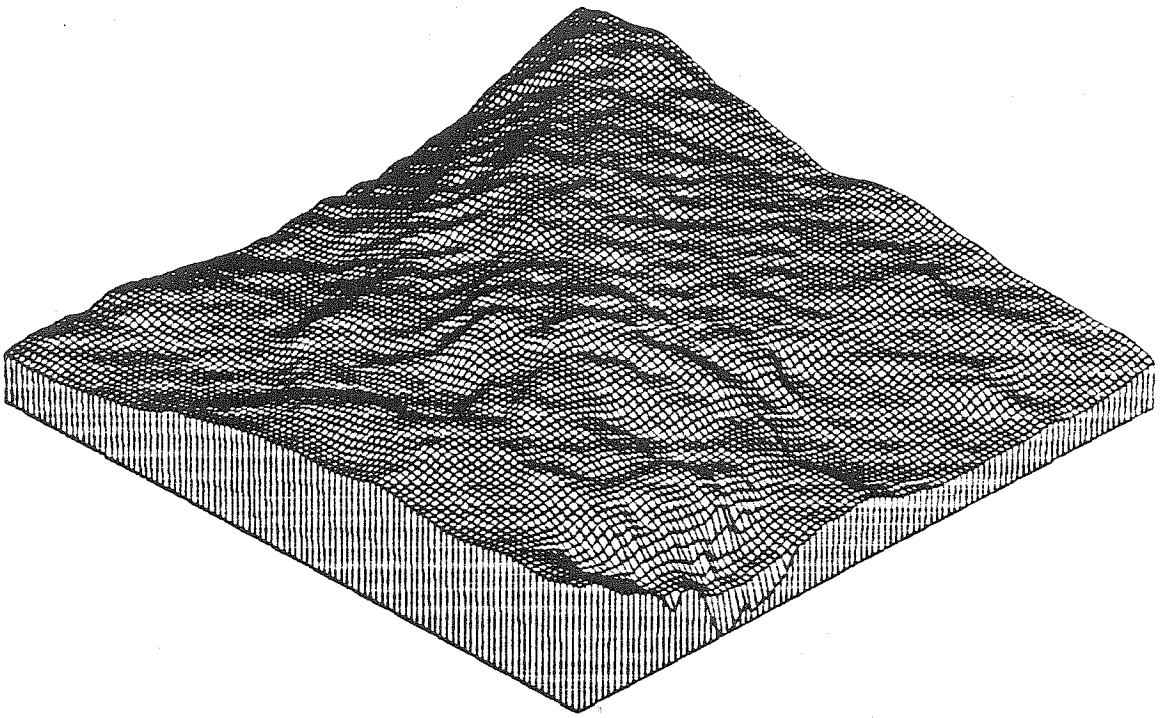

b
a
0
8
16
X [mm]
Y [mm]

Figure 3.9c. Perspective plot of *Spriggina* topography . View is from the lower right at an elevation of 45°, with a vertical exaggeration of 2 : 1.



which can cause SOR to diverge. Smoothing the SSIPSF-PI estimate with one pass of a 3 by 3 boxcar filter before iteration prevents this problem. When Gaussian noise with a standard deviation of $b_0/128$ is added to the image before quantization, two passes of the boxcar are necessary, but iteration is straightforward thereafter.

Contamination of the image with coherent noise results in coherent noise in the solution but does not make iteration more difficult. Unless the noise wavevector is almost perpendicular to the phase plane, the result is roughly the addition of the corresponding coherent noise (*i.e.*, the sinusoidal surface whose image would be the noise in the data) to the reconstructed topography. In this case one can show that the relation between the amplitude Δb of noise in the brightness and the amplitude Δz of the elevation error is approximately:

$$\Delta z \simeq \frac{\lambda}{2\pi} \frac{\sec\varphi \cot g \Delta b}{\sqrt{1-(\sec\varphi \cot g \Delta b)^2}}, \tag{3.2}$$

where λ is the wavelength and φ is the angle from the phase plane to the wavevector (the direction of maximum slope) of the noise. The case indicated in the table, for a (quantized) image plus noise with $\Delta b = b_0/128$, $\lambda = 10$ pixels (3 km), and $\varphi = 45°$, fulfills this expectation. Clearly, coherent noise is a serious potential concern, since it is common in digital images and even moderate amounts may lead to large errors in z if λ or $\sec\varphi$ are large. Coherent noise is often extremely difficult to eradicate from an image (if it is not exactly harmonic), but fortunately is relatively easy to detect.

As a last test case, an unquantized test image was degraded by convolution with a 3 by 3 boxcar before PC inversion. The final rms residual increases only slightly, but the within-rows residual is trebled (even with respect to a smoothed version of the reference topography). The severity of the error due to blur will in general depend not only on the degree of blurring but on the spatial-frequency spectrum of the topography.

I conclude this section with an application of the PC algorithm to a real image.

Although the metric accuracy of the result cannot be ascertained, it is nonetheless of interest because it demonstrates the ability of the algorithm to function on a real dataset with unknown normalization, substantial noise, and much larger slopes than the Utah topography. It also shows that PC is useful for photomacrography as well as for remote sensing.

Figure 3.9a shows the fossil organism *Spriggina*, from the Pound Quartzite of South Australia. The figure is a 200 by 200 pixel subarray extracted from an 800 by 800 pixel digital image obtained with the PFUEI charge-coupled-device camera at Caltech. To create a surface with a known photometric function, a high-resolution latex replica of the fossil was exposed to MgO smoke until uniformly coated; this preparation is, to excellent approximation, a Lambert scatterer. A bare filament desk lamp bulb some four meters from the replica was used to provide nearly point-source illumination at a phase angle of 45°. The dark current value (based on a masked region) was subtracted from the raw image data, and the result ratioed to an image of a blank field.

The resulting flat-fielded image was convolved three times with a 3 by 3 boxcar filter to reduce the highly visible noise, and alternate samples extracted from a 400 by 400 region to reduce the dataset to tractable size. The PC algorithm was applied to the resulting image, using the area roughness criterion and a penalty number of 10^4 as usual. The method used to normalize the image, based on the estimated orientation of the mean plane, is described in Appendix A. As with the artificially noisy pseudoimages, there was a tendency for the SOR iteration to become divergent at the highest resolutions and in the earliest linearizations. Progress could nonetheless be made by discarding $\{\Delta z\}$ whenever divergence occurred, smoothing $\{z\}$ with a 3 by 3 boxcar, and then relinearizing.

Figure 3.9b is a greyscale representation of the PC altitude estimate obtained

after $\sim$ 6800 operations per pixel, and Figure 3.9c shows the same data in perspective plot form, with a vertical exaggeration of 2 : 1. The $\sim$ 0.9 mm relief obtained for the fossil is in good agreement with direct inspection of the replica.

4. Discussion

The previous section indicates the power of the finite element PC algorithm. At an expenditure of $\lesssim 10^4$ operations per pixel, it can yield two-dimensional topographic information about surfaces of realistic complexity to an accuracy of $\sim$ 20% (rms error over rms topography) even in the presence of substantial noise. The error within profiles aligned with the direction of illumination is much lower: $\sim$ 2–7%, depending on the phase angle. This within-rows error may, in fact, be of interest in real problems. If it can be determined that the image contains a level or near-level region, then adjustments of the profiles made so that they match there can be extended to the rest of the dataset. (This is a kind of *ex post facto* application of a boundary condition to the topography; given the nonlinearity of the problem, it cannot be expected to yield a completely accurate result, but it can help substantially.) Moreover, for some purposes, such as the estimation of the fractal dimension of a fracture surface (Mandelbrot *et al.* 1984), the desired product may be an ensemble of profiles rather than a two-dimensional surface. This information could be obtained much more rapidly and for many more profiles by photoclinometry than by cutting and measuring the sample.

On the basis of the tests carried out, it is unlikely that the two-dimensional PC algorithm could be completely automated. Judgement on the part of the operator was required to decide when iteration at reduced resolution was no longer fruitful; strict application of the stopping criteria of Brandt (1976) would lead the algorithm into an endless loop between the lowest and next-lowest resolution. As argued above, this failure of the stopping criteria (which *have* been automated for other problems)

is almost certainly a consequence of the simplified but incorrect scheme used to inject the brightness measurements onto the coarser mesh. Additionally, the operator must be alert for divergence of SOR iteration and either attempt to treat it by smoothing away the erroneous high-frequency components of the topography, or decide to abandon the inversion. No foolproof test to distinguish treatable from fatal divergence was discovered. The requirement for supervision should not be seen as a critical shortcoming of the method, however. Both photogrammetry and existing photoclinometric methods (*e.g.*, Davis and Soderblom 1982; 1984; Howard *et al.* 1982) also require supervision, and they produce only separated tiepoints and one-dimensional profiles, respectively, rather than topography over the entire image field.

Application of any photoclinometric method to real remote-sensing data will be subject to several potential problems that are not present in the analysis of pseudoimages and laboratory-scale data. I will briefly discuss some of these and in some cases will suggest remedies. First, we may not know the proper photometric function to use. I argue that knowledge of the albedo, along with knowledge of the image system calibration, is not strictly necessary for successful PC. The self-normalization algorithm presented in Appendix A has been shown to estimate the overall normalization very well, by requiring that the reconstructed surface have the correct average orientation. Variation in the *form* of the photometric function is potentially more serious. Howard *et al.* (1982) have investigated the Minnaert photometric function (3.1) in some detail, finding that misestimation of the overall brightness normalization leads to large errors in estimated slope, but that variation of the exponent k has a much smaller effect. Davis and Soderblom (1984) find that a Lommel-Seeliger Lambert photometric function is less sensitive to misestimation of its photometric parameter than the Minnaert function. Recent work by Wilson *et al.* (1985) suggests that for Hapke's (1984) photometric function, the only parameter with a significant

effect is the rms slope (at unresolved lengthscales) T. Other forms of the photometric function have yet to be examined. For all these photometric functions there is the hope that spatially averaged observations at a variety of geometries will suffice to constrain the parameters needed for interpretation of the resolved brightness variations.

A related and potentially much more serious concern is the effect of atmospheric scattering. This may be very complex, but heuristically the two main effects will be attenuation of the light transmitted to and then from the surface, and scattering from the atmosphere into the camera lens. We thus have a more-or-less linear transformation of the brightness $b' \simeq be^{-\tau(\sec\theta+\sec\phi)} + b_{atm}$, where τ is the normal optical depth and b_{atm} is the brightness (appropriately normalized) of the light scattered from the atmosphere. Clearly, if either quantity is large, disaster may ensue. To the extent that the photometric function is linearizable, the offset leads to a net tilt of the recovered surface topography towards the light source, and the attenuation to a reduction of the amplitude calculated. If we use the self-normalization technique, the entire expression will be further multiplied by a factor less than unity, so that the net tilt is removed but the topography is even more de-emphasized. Of course, there will also be nonlinear effects leading to a distortion of the recovered surface.

The only viable solutions to the atmospheric scattering problem are: restrict one's attention to planets without atmospheres, or attempt to model the atmospheric effects and hence remove them. Such modeling can vary enormously in sophistication, from locating a "black" region in the image (a shadow or low-reflectivity surface such as a body of water) and identifying its brightness as b_{atm}, to a full radiative-transfer calculation at each point in the image, taking into account multiple scattering, diffuse illumination of the surface by the sky, and so on. In the absence of additional information, all such models require the enabling assumption that the atmospheric

properties are uniform (so that scattering depends on position at most in a known way, through the varying illumination geometry). Ken Herkenhoff (personal communication, 1986) is currently working on a radiative-transfer model of the atmosphere in the Martian polar regions near the terminator. In the near future this model will be used to conduct a photoclinometric investigation of the topography of the polar layered terrain. Where possible, comparison will be made with photogrammetric data to assess the accuracy of the result.

The most formidable obstacle to practical photoclinometry is almost certainly the possibility of spatially varying surface photometric properties, in particular the albedo. Again, we have several choices: give up and apply PC only to bodies certified to be bland (if such can be found), use more than one image and solve for both the albedos and the altitudes, or attempt to model the albedo variations. The second technique goes by the rather unfortunate name of "photometric stereo." A finite element algorithm for photometric stereo could be constructed as a relatively straightforward generalization of the PC code. It should be emphasized that, despite its requiring two images, photometric stereo may still have substantial advantages over genuine photogrammetric stereo. As mentioned before, stereo yields only point measurements of the elevation, and these may be few if the images are bland and hard to cross-correlate. Photometric stereo would yield a two-dimensional topographic dataset *plus* a map of the surface albedo. Also, overlapping images may be easier to obtain with different illumination geometries than with different observation geometries, especially from sun-synchronous satellites. Although the look angle and time of day at which a region is imaged are fixed, the azimuth of solar illumination will vary with the seasons (R. J. P. Lyon, personal communication, 1985). Comparison of images made at different times of the year could give a topographic solution of better quality than could be obtained for a uniform surface from one image. Each

image constrains the surface in the cross-phase-plane direction to which the other is least sensitive. Another interesting prospect for photometric stereo is to use a third image (or assume the albedo distribution is known) to obtain an *overdetermined* set of equations for the elevations. The artificial roughness criterion would not be needed, and data of very high accuracy might be produced this way.

Short of writing a new program to do photometric stereo, there are several possible ways of attempting to correct the image to what it would have been for a uniform albedo surface, *before* applying the existing PC algorithm. The assumption of a uniform surface within the PC program need not be changed. (Were the form of the photometric function to vary spatially to a significant extent, this precorrection technique would not be feasible.)

The simplest way to attempt correction of the brightnesses, which should nonetheless be taken seriously, is *ad libitum* by a photogeologist working interactively with the PC program. The human would examine both the image and the previous PC altitude estimates and attempt to identify geologically plausible albedo units, adjusting their relative albedos to produce the most reasonable revised altitudes. As a concrete example, a systematic slope in crater floors away from the light source could be corrected by positing that the craters are filled with a low-albedo unit and rescaling the input brightnesses upward. This technique has been used with some success by Mouginis-Mark and Wilson (1981) for Mercury. Automation of the heuristics for albedo-unit selection is a more speculative possibility (Bruce Murray, personal communication, 1986).

In some cases, it may be possible to extract the needed albedo information from a multispectral dataset. Eliason *et al.* (1981) report the separation of Landsat data into spectral-albedo images normalized to a level surface and a "topographic modulation" image, which is precisely the kind of albedo normalized image we require.

Their method involved the examination of statistics for the ratios of the images in different spectral bands. A finite number of clusters of pixels with similar ratios were identified as discrete surface units, and then each pixel was assumed to be paved with the unit it most closely resembled. To the extent that the photometric function can be linearized, the mean brightness of all the pixels in any given unit (for each spectral band) is an estimate of the brightness of a level surface covered with that unit. Then the level-surface spectral albedo maps may be constructed by replacing the actual brightnesses of each pixel with the level-surface (mean) brightnesses of the unit to which it belongs. The topographic modulation is given by the ratio of the pixel's actual brightness to that for a level surface.

A more sophisticated variant of this method due to Greg Ojakangas (unpublished) allows each pixel to contain an unresolved mixture of several units. In his method, the statistics of band ratios are examined and a set of *end member* units identified such that the color ratios of every pixel can be expressed as an area-weighted average of the color ratios of the end members. (All pixels must have colors lying between those of the end members, in order that none of the areas be negative.) Each pixel is then considered in turn. Given n spectral bands with distinct properties, there are $n+1$ equations of constraint: n brightnesses, plus the requirement that the partial areas of all end members in a given pixel sum to the total pixel area. We can therefore invert the system of linear equations to find the relative proportions of n different end member units, plus the topographic modulation by which their level-surface brightnesses must be multiplied. The key difficulty here is the selection of appropriate end members. For a meaningful solution to be obtained, the number of end members must not exceed the number of spectral channels in the imaging sytem (and may have to be less, if some of the spectral bands are redundant). At the same time, however, the set of end members must adequately describe the full range of

surface colors encountered, or nonphysical negative areas will result. It is thus advisable to break up large images into a number of relatively small regions, with only a few distinct surface units in each. Despite these caveats, when a suitable set of end members *can* be chosen, the linear inverse method produces impressive results.

Finally, the inclusion of *a priori* topographic information in the PC algorithm may turn out to be extremely helpful, particularly in modeling out atmospheric effects and (low spatial frequency) albedo variations. Because the PC problem has been formulated as the constrained minimization of the roughness function, the addition of further constraints is trivial. Furthermore, in the penalty method formulation it is straightforward weight each constraint according to its importance. The most useful constraints would undoubtedly be *benchmarks*: points where the elevation is known as a result of sterometry, radar altimetry, or whatever means. "By eye" estimates of the local topographic strike in some elements could also be incorporated. Appendix A includes a discussion of one way in which these constraints could be implemented in the finite element algorithm. Here I consider their potential utility.

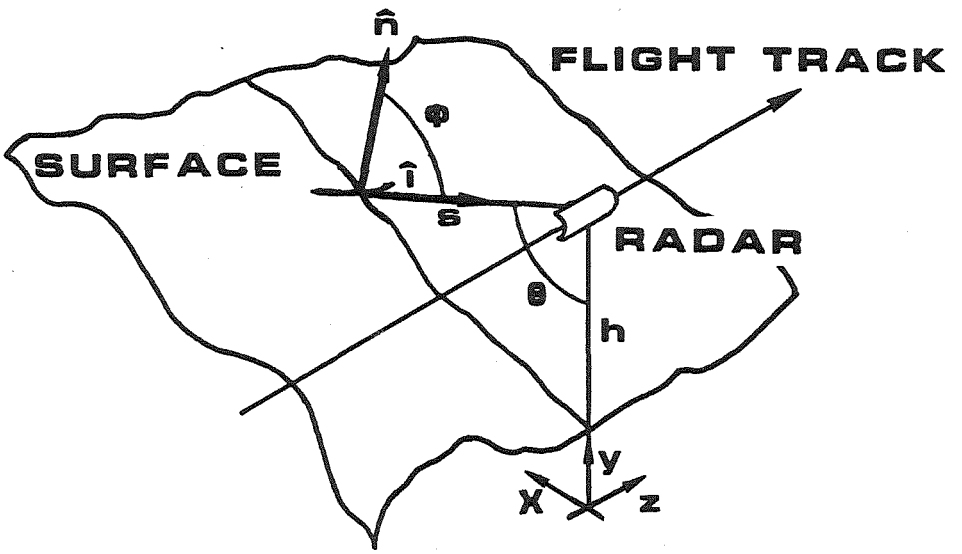
A priori information would, by itself, contribute to the accuracy of the PC algorithm, but its utility might be limited by the resolution at which it was available. As an extreme case, a closely spaced set of benchmarks (such as an altimetry track) spanning the image could be used to suppress the phase plane striping effect. A smaller number of data points would be correspondingly less effective. If one is interested in modeling atmospheric or albedo effects, on the other hand, even a small number of benchmarks could be very useful. Consider the simple atmospheric scattering model $b' \simeq be^{-\tau(\sec\theta+\sec\phi)} + b_{atm}$ discussed above. The PC algorithm may readily be extended to adjust the parameters τ and b_{atm} along with the nodal altitudes, but in the absence of other constraints the results would be disastrous. The roughness would be minimized by an unphysical solution with $\{z\} \to 0$, $\tau \to -\infty$. This cor-

responds to a nearly flat surface whose miniscule brightness variations are amplified to the observed values by the atmosphere! Introduction of two or more benchmarks (only the differences between altitudes are significant in PC) in this situation provides information about the actual amplitude of the topography, and should enable a meaningful solution. A spatially varying atmosphere model may likewise be constrained, provided benchmarks are available with a fine enough spacing and a broad enough coverage. Similar considerations hold for models of spatially varying albedo (though the albedo is less likely than the atmospere to be fit by a slowly varying function). In the absence of *a priori* constraints, minimizing the rougness will drive as much of the brightness variation as possible into the albedo. Addition of benchmarks would help ensure the correct scale of topography.

I conclude this paper with a brief description of the generalization of the two-dimensional finite element method to the analysis of topography from side-looking radar (SLAR) images, known as radarclinometry[6]. This name was proposed by Wildey (1984) to emphasize that, like photoclinometry, the method derives shape information from satisfying a set of "brightness" measurements, not from the intrinsic range-finding property of radar. Radarclinometry (henceforth RC) is of interest for a number of reasons. First, all SLAR images are subject to geometric distortion due to the way they are formed (explained below) which makes geologic interpretation difficult. Thus, a technique that takes a single SLAR image and provides even an approximate rectification of the geometry is of practical use. Second, RC may be almost the *only* means of obtaining topographic information from proposed missions to Venus and Titan. The Magellan spacecraft to Venus will carry a synthetic aperture radar capable of high resolution, but not a radar altimeter. Although overlapping radar images can be combined to yield stereo information, identifying corresponding

[6] A radarclinometry algorithm could also be applied to side-looking sonar without further modification, though its utility in this context is somewhat dubious.

Figure 4.1. Coordinate system for radarclinometry . Cylindrical coordinates are centered on the flight track of the radar, with the *azimuth* z along the track, *slant range* s measured obliquely to a point on the surface, and *elevation angle* θ up from the nadir. A cartesian coordinate system with z axis parallel to that above is centered in the mean plane a distance h below the flight track. The task of two-dimensional radarclinometry is to find the surface $\theta(s,z)$ whose normal vector $\hat{n}$ is consistent with the observed radar brightness at every point (s,z), then to transform this description of the surface to one in terms of $y(x,z)$.


n̂
FLIGHT TRACK
Φ
SURFACE
î
s
RADAR
θ
h
y
X
z

surface features is if anything more difficult than for visible images (Wildey 1986*b*). Furthermore, the baseline Magellan mission allows only for the surface of Venus to be imaged once. RC is thus an attractive way of obtaining topographic data for Venus at much higher resolution than did Pioneer Venus. It is even more attractive for use with the proposed Titan Imaging Radar on the Cassini mission. Although Cassini will orbit Saturn, it will encounter Titan only in a finite number of rapid fly-bys. Opportunities for stereo SLAR imaging will thus be extremely limited, and a radar altimiter (if one is included) will yield only a series of one-dimensional profiles of the body.

SLAR differs radically from imaging at visible wavelengths in its geometry (Figure 4.1), a difference that affects most aspects of radarclinometry. The method is *monostatic*, *i.e.*, the illumination comes from the location of the detector. There is thus only one ray vector $\hat{\imath}$, and one photometric angle $\phi = \hat{n} \cdot \hat{\imath}$, where $\hat{n}$ is the unit normal to the surface. The photometric function $b(\phi)$ is thus somewhat simplified. More importantly, although a combination of doppler and time information is used to determine the *azimuth* coordinate z of features measured along the flight track, location in the transverse direction is determined from the delay required for return of an echo. The corresponding coordinate is thus the *slant range* s. Thus the better a region of the surface approximates a wavefront of constant s, the more compressed its echo will be in time, and hence the more compressed it will appear on the SLAR image. This is the characteristic distortion of SLAR, with slopes toward the radar foreshortened, and those facing away stretched out. If we can determine the cartesian (x, y, z) coordinates of each point (s, z) in the image, we can make both a topographic dataset $y(x, z)$ and a geometrically corrected image $b(x, z)$.

Despite his success with the PC problem, the main worker in the field, Wildey (1984, 1986*a*), abandoned the two-dimensional approach for radarclinometry in favor

of the older technique of integration along one-dimensional profiles. He requires that the user determine the direction of strike at the first pixel and then propagates this quantity by making an assumption about the second derivatives of the surface known as "local cylindricity." The path integral approach would appear to be necessitated by the fact that at each pixel we must iterate the values of *both* x and y in order to satisfy the brightness constraint on the orientation at fixed slant range s. The surface $y(x,z)$ is thus defined implicitly, rather than explicitly, by the RC equations.

I show in Appendix B how this apparent difficulty may be removed, and the two-dimensional nature of the problem exploited as in photoclinometry. As in photoclinometry, boundary conditions are formally required but not available, so a roughness criterion must be used. The trick is to work in the cylindrical coordinate system imposed by the radar process, solving for a surface described in the form $\theta(s,z)$, where θ is the *elevation angle.* Global solution is possible because s appears only explicitly in the equations. Once the surface is known in cylindrical coordinates, it is straightforward to transform to the desired cartesian coordinates. The only drawbacks to this formulation are that the descriptions of the surface orientation $\hat{n}$ and its derivatives are quite complex, that the area roughness function cannot be imposed precisely, and that the finite element approximation to $\theta(s,z)$ does not represent a level surface well when the mesh is coarse. Other potential problems are the presence of noise (coherent and speckle) in synthetic aperture radar images, and the need for at least relative radiometric calibration between pixels of an image. None of these should prove an insurmountable obstacle to the implementation of two-dimensional RC in the near future.

Acknowledgements

I thank J. Kirschvink for stimulating my initial interest in photoclinometry and for many valuable discussions since then. I also thank J. Sepkoski for his fortitude

in presenting early PC results for which he should not have been held responsible, L. Soderblom for providing the digital topographic dataset, G. Garneau for help at IPL, B. Runnegar for the latex *Spriggina* replica, and especially J. Westphal for use of the PFUEI CCD camera and associated computer facilities.

Appendix A: Explicit Forms of the Photometric Function and Roughness

I derive the explicit forms of the Minnaert photometric function (3.1), and the contribution of an element to each of the three roughness criteria (2.3), in terms of the nodal altitudes. For convenience, I restate the continuous versions of the functions here:

$$b = F(g)(\cos\theta)^{k(g)}(\cos\phi)^{k(g)-1} \qquad (A.1a)$$

$$= F(g)(\hat{n}\cdot\hat{\imath})^{k(g)}(\hat{n}\cdot\hat{e})^{k(g)-1}, \qquad (A.1b)$$

where the unit vectors $\hat{n}$ normal to the surface, $\hat{\imath}$ toward the source of illumination, and $\hat{e}$ toward the observer, along with the angles ϕ, θ, and g are defined in Figure 2.1, and

$$S = \begin{cases} \iint (z-Z)^2\,dx\,dy, & \text{the rms altitude} \quad (A.2a) \\ \iint \sqrt{1+\left(\dfrac{\partial(z-Z)}{\partial x}\right)^2+\left(\dfrac{\partial(z-Z)}{\partial y}\right)^2}\,dx\,dy, & \text{the area, or} \quad (A.2b) \\ \iint -(c(z-Z)+e^{-1})\ln(c(z-Z)+e^{-1})\,dx\,dy, & \text{the entropy.} \quad (A.2c) \end{cases}$$

where Z is a reference surface for the topography z. Adopting the local numbering of the nodes in an element e shown in Figure A.1, we seek $b^e(z_1^e, z_2^e, z_3^e, z_4^e)$ such that the brightness constraints can be written $b^e = B^e$. The area criterion of roughness (A.2b) involves the derivatives of z, so for it we also seek a division according to elements $S^e(z_1^e, z_2^e, z_3^e, z_4^e)$ such that $S = \sum_e S^e$. From the explicit representations, we can calculate the first derivatives of b^e and the first and second derivatives of S^e.

These quantities may then be *assembled* (summed) into the locations in the matrix equation (2.9) corresponding to the global numbering of the nodes.

The rms-altitude and entropy roughness criteria do not involve the derivatives of z, so it is more convenient to write them as a sum over (globally numbered) nodes $S = \sum_i S_i(z_i)$; clearly, the functional form S_i is exactly that given above. This is a departure from the strict finite element formulation, but it is justified by the simplification to which it leads and by the fact that the method is not sensitive to the exact nature of S.

Having chosen a one-point Gauss quadrature scheme (cf. equation 2.4), we approximate the integrals over the element implicit in b^e and S^e by the value of the integrand at the center. Defining:

$$\Delta_1 = z_3^e - z_1^e, \quad \Delta_2 = z_2^e - z_4^e, \tag{A.3}$$

and analogously,

$$\Delta_1' = (z_3^e - Z_3^e) - (z_1^e - Z_1^e), \quad \Delta_2' = (z_2^e - Z_2^e) - (z_4^e - Z_4^e), \tag{A.4}$$

we write the derivatives at the central point given by bilinear interpolation. It is convenient to work in the rotated (x_1, x_2, z) coordinate system defined in Figure A.1. Then

$$\frac{\partial z}{\partial x_1} = \frac{-\Delta_1}{\sqrt{2}}, \quad \frac{\partial z}{\partial x_2} = \frac{\Delta_2}{\sqrt{2}} \tag{A.5}$$

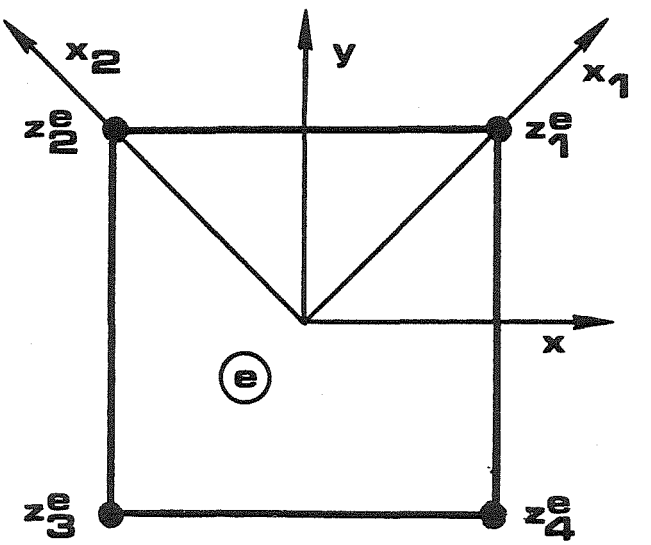
and

$$\frac{\partial (z - Z)}{\partial x_1} = \frac{-\Delta_1'}{\sqrt{2}}, \quad \frac{\partial (z - Z)}{\partial x_2} = \frac{\Delta_2'}{\sqrt{2}} \tag{A.6}$$

Realizing that $\left(\frac{\partial (z-Z)}{\partial x_1}\right)^2 + \left(\frac{\partial (z-Z)}{\partial x_2}\right)^2 = \left(\frac{\partial (z-Z)}{\partial x}\right)^2 + \left(\frac{\partial (z-Z)}{\partial y}\right)^2$ is an invariant, we can immediately write the element contribution to (A.2b):

$$S^e = \frac{1}{\sqrt{2}} \sqrt{2 + {\Delta_1'}^2 + {\Delta_2'}^2}. \tag{A.7}$$

Figure A.1. Local node numbering convention . For the purpose of calculating element contributions to S and b, it is convenient to renumber the four nodes in a given element as shown. Rotated coordinate system (x_1, x_2, z) simplifies the form of the functions.


x2
y
x1
z2e
z1e
x
e
z3e
z4e

To eliminate unnecessary arithmetic in the PC algorithm, the factor of $\frac{1}{\sqrt{2}}$ was omitted from the definition of S^e. The derivatives of S^e with respect to the nodal altitudes may be obtained by the chain rule from the two first derivatives and three second derivatives with respect to the Δ_i'.

This leaves the brightness. Defining unit vectors $\hat{x}_1$, $\hat{x}_2$, $\hat{z}$, we obtain for the unit normal to the surface

$$\begin{aligned}\hat{n} &= \frac{-(\partial z/\partial x_1)\hat{x}_1 - (\partial z/\partial x_2)\hat{x}_2 + \hat{z}}{\sqrt{(\partial z/\partial x_1)^2 + (\partial z/\partial x_2)^2 + 1}} \\ &= \frac{\Delta_1\hat{x} - \Delta_2\hat{x}_2 + \sqrt{2}}{\sqrt{\Delta_1^2 + \Delta_2^2 + 2}}.\end{aligned} \tag{A.8}$$

Since we observe along the z axis, and the light source is in the (x, z) plane at a phase angle g, we have $\hat{e} = \hat{z}$ and $\hat{\imath} = (\sin g\hat{x} + \cos g\hat{z}) = (\frac{1}{\sqrt{2}}\sin g(\hat{x}_1 - \hat{x}_2) + \cos g\hat{z})$. The photometric function may now be written as:

$$b^e = F(g)\left(\frac{1}{\sqrt{2}}\frac{2\cos g + (\Delta_1 + \Delta_2)\sin g}{\sqrt{2 + \Delta_1^2 + \Delta_2^2}}\right)^k \left(\frac{\sqrt{2}}{\sqrt{2 + \Delta_1^2 + \Delta_2^2}}\right)^{k-1}.$$

Again, the work done in the PC algorithm was reduced by defining b^e to be this quantity, divided by the constant factor $\frac{F(g)}{\sqrt{2}}$. The observed brightnesses were of course rescaled in advance by the same amount. When the image is photometrically calibrated and one knows the surface albedo (*e.g.*, for a pseudoimage), this rescaling is straightforward. Tests showed that it could also be estimated accurately by a simple iterative technique exploiting the relatively weak dependence of b on the transverse gradient $\frac{\partial z}{\partial y}$. We make the assumption that, for the purpose of determining the normalization β such that $B^e = \beta\overline{B^e}$, where the $\overline{B^e}$ are the raw data, we can approximate $\frac{\partial z}{\partial y} \simeq \frac{\partial Z}{\partial y}$ within the photometric function. The equation $b^e(\frac{\partial z}{\partial x}, \frac{\partial Z}{\partial y}) = \beta\overline{B^e}$ may be inverted for each pixel to yield $\frac{\partial z^e}{\partial x} \simeq b^{-1}(\beta\overline{B^e}; \frac{\partial Z}{\partial y})$. The derivative of the gradient with respect to β may also be obtained. Then we use Newton's method to solve the equation $\frac{1}{MN}\sum_e b^{-1}(\beta\overline{B^e}; \frac{\partial Z}{\partial y}) = \frac{\partial Z}{\partial x}$ for β. That is, we require that the

normalization lead to the correct in-phase-plane gradient, averaged over the image. In practice, two or three iterations suffice to converge to an approximate β accurate to better than 0.2%.

To conclude this section, I show how two classes of *a priori* constraint, benchmarks and strike estimates, may be included in the finite element PC algorithm.The formulation of the PC problem as a constrained minimization carried out by the penalty method makes the introduction of additional constraints of variable weight entirely straightforward. Equation (2.7) may be generalized to read:

$$\delta\Big(\frac{1}{\alpha}S + \frac{1}{2}(\langle B\rangle - \langle b\rangle)(\{B\} - \{b\}) + \sum_m \frac{w_m}{2}(z(x_m, y_m) - \zeta_m)^2 + \sum_n \frac{w_n}{2}\sin^2(\phi(x_n, y_n) - \varphi_n)\Big) = 0, \quad (A.10)$$

where the first sum is over benchmarks of elevation ζ_m located at coordinates (x_m, y_m), and the second is over points (x_n, y_n) where the strike is estimated *a priori* to be φ_m, compared to the modeled value ϕ. The w_m and w_n are the weights with which these constraints are to be applied.

The functions $z(x, y)$ and $\phi(x, y)$ needed to evaluate equation (A.10) for any choice of tiepoints may be obtained using the interpolating functions chosen for the finite element scheme. A great simplification is possible, however, if we restrict our attention to benchmarks located exactly at nodes, and strike estimates in the centers of elements. Use of the constraints during iteration on the coarser meshes is thereby precluded, but I argue that this is not a drawback. Strike estimates in particular, and to a lesser extent benchmarks, probably refer to the finest scale features in the image, and thus ought not to be utilized during iteration at low resolution. (If these constraints are being used to facilitate fitting atmospheric or albedo parameters, such parameters can be held fixed except when working on the finest mesh).

With this in mind, let us define $\{\zeta\}$ as the vector whose ith entry is the benchmark altitude ζ located at node i, and $[K_1]$ as the diagonal matrix whose (i, i)

entry is the weight w to be attributed to this benchmark; if there is no datum for node i, let $K_{1ii} = \zeta_i = 0$. Similarly, let φ^e be the *a priori* strike angle at the center of element e, measured anticlockwise from the positive x axis. The corresponding weight, w^e, vanishes if φ^e is unknown. The equation (2.9) for a Newton-Raphson step then becomes:

$$\Big([K^k] + [K_1] + [K_2^k]\Big)\{\Delta z^k\} = \{E_1^k\} + \{E_2^k\} + \{E^k\}, \qquad (A.11a)$$

where

$$\{E_1^k\} \equiv -[K_1]\Big(\{z^k\} - \{z_0\}\Big) \qquad (A.11b)$$

$$[K_2^k] \equiv \left[\frac{\partial^2 \sum_e w^e \sin^2(\phi(\overline{x^e}, \overline{y^e}) - \varphi^e)}{\partial\{z\}\partial\langle z\rangle}\right], \qquad (A.11c)$$

$$\{E_2^k\} \equiv -\left\{\frac{\partial \sum_e w^e \sin^2(\phi(\overline{x^e}, \overline{y^e}) - \varphi^e)}{\partial\{z\}}\right\} \qquad (A.11d)$$

and, in terms of the finite element representation,

$$\sin^2(\phi(\overline{x^e}, \overline{y^e}) - \varphi^e) \equiv \frac{(\Delta_1 \cos(\varphi^e - \frac{\pi}{4}) - \Delta_2 \sin(\varphi^e - \frac{\pi}{4}))^2}{\Delta_1^2 + \Delta_2^2}. \qquad (A.11e)$$

The function $\sin^2(\phi - \varphi)$ was chosen to be positive semidefinite, vanishing when $\phi = \varphi$, and dependent only on the orientation, not the magnitude, of the local surface gradient. A case can be made that if the magnitude of the gradient is small, the orientation is less likely to be significant. Omission of the normalization in the denominator of equation (A.11e) would result in an appropriate slope-dependent weighting of the strike constraint (as well as simplifying the computation).

Appendix B: Finite-Element Formulation of Radarclinometry

Leaving aside such "practical" problems as radiometric calibration (Wildey 1984) and coherent noise, radarclinometry differs from photoclinometry in its unusual geometry (Figure 4.1). The radar unit moves along a more-or-less linear flight track above the surface of a planet, and constructs an "image" of brightness B as a function of slant range s (related to echo delay by the speed of light c) and azimuth z (decoded from

time and doppler-shift information). The brightness of the reflected radiation from a point on the surface depends on the single angle $\phi = \hat{n} \cdot \hat{\imath}$. The brightness function $b(\phi)$ may be assumed known for the purposes of formulating our algorithm.

We establish cartesian coordinates, in which we wish to describe the scattering surface; keeping the z axis parallel to the flight path, we make the y axis vertical. If the flight path is a distance h above the (x, z) plane, which is assumed to be the mean plane of the ground, the two coordinate systems are related by:

$$x = s \sin\theta, \quad y = h - s\cos\theta, \tag{B.1}$$

and

$$s = \sqrt{x^2 + (h-y)^2}, \quad \theta = \tan^{-1}\left(\frac{x}{h-y}\right). \tag{B.2}$$

By analogy with the PC problem, which can be written schematically

$$\text{Find} \quad z(x,y), \qquad \text{such that} \quad b(\hat{n}(x,y)) = B(x,y); \tag{B.3}$$

we might naively attempt to formulate the RC problem as:

$$\text{Find} \quad y(x,z), \qquad \text{such that} \quad b(\hat{n}(x,z) \cdot \hat{\imath}(x,y)) = B(s(x,y),z). \tag{B.4}$$

Unfortunately, y appears implicitly in this problem, and a solution is not forthcoming. The way out is to formulate the problem instead as:

$$\text{Find} \quad \theta(s,z), \qquad \text{such that} \quad b(\hat{n}(s,z) \cdot \hat{\imath}(s)) = B(s,z). \tag{B.5}$$

This is an explicit equation for $\theta(s,z)$, and hence may be solved as was the PC problem. Of course, the RC problem is underdetermined for exactly the like reason as photoclinometry, so we must apply a roughness criterion to winnow the multiplicity of solutions. Once we have the solution, we may express it in more conventional form $y(x,z)$ by means of the transformation (B.1).

We pay the price for this approach in the following ways: 1) Calculation of $\hat{n} \cdot \hat{\imath}$ from the "surface" $\theta(s,z)$ is more involved than calculating $\hat{n} \cdot \hat{z}$ in the PC problem. This is not, however, a fundamental obstacle. 2) The roughness criterion cannot (in practice) be applied precisely. An approximation equivalent to the use of $dx\,dy$ instead of $dX\,dY$ in equation (A.2b) must be made. 3) As the mesh becomes large, the finite element interpolation fails to represent the mean plane of the ground well, since the linear interpolation $\theta = a + bs$ describes a segment of an Archimedean spiral. Nonetheless, let us proceed.

We divide the (s,z) "plane" into rectangular elements, and adopt the usual local numbering in each. Define

$$\overline{s^e} = \frac{1}{2}(s_1^e + s_2^e), \quad \Delta s^e = s_1^e - s_2^e, \tag{B.6}$$

and

$$\overline{z^e} = \frac{1}{2}(z_1^e + z_4^e), \quad \Delta z^e = z_1^e - z_4^e, \tag{B.7}$$

We will also need the three combinations of the nodal angles:

$$\overline{\theta^e} = \frac{1}{4}(\theta_1^e+\theta_2^e+\theta_3^e+\theta_4^e), \quad \Delta_1 = \frac{1}{2}(\theta_1^e-\theta_2^e-\theta_3^e+\theta_4^e), \quad \Delta_2 = \frac{1}{2}(\theta_1^e+\theta_2^e-\theta_3^e-\theta_4^e). \tag{B.8}$$

It is convenient to choose elements of width $\Delta z^e = 1$, and with Δs^e such that if the element lay in the mean plane, it would have unit cartesian depth: $\Delta x_0^e = \left.\frac{\partial x}{\partial s}\right|_{y=0} \Delta s^e \simeq \frac{\overline{s^e}}{\sqrt{\overline{s^e}^2 - h^2}} \Delta s^e = 1$.

To evaluate the brightness function and roughness criterion, we want to express the cartesian components of the unit normal $\hat{n}$, and its derivatives in terms of the finite element approximation. Then the brightness and roughness derivatives may be obtained using the chain rule. We start with

$$n_y = \frac{1}{\sqrt{1 + (\partial y/\partial x)^2 + (\partial y/\partial z)^2}}, \quad n_x = -n_y \frac{\partial y}{\partial x}, \quad n_0 = -n_x \sin\overline{\theta} + n_y \cos\overline{\theta}. \tag{B.9}$$

The brightness is a function only of n_0, the component of $\hat{n}$ toward the radar, while it is n_y that will appear in the area roughness criterion. Using the chain rule, we express the cartesian derivatives as:

$$\frac{\partial y}{\partial x} = \frac{\overline{s}\sin\overline{\theta}(\partial\theta/\partial s) - \cos\overline{\theta}}{\overline{s}\cos\overline{\theta}(\partial\theta/\partial s) + \sin\overline{\theta}}, \tag{B.10a}$$

and

$$\frac{\partial y}{\partial z} = \frac{\overline{s}(\partial\theta/\partial z)}{\overline{s}\cos\overline{\theta}(\partial\theta/\partial s) + \sin\overline{\theta}}. \tag{B.10b}$$

The finite element approximations for the derivatives of θ at the center of the element are:

$$\frac{\partial\theta}{\partial s} = \frac{\sigma}{\overline{s}}\Delta_1, \quad \frac{\partial\theta}{\partial z} = \Delta_2. \tag{B.11}$$

where the denominator Δz has dropped out, and we have $\sigma = \overline{s}\frac{\Delta x_0}{\Delta s} \simeq \frac{\overline{s}^2}{\sqrt{\overline{s}^2 - h^2}}$, a known quantity. In terms of these approximations, we get:

$$\frac{\partial y}{\partial x} = \frac{\sigma\sin\overline{\theta}\Delta_1 - \cos\overline{\theta}}{\sigma\cos\overline{\theta}\Delta_1 + \sin\overline{\theta}}, \tag{B.12a}$$

and

$$\frac{\partial y}{\partial z} = \frac{\overline{s}\Delta_2}{\sigma\cos\overline{\theta}\Delta_1 + \sin\overline{\theta}}. \tag{B.12b}$$

Hence, it follows after some labor that

$$n_0 = \frac{\sigma\Delta_1}{\sqrt{1 + (\sigma\Delta_1)^2 + (\overline{s}\Delta_2)^2}}, \tag{B.13a}$$

for the component of the normal vector appearing in the brightness, and

$$n_y = \frac{\sigma\Delta_1\cos\overline{\theta} + \sin\overline{\theta}}{\sqrt{1 + (\sigma\Delta_1)^2 + (\overline{s}\Delta_2)^2}}. \tag{B.13b}$$

These equations are somewhat surprising; the former is entirely independent of $\overline{\theta}$, while clearly $\frac{\partial n_y/\partial\overline{\theta}}{\partial n_y/\partial\Delta_1} \sim \frac{1}{\sigma} - \Delta_1\tan\overline{\theta}$ is a small quantity (and will appear in the roughness criterion multiplied by $\frac{1}{\alpha}$). We are thus in the fortunate position of being able to neglect derivatives with respect to $\overline{\theta}$. As in the case of PC, the brightness and roughness will have only two independent first derivatives and three second derivatives. (These derivatives will be more expensive to compute, however, since they

require the evaluation of trigonometric functions.) Calculation of the partial derivatives of the unit normal is left as an exercise for the committee.

We next need to address the problem of applying the roughness criterion in radarclinometry. The rms z and entropy criteria can be approximated by sums over nodes as in PC; it is only the area criterion that appears problematic. The difficulty is that the quantity $\frac{1}{n_y}$ is not precisely the soap-film area of the element of surface, but the ratio of this area to its projection on the (x, z) plane. In the PC problem we minimized the sum over elements of the equivalent component $\frac{1}{n_z}$, and this was what was desired *if* the reference surface was $z = 0$, since the area of each element in that plane was unity. When an oblique reference surface was chosen, we ought to minimize $\frac{1}{n_\perp}$ per unit area in the oblique reference plane, $n_\perp$ being the component of $\hat{n}$ perpendicular to that plane. Instead we approximated $n_\perp$ and integrated over the plane $z = 0$.

In the case of radarclinometry, the problem is more severe. Were we to minimize $\sum_e \frac{1}{n_y^e}$ we would, in fact, be minimizing the soap-film area per unit area in the (s, z) "plane," *i.e.*, driving the surface towards tangency with the line of sight. The *correct* modification is to minimize $(\sum_e \frac{\Delta x^e \Delta y^e}{n_y^e})/(\sum_e \Delta x^e \Delta y^e)$. Unfortunately, the necessary normalization in the denominator makes the contribution at each element depend on all the nodes. The sparsity and bandedness on which practical computation depend are destroyed.

An approximate solution, in the spirit of that offered for the PC problem, is to substitute Δx_0^e for Δx^e in the above criterion. Since Δx_0^e does not depend on the nodal θ_i, the normalization in the denominator is a constant and does not enter into the minimization. The roughness criterion will now be over- or under-vigorously applied to a given element accordingly as it tilts away from or toward the radar ($\frac{\Delta x_0^e}{\Delta x^e} > 1$ or < 1.). Having already assumed small slopes in several places, we should

not be overly worried by this.

References

ALBERT, A. (1972). *Regression and the Moore-Penrose Pseudo-Inverse*, Associated Press, NY, pp. 119–123.

ANDREWS, H. C. AND B. R. HUNT (1977). *Digital Image Restoration*, Prentice-Hall, Englewood Cliffs, NJ.

BONNER, W. J. (1960). Photoclinometry from Oblique Photography. *U.S. Geol. Soc. Rept.* 1–30.

BONNER, W. J., AND R. A. SCHMALL (1973). A Photometric Technique for Determining Planetary Slopes from Orbital Photographs. *U.S. Geol. Surv. Prof. Paper.* **812-A**, A1–A16.

BRANDT, A. (1977). Multi-Level Adaptive Solutions to Boundary Value Problems. *Math. Comp.* **31**, 333–390.

BROWN, C. M. (1984). Computer Vision and Natural Constraints. *Science.* **224**, 143–156.

DALE, E. D. (1962). *The Application of the van Diggelen Method of Slope Analysis to Lunar Domes and Wrinkle Ridges*, Ph.D. Thesis (Unpubl.), Univ. of Manchester, Manchester.

DAVIS, P. A., AND A. S. MCEWEN (1984). Photoclinometry: Analysis of Inherent Errors and Implications for Topographic Measurements. *LPSC Abstracts.* XV, 194–195.

DAVIS, P. A., AND L. A. SODERBLOM (1982). Rapid Extraction of Relative Topography from Viking Orbiter Images II. — Application to Irregular Topographic Features. *Rep. Prog. Planet. Geol.* NASA TM 85127, 263–265.

DAVIS, P. A., AND L. A. SODERBLOM (1984). Modeling Crater Topography and

Albedo from Monoscopic Viking Orbiter Images. I. Methodology. *J. Geophys. Res.* **89**, 9449–9457.

DAVIS, P. A., L. A. SODERBLOM, AND E. M. ELIASON (1982). Rapid Estimation of Martian Topography from Viking Orbiter Image Photometry. *Rep. Prog. Planet. Geol.* NASA TM 85127, 331–332.

ELIASON, P. T., L. A. SODERBLOM, AND P. S. CHAVEZ, JR. (1981). Extraction of Topographic and Spectral Albedo Information from Multispectral Images. *Photogram. E. R. S.* **48**, 1471–1579.

GLAESSNER, M. F. (1961). Pre-Cambrian Animals. *Scient. Amer.* **204**, 72–78.

HAPKE, B. W. (1963). A Theoretical Photometric Function for the Lunar Surface. *J. Geophys. Res.* **68**, 4571–4586.

HAPKE, B. W. (1966). An Improved Theoretical Lunar Photometric Function. *Astron. J.* **71**, 333–339.

HAPKE, B. W. (1984). Bidirectional Reflectance Spectroscopy. 3. Correction for Macroscopic Roughness. *Icarus.* **59**, 41–59.

HESTENES, M. R., AND E. STEIFEL (1952). Methods of Conjugate Gradients for Solving Linear Systems. *Nat. Bur. Standards J. Res.* **49**, 409–436.

HOWARD, A. D., K. R. BLASIUS, AND J. A. CUTTS (1982). Photoclinometric Determination of the Topography of the Martian North Polar Cap. *Icarus.* **50**, 245–258.

IKEUCHI, K. AND B. K. P. HORN (1981). Numerical Shape from Shading and Occluding Boundaries. *Artif. Intel.* **17**, 141–182.

KERSHAW, D. S. (1978). The Incomplete Cholesky-Conjugate Gradient Method for the Iterative Solution of Systems of Linear Equations. *J. Comp. Phys.* **26**, 43–65.

KIRK, R. L. (1984). A Finite-Element Approach to Two-Dimensional Photocli-

nometry. *Bul. Amer. Astron. Soc.* **16**, 709.

KIRSCHVINK, J. L., R. L. KIRK, AND J. J. SEPKOSKI, JR. (1982). Digital Image Enhancement of Ediacaran Fossils: A First Try. *Geol. Soc. Amer. Abstr. with Programs.* **14**, 530.

LEWIN, R. (1984). Alien Beings Here on Earth. *Science.* **223**, 39.

LUCCHITTA, B. K., AND N. A. GAMBELL (1969). Evaluation of Photoclinometric Profile Derivations. In *Analysis of Apollo 8 Photography and Visual Observations*, NASA SP-201. 51–59.

MANDELBROT, B. B. (1982). *The Fractal Geometry of Nature*, W. H. Freeman, NY.

MANDELBROT, B. B., D. E. PASSOJA, AND A. J. PAULLAY (1984). Fractal Character of Fracture Surfaces. *Nature.* **308**, 721–722.

MCCAULEY, J. F. (1965). *Terrain Analysis of the Lunar Equatorial Belt.* U.S. Geol. Surv. open-file report. 44pp.

MCEWEN, A. S. (1985). Topography and Albedo of Ius Chasma, Mars. *LPSC Abstracts.* **XVI**, 528-529.

MEIJERINK, J. A., AND H. A. VAN DER VORST (1977). An Iterative Solution Method for Linear Systems of Which the Coefficient Matrix is a Symmetric M-Matrix. *Math. Comp.* **31**, 148–162.

MOORE, J. M., A. S. MCEWEN, E. F. ALBIN, AND R. GREELEY (1985). Topographic Evidence for Shield Volcanism on Io: Implications for Composition and Lithospheric Thickness. *LPSC Abstracts.* **XVI**, 575-576.

MOUGINIS-MARK, P., AND L. WILSON (1981). MERC — A Fortran IV Program for the Production of Topographic Data for the Planet Mercury. *Comput. Geosci.* **7**, 35–45.

ORTEGA, J. M. (1970). *Iterative Solution of Nonlinear Equations in Several Vari-*

ables, Associated Press, NY.

PASSEY, Q. R., AND E. M. SHOEMAKER (1982). Craters and Basins on Ganymede and Callisto: Morphological Indicators of Crustal Evolution. In *Satellites of Jupiter* (D. Morrison, Ed.), Univ. of Arizona Press, Tucson, pp. 379–434.

PEITGEN, H.-O., AND P. H. RICHTER (1986). *The Beauty of Fractals: Images of Complex Dynamical Systems*, Springer-Verlag, Berlin.

PRATT, W. K. (1978). *Digital Image Processing*, Wiley-Interscience, NY, p. 207.

PRESS, W. H., B. P. FLANNERY, S. A. TEUKOLSKY, AND W. T. VETTERLING (1986). *Numerical Recipes*, Cambridge Univ. Press, Cambridge.

RINDFLEISCH, T. C. (1965). A Photometric Method for Deriving Lunar Topographic Information. *JPL Technical Report 32-786.*

RINDFLEISCH, T. C. (1966). Photometric Method for Lunar Topography. *Photogram. Eng.* **32**, 262-277.

ROWAN, L. C., J. F. MCCAULEY, AND E. A. HOLM (1971). Lunar Terrain Mapping and Relative Roughness Analysis. *U.S. Geol. Surv. Prof. Paper.* **599-G**, G1–G32.

SQUYRES, S. W. (1981). The Topography of Ganymede's Grooved Terrain. *Icarus.* **46**, 156–168.

STASA, F. L. (1985). *Finite Element Analysis for Engineers*, Holt Rinehart, NY.

TYLER, G. L., R. A. SIMPSON, AND H. J. MOORE (1971). Lunar Slope Distributions: Comparison of Bi-static Radar and Photographic Results. *J. Geophys. Res.* **76**, 2790–2795.

VAN DIGGELEN, J. (1951). A Photometric Investigation of the Slopes and Heights of the Ranges of Hills in the Maria of the Moon. *Netherlands Astron. Inst. Bull.* **11**, 283–289.

VEVERKA, J. (1973). The Photometric Properties of Natural Snow and of Snow-Covered Planets. *Icarus.* **20**, 304–310.

WATSON, K. (1968). Photoclinometry from Spacecraft Images. *U.S. Geol. Surv. Prof. Paper.* **599-B**, B1–B10.

WILDEY, R. L. (1975). Generalized Photoclinometry for Mariner 9. *Icarus.* **25**, 613–626.

WILDEY, R. L. (1984). Topography from Single Radar Images. *Science.* **224**, 153–156.

WILDEY, R. L. (1986a). Radarclinometry for the VRM. *Photogram. E. R. S.* **52**, 41.

WILDEY, R. L. (1986*b*). Obstacles Facing the Venus Radar Mapper — The Implications of Gestalt Formation in Stereo-Radargrammetry. *Earth Moon and Planets.* **35**, 47–54.

WILHELMS, D. E. (1963). A Photometric Technique for Measurement of Lunar Slopes. In *Studies for Space Flight Program*, Part D of *Astrogeologic Studies Annual Progress Report.* U.S. Geol. Surv. open-file report. 1–12.

WILSON, L., J. S. HAMPTON, AND H. C. BALEN (1985). Photoclinometry of Terrestrial and Planetary Surfaces. *LPSC Abstracts.* **XVI**, 912–913.

ZINKIEWICZ, O. C. (1977). *The Finite Element Method*, 3rd ed., McGraw-Hill, NY.