SOME PROBLEMS IN MANTLE STRUCTURE AND DYNAMICS

PART 1: Inversion for Depth Variation of Spectra of Mantle Compressional and Shear Velocity Heterogeneity.

PART 2: Physical Model of Source Region of Subduction Zone Volcanism.

Thesis by
John Huw Davies.

In Partial Fulfillment of the Requirements
of the Degree of Doctor of Philosophy.

California Institute of Technology
Pasadena, California.

1990

(submitted December 14th, 1989).

Acknowledgements

My time in the Seismological Laboratory has been one of great pleasure. I am very grateful to all the faculty of the Division who have allowed me latitude in investigating problems of my choosing. Foremost I would like to thank my two thesis advisors. Professor Rob Clayton has been a generous, fair and understanding thesis advisor, he was ready to help me no matter how busy he was. For this I shall be ever grateful. Professor Dave Stevenson has been a most stimulating and thoughtful advisor. He nurtured my curiosity but kept my thoughts firmly on the Earth (my feet on the ground!). I should also like to thank both Professor Anderson and Professor Hager for encouraging me generally, and specifically to pursue the subduction zone modeling.

I have made a great many friends during my time at Caltech. In particular I wish to thank Oli Gudmundsson. Oli was a true friend. I would also like to thank Oli for the fun time we had collaborating on the method that is the basis of the first part of this thesis. I am very grateful to Oli for his genuinely unselfish and open attitude. The P-wave analysis using this method was also undertaken with Oli.

Many have helped me in my goal of completing my thesis over the years. I should like to mention office-mates, David Scott, Sally Rigden, Holly Eissler, Cathy Smither, Bruce Worden, David Wald and Joanne Yoshimura who have all suffered humourously my questions and chatter. They not only helped me with my work but also helped make my time both pleasant and rewarding. Other

students whom I would like to thank for their friendship and help include Luçiana Astiz, Steve Grand, Hua-Wei Zhou, Tom Duffy, Scott King, Phil Ihinger, Mark Fahnestock and David Bell.

Over the years I have had encouragement and help from numerous friends and relatives. I would like to thank Bedwyr for his support and friendship. This thesis belongs to Ann as much as myself; without her help it would be much poorer and would have been much more work. Even more than her practical help; her love, friendship and moral support have been critical in keeping me going over these last few hard months. Lastly I should like to thank my parents for their whole-hearted support over the years. I dedicate this thesis to Ann and my parents.

"I mam, dadi a'ng nghariad. Diolch."

Abstract

Part 1. The scatter in ISC P- and S-wave travel-time residuals was inverted to give a measure of the data-set's incoherent noise and the depth variation of the spectra of the Earth's seismic heterogeneity. The P- and S-wave models are similar in pattern with most of their power shallower than 400km. The power generally decreases with depth and is lowest around 1500-2600km depth. Both models show a slight increase locally below the 670km discontinuity. The long-wavelength half-width (l<50) is around 500km through the upper mantle increasing to around 1200km in the lowermost mantle. The variance in the travel-time residuals requires that $(\delta \ln V_S/\delta \ln V_P) \approx 2$, if they are correlated. Our results suggest values as high as 5 from 60-1400km; these could be correct but our preferred explanation is that it's a result of poor depth resolution of the shallowest layer and a difference in the spectral resolution of the two studies.

Part 2. Thermal modeling of a generic subduction zone suggests that the oceanic crust does not melt extensively to be the source of subduction zone magmas. The slab dehydrates and the water is transported laterally into the wedge by a mechanism involving transport fixed in amphiboles and vertical porous flow when free. This water generates melting at the amphibole saturated solidus. Melting reaches a maximum at the hottest geotherm, which also caps the source region. Melts depart the source region in cracks whose direction of propagation is perpendicular to the least compressive stress. For a corner flow regime this leads to focussing

of melt towards the wedge corner. The model correctly predicts the location of the volcanic front. The melt and residue provide buoyancy that leads to local flow reversal and modulates the volcanism with a period of ≈1Ma. Estimates suggest more water is subducted than reappears in extrusive volcanics. We suggest the excess water is stored in melts trapped deeper in the section that later become the precursors of granitic and tonalitic plutons.

Table of Contents

INTRODUCTION

This thesis is in two parts. In the first part we assess the spectrum of mantle heterogeneities as a function of depth through the mantle. In the second part we develop a physical model of the source region of subduction zone volcanism. These are two disparate attempts to understand the structure and dynamics of the mantle.

Present day seismic data and computational power only allow large scale length deterministic inversions, while many of the features of high Rayleigh number convection are known to be of a very short scale-length. This method gives us the ability to investigate the whole spectrum of seismic heterogeneity and can potentially provide unique constraints on convection modeling. Given the expected chaotic nature of mantle convection and the restriction to generic convection models, the statistical nature of these models is not a major limitation.

Subduction zone magmas (SZM) are one of the few direct samples of the mantle. Their undeniable association with one of the major features of the plate tectonic system suggests that understanding their generation could lead to an improved understanding of subduction in general. SZM are believed to be one of the agents for generating continental crust. Hence, better understanding of them would lead not only to an understanding of the evolution of the mantle through time, but also to an understanding of the production of continental crust. The wealth of geophysical, geochemical and petrological data appertaining to

subduction zones will also provide tight constraints on models of mantle convection and Earth evolution.

Only through developing a coherent understanding of the many different phenomena due to mantle convection will we develop a well-constrained dynamic model to explain plate tectonics. This thesis is one small contribution towards achieving this goal.

PART 1

Inversion of Depth Variation of Spectra of Mantle Shear and Compressional Wave Structures

ABSTRACT

We have developed a method of inverting the scatter in body-wave travel-time residuals to obtain the spectra of seismic velocity heterogeneity as a function of depth through the Earth's mantle. This method also gives a measure of the incoherent scatter ("noise") in the data set.

This method was applied to teleseismic ISC P- and S-wave data. The data were inverted to give a finely and coarsely parameterized model in depth for the P-waves and a coarse model for the S-waves.

The incoherent scatter suggests that the ultimate signal-to-noise ratio of both P- and S-wave data-sets is around 2 and 1.5 respectively. Evaluations of noise in data for deterministic inversions can be made, including that due to structure smaller than the grid size of the inversion, and therefore the problem of aliasing can be estimated.

We find that most of the power affecting travel-time residuals is located in the upper-most mantle. About half of the power has harmonic degree $l > 50$. The power with $l < 50$ has a half-width of around 500km in the upper half of the mantle, increasing in the lower half to ≈1200km. The P and S half-widths are equal within the error bounds of the studies.

There is poor resolution for the upper 60km, and the large-scale power is small here. Total, short and long wavelength power are maximum between 60 and 300km in the P-wave study. Total and short-scale power is also

a maximum for the shear wave study at this depth; but the long wavelength power reaches its maximum in the depth range 300-540 km. The ratio of power in the two studies is poorly defined in the shallowest bin and the lowermost mantle. For the rest of the mantle the ratio of $\delta ln(V_S)/\delta ln(V_P)$ *is greater than 4 and up to 7. Other seismic studies have suggested a value of 2. We suggest that this difference could be real because this study is the first to locate lateral variations at these depths for both S- and P-waves, but more probably it is a function of the large damping and poor resolution of the shallowest layer in the S-wave study and the difference between the spectral resolution of the two studies. Other studies were sensitive either to the lower mantle, or to the crust and uppermost mantle. We suggest that such high ratios (2 or 5) could be the result of either a change in mineral response to temperature due to high pressure or a reflection of compositional variations. A ratio as high as 5 would have a significant bearing on surface wave and free oscillation studies; where they presently assume correlation with a ratio between 1.25 and 2.5. The signal weighted value of this ratio is* $\approx$ *2 assuming that the P and S fields are perfectly correlated, and that the sampling is identical. All these are much larger than the value of* ≈ 1 *that the atmospheric pressure experiments suggest.*

The pattern of variation of the heterogeneity in both studies is very similar. It decreases by an order of magnitude or more in going from the upper mantle to the lower mantle. In the high resolution P-wave study this drop is seen to occur abruptly at around 300km depth. This is supported but not resolved by the two coarse models. We suggest that this is a reflection of the ocean continent dichotomy; and it suggests that this transition occurs at a relatively uniform depth globally. Interestingly both studies show a blip in the decreasing trend from upper to lower mantle just at the 670 km discontinuity. This blip again could be an artifact of the heavy damping of this poorly

resolved depth. There is more heterogeneity below than above this discontinuity. The P-wave study also shows an increase in heterogeneity at the CMB of a short half-width ≤ 400km. Our S-wave data-set does not sample near the CMB, since at those epicentral distances it is contaminated with SKS arrivals and is ignored.

There is no evidence of periodicity in the travel-time auto-covariance; suggesting that there is little or no periodicity in the underlying convection. Equally, the short half-width through most of the mantle suggests high Rayleigh number convection, with its attendant small-scale structures. The large difference in power between the upper and lower mantle suggests either a change in convective regime or a difference in the thermal sensitivity of the elastic constants for the characteristic minerals in both layers.

The increased short-scale power at both the top and bottom of the mantle suggests that a large part of the seismic signature at these boundaries is compositional since one would expect a red spectra for a solely thermal boundary layer. If the temperature variations predicted from the seismic variations drive the flow through thermal buoyancy forces and correlate with the flow, then heat flow constraints can be satisfied with reasonable estimates of viscosity. The length scale and magnitude of temperature variations in the lower mantle suggest that there is less than one order of magnitude variation laterally in long-wavelength viscosity, hence models deriving the long-wavelength geoid and plate motions from internal buoyancy forces, assuming spherical symmetric viscosity , are most probably valid.

1.1 Introduction

Earth Science has undergone a major advance over the last two decades through the unifying framework provided by plate tectonics even though it is only a kinematic theory. Since its earliest days, people have searched for a dynamic model. It was rapidly accepted that this involved convection of the Earth's mantle; if only because that was the only way to transport the heat generated by radioactivity to the surface. In the last two decades we have not solved the mystery of how plate tectonics is powered. Seismology, which has given us our most detailed information on the Earth's interior, was, until this decade limited to radial, spherically-symmetric Earth models [*Dziewonski and Anderson*, 1981; *Jeffereys*, 1939]. Convection is driven by lateral temperature differences and hence spherically-symmetric models give us no information on the location of these driving forces. Seismology gave us only limited information on the dynamic state of the Earth's interior e.g., inhomogeneity parameter [*Masters*, 1979]. Now there are models that attempt to address the degree of lateral variation in velocity in the mantle and hence potentially locate the driving forces directly [*Clayton and Comer*, 1983; *Dziewonski*, 1984].

Modeling of mantle convection has advanced from 2D cartesian, Boussinesq models of convection with a uniform viscosity, incompressible fluid to modeling of convection in 3D spherical geometry, with temperature, pressure and stress dependent viscosity, in a compressible fluid. Even with these advances the models are still generic in that they do not claim to know the initial or boundary

conditions well enough to be able to predict the actual behavior of the Earth. Most probably this will never be possible since mantle convection is most probably chaotic [*Stewart and Turcotte*, 1989]. If the initial conditions are sufficiently well known then reasonable predictions might be possible over short time frames but in general we shall most probably only be able to compare the statistics of convection modelling with observations on the real Earth. Hence, for addressing the question of Earth dynamics we suggest that a seismic model attempting to derive the statistics of seismic velocity variations is not a limitation. In fact, it has the added advantage that it can image the whole spectrum. Due to the reduction in model parameters, we should be able to obtain more robust estimates than possible from deterministic models of the statistics of velocity variations in the mantle. It is likely that the primary features of mantle convection will have a small scale-length. It is still impossible to deterministically image these scales, even with high resolution, tomographic techniques.

The earliest models of the Earth's velocity field were feats of statistics and computation and have remained a reasonable representation of the Earth's velocity field [*Jeffereys*, 1939; *Jeffereys and Bullen*, 1940]. These models and others were radially varying but spherically symmetric [*Dziewonski and Anderson*, 1981; *Herrin*, 1968]. They provided the backbone on which identification of phases was based and are the reference models from which aspherical perturbations are calculated. In this study we shall use the original Jeffereys-Bullen (JB) model as our reference model; it is a very reasonable model for teleseismic epicentral distances. In figure 1 we

show the travel-time curves for P and S-waves and interfering phases. We observe that the P phase is separate until PcP meets it asymptotically at close to 100°; hence there are no crossing phases. For the S-phase by contrast we notice that it is crossed by ScS and SKS at around 80°. In figure 2 we show the S-wave travel-time residuals as a function of epicentral distance. For teleseismic rays ($\Delta >$ 30°) we see that there is little bias in the reference model i.e., we have a symmetric distribution. At 80° the interfering SKS arrivals identify themselves as a distinct streak nearly vertically across this pattern.

As discussed above, the interest has moved away from developing spherically symmetric Earth models to evaluating aspherical structure. All these models to date have been deterministic, and have expressed their results as 3D models of velocity variations throughout the Earth's mantle and/or core. In analogy with medical imaging and its CAT (Computer Aided Tomography) scanner, this field has been called "tomography". In both medical and seismological fields, data representing line integrals through 3D space are converted into 3D models that can be viewed by taking slices ("tomo") through the model. In CAT scanning the rays are X-rays and the images are of X-ray absorption; while in seismology the rays are "seismic rays" following the path of seismic energy from the source to the receiver; while the images are the seismic velocity or relative attenuation.

Seismic tomographers face many issues in attempting to develop the best possible images of the Earth. These include deciding what parameterization to adopt for the model space. First, this con

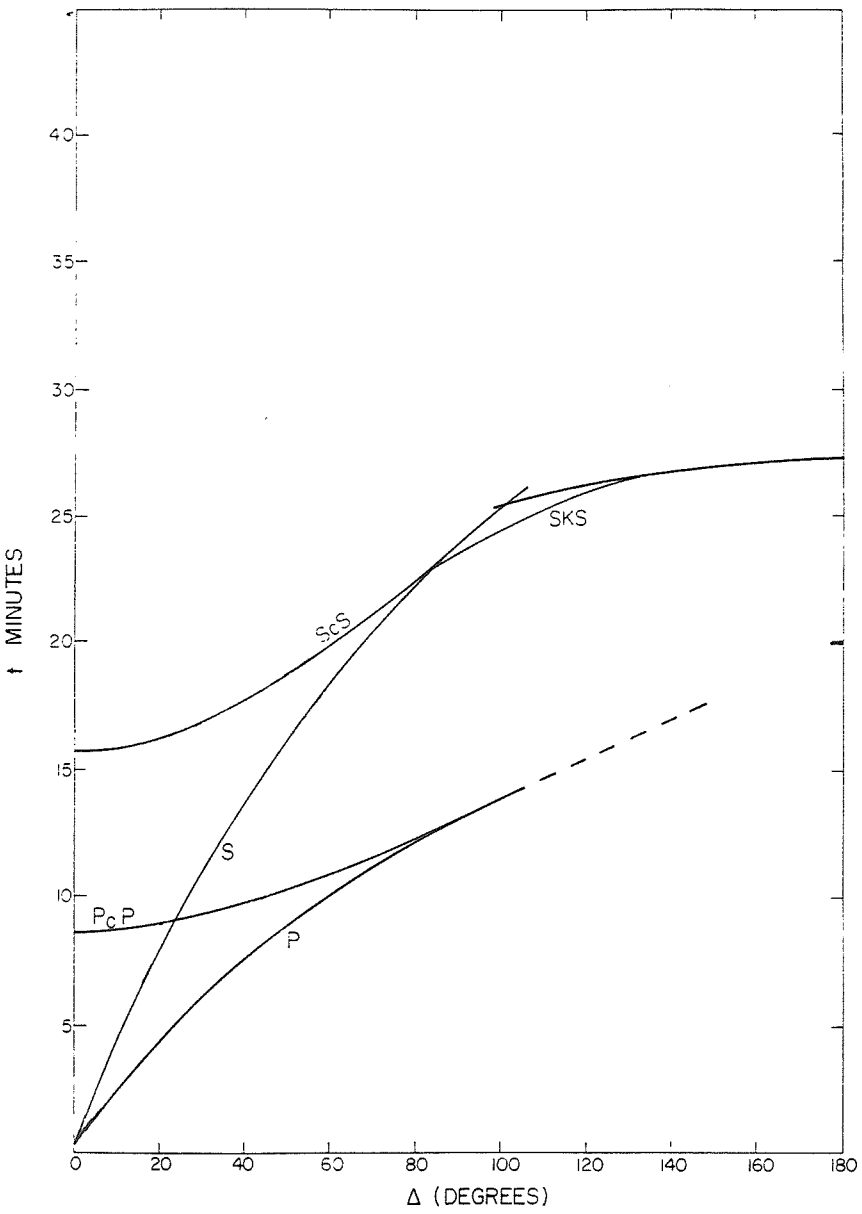


Figure 1.1 Plot of travel-time versus epicentral distance for the S, P and other interfering phases, e.g., SKS and PcP.

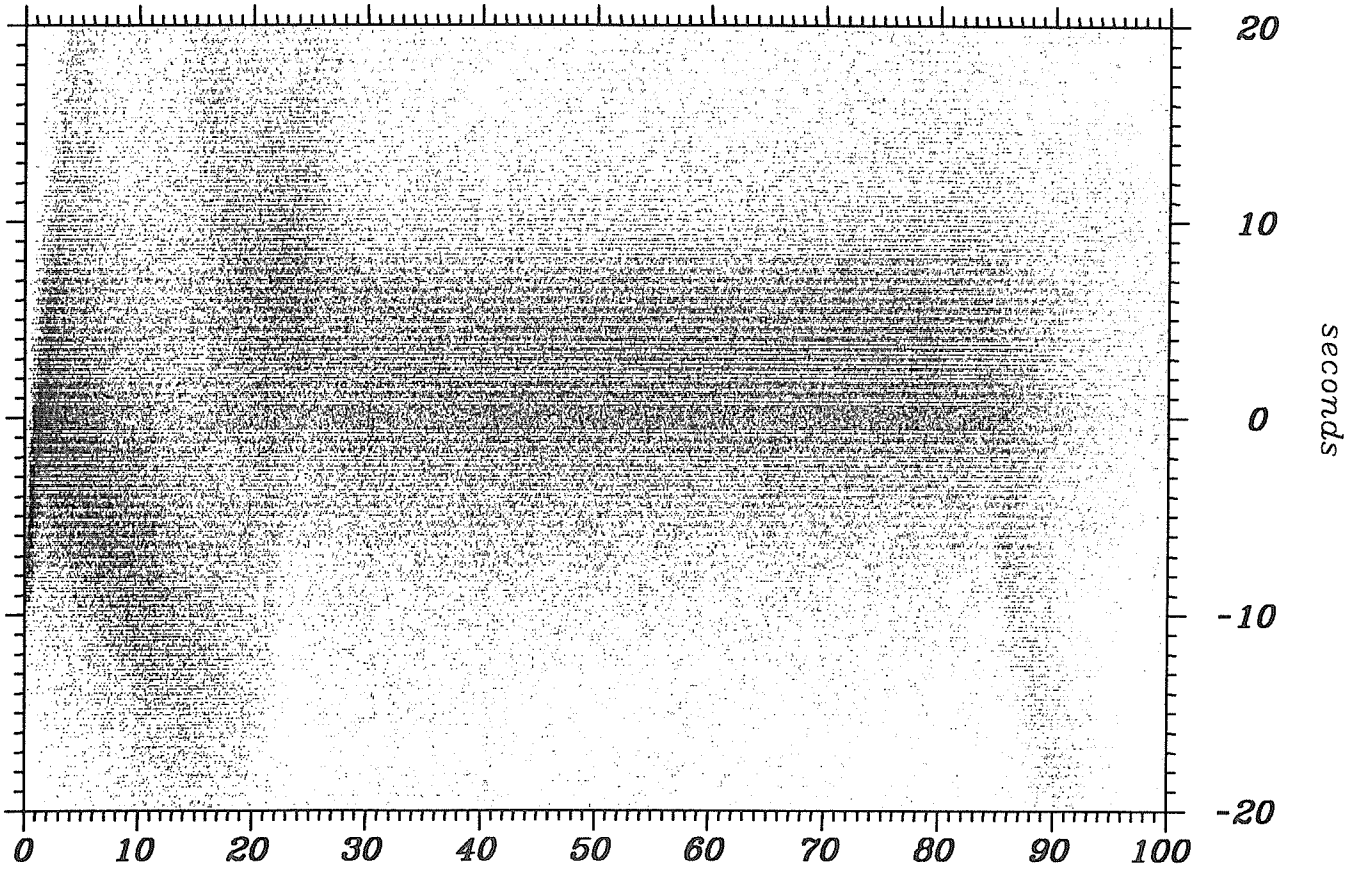


Figure 1.2 Plot of Jeffereys-Bullen shear wave travel-time residuals as a function of epicentral distance. Notice that their distribution is uniform over teleseismic distances (30° to 100°) with the exception of the vertical streak due to SKS from ≈ 80° to 90°.

cerns what physics to include; e.g., whether the mantle should be isotropic or anisotropic, and whether density and P and S velocities are correlated. Second, it includes what cell sizes to have at different depths through the mantle. This second point could be addressed more efficiently if one knew before hand what the magnitude and scale-length of the perturbations were. One also needs to judge how to damp the models and trade-off model errors and resolution. To estimate the errors in the model one needs to decide what the errors are in the data. Many workers use huge data-sets (e.g., the catalogue of the International Seismological Society (ISC)) that make it impractical to personally examine all the original data (due to its volume [millions of picks], and inaccessibility); therefore it is difficult to make a detailed assessment of the errors in the data-set. Even workers who use much fewer data have difficulty in deciding what errors arise from mislocation of the sources, etc. Hence, any method allowing an estimate of errors will also help in deterministic inversions.

With mantle body-wave tomography there have been two main model parameterization schemes. First, cell models where the model space is divided into constant value cells. Second, the model space is parameterized by a truncated harmonic expansion. The first method has been applied to iterative back-projection techniques that have allowed the inversion for very large model spaces. The second method has generally been applied to one-step large linear matrix inversion methods. A third method, which has yet to be used in 3D body-wave tomographic inversions, is the Backus-Gilbert method. Here the model, at a point, is made from a sum over the data. The result can be derived at any point in the model where the

resolution and error are traded-off to give the optimum result at this point. This is an expensive method to apply to large data-sets and to examine the whole of a 3D model. Therefore the Backus-Gilbert method is presently of limited use in 3D studies; it has been applied to surface wave studies of upper mantle heterogeneity [*Tanimoto*, 1985; 1986]. Since only the first two methods are presently practical for achieving high resolution images, it would be useful to have methods for evaluating the magnitude and scale-length of seismic variations that would allow better parameterization of model spaces, which would lead to better deterministic images of the structures. This could lead to the development of models where the trade-off between error and resolution is made uniform so that the global damping parameters of the cheaper generalized inverse methods can be applied and give equivalent local trade-off as provided by the Backus-Gilbert method.

Other studies of aspherical structure include observations from scattering studies, splitting of the Earth's free oscillation modes, many regional body-wave waveform studies, and phase differences between reference synthetic and observed surface waves. The scattering studies have been concentrated on localized features underneath arrays, and near the core-mantle boundary. The splitting studies give information on only the even harmonic degree structure of the asphericity, and have yet to be inverted directly for structure beyond harmonic degree 4. The surface wave studies have had limited resolution in the mantle until the overtones were used, these have led to the whole Earth being imaged [*Tanimoto*, 1989a]. Assumptions are made regarding density and P-wave velocity varia-

tions in the surface-wave and free oscillation studies. They are usually assumed to correlate with the S-wave variations; and a constant value is assumed for the ratios throughout the mantle. These surface wave studies have produced the only global images of the Earth's upper mantle.

We propose a robust method to arrive at an indication of the Earth's 3D structure and an indication of the incoherent noise in the data-set. This model is intermediate between the earliest spherically symmetric models and the most recent 3D models. Simplistically the earliest models can be thought of as describing the Earth as onion layers where the velocity is constant in each layer. The deterministic 3D models can be thought of as dividing the layers up into cells and evaluating the best constant value to each cell; i.e. each layer would look spotty and more importantly the spots can be located relative to features on the surface. Continuing the analogy, our method describes the spectrum for the seismic velocity variations in each layer; i.e. defining a measure of the magnitude of the velocity variations and their scale-length: each layer could be described by two numbers. One describes how dark and light the spots are and the other describes how large they are. Note that we have no information on where in the layer the spots are dark or light. This method provides information only on the statistical variations of the velocity field. The method is more robust, because it concentrates on getting a simpler description of the velocity fields, i.e., the spectrum. It doesn't tell us whether the shear wave velocities are fast or slow at 1200km below North America, just that globally the variations are of scale-length 500 km and are of order $2\ 10^{-2}$ km/s.

The heart of this method is evaluating the scatter in travel-time residuals of rays that travel similar paths. Consider two earthquakes occurring at an identical location. They should have the same travel time to all stations. Any difference between the travel-time to identical stations from events at supposedly identical locations must be a reflection of errors in picking the arrival time of the phase, errors in locating the source, errors in the clocks at the stations, etc. Consider now an Earth with laterally uniform velocities at all depths except for one layer where the velocity is lower by a constant amount in one hemispherical patch. Consider two rays that are close together and travel through the same half of the patch, they will record the same residual. Now consider that the rays are spaced sufficiently apart so that one ray goes through this patch and the other doesn't; they will have different residuals. By discovering at what separation the difference appears, we can estimate the size of the patch. From the magnitude of the difference in the residuals we can also evaluate the difference in the velocity of the patch away from the reference; assuming of course that we know its depth variation. We can investigate the depth variation of the velocity variations by comparing the scatter in bundles of rays travelling different distances, since rays travelling further penetrate deeper into the Earth.

Due to the problems caused by triplications, we only used teleseismic phases. Hence, all the rays have very similar path lengths in the upper mantle; and therefore we would have poor resolution of the variation of heterogeneity in the upper mantle. Earthquakes though occur at all depths in the upper mantle and we can gain some

resolution of the variations in the upper mantle by comparing the scatter in bundles of rays that arise from different source depths. The deeper source depth bundle is missing one leg in the upper mantle, and hence doesn't pick up the scatter due to heterogeneities along this leg.

This method was first presented at the SSA meeting in Hawaii in May, 1988 [*Davies et al.*, 1988]. The spherical geometry formulation was first presented at the Fall AGU meeting in San Francisco [*Gudmundsson et al.*, 1988]; while the results for the compressional study were first published in *Gudmundsson et al.*, [1989] and repeated in *Gudmundsson* [1989]. This is the first presentation of the results for the shear velocities and the comparison between the compressional and shear wave results.

1.2 Data Analysis

The Earth was divided into a grid of equal area cells. The cell boundaries are along lines of longitude and latitude while the equator was a line of symmetry. An example of an equal area grid is presented in figure 1.3. Grids were generated at scales varying all the way from one box for the whole Earth to boxes that are 1^{o} on a side for S, and 0.5^{o} on a side for P. Rays which shared the same epicentral distance (the distance between source and receiver) bin and also the same source depth bin were considered together. A summary ray is a collection of rays that share the same epicentral distance and depth bins, and also the same receiver and source cells. figure 1.3 presents three summary rays that belong to the same epicentral and depth bin. The bins in the shear-wave study were 4^{o}

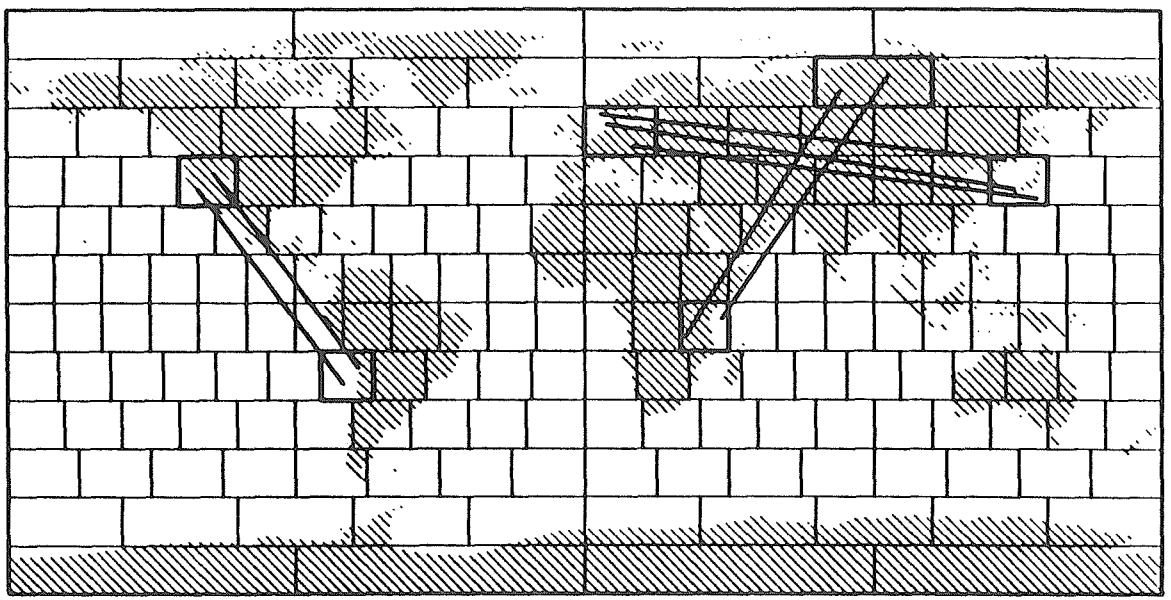


Figure 1.3 An equal area grid with 3 summary rays. These 3 summary rays belong to the same bin in epicentral distance and potentially the same source depth bin. Variance of residuals of component rays of these summary rays are contributors to the datum for this grid-size, epicentral distance and source depth bin. From *Gudmundsson et al.* [1989].

wide in epicentral distance, extending from 31° to 75°; while they were 2° wide in the compressional wave study extending from 30° to 100°. The bins in depth were 0-32km, 34-60km, 60km-100km, 100-200km, 200-400km and 400-700km, in both studies. Δ represents epicentral distance, while Z represents source depth.

Scale was defined by two conventions. First, the scale of a grid was defined to be the length of the side of a flat square box that contained the same average surface area as a typical cell from the grid. This measure is always reported in kilometers, and is primarily used in the figures for the S-wave data. The second convention describes the scale by the equivalent angular radius of a spherical cap that has the same area as a typical cell; this scale is always reported in an angular measure, degrees or radians, and is primarily used in the figures for the P-wave data. For small box size these measures can be converted with reasonable accuracy, 1° is $\approx$ 110km. At large box size the conversion deteriorates since squares and circles of the same length side or diameter don't have the same area. θ represents scale.

We used data collected by the ISC from 1964-1986. We selected S- and P-wave arrivals from events that had at least 50 phases reported because these would be better located. Any event located at 0 km or 33km depth was ignored. ISC frequently assigns poorly located events to these depths.

We used only shear wave arrivals out to 75° to avoid the contamination of S-wave picks mentioned above. The contamination is the result of mis-picking SKS, which crosses S at around 80°; see figure 1.1. The contamination can be observed in figure 1.2 as the

streak that is directed down and slightly to the right at this distance. Not having data out to 100° means that we don't sample the lower 450km of the mantle. Hence we have no information in this analysis on the S-wave structure of the Core-Mantle boundary (CMB) or D ", the distinctive boundary layer just above the CMB. We only selected shear-wave arrivals whose JB travel-time residual (δt) (i.e. the actual travel-time less the travel-time predicted for that ray through the reference model) were greater than -9 s. or less than 11s. These bounds allowed all rays that were within 2.5 standard deviation from the mean (+1.0s). The distribution of the travel-time residuals is better modeled by an exponential distribution than a Gaussian distribution, and this selection is like 2 e folding distances of the exponential distribution away from the mean. This condition on residuals removed all wildly incorrect residuals. For the P-wave study we selected JB residuals such that $|\delta t| < 4$s, this is closer to 3.5 standard deviations. While this procedure removes some of the more extreme noise it might also be eliminating some of the largest signal. This selection suggests that we are possibly underestimating the signal, but since more than 95% of the data is included this won't be significant.

To measure the degree of correlation between the travel-time residuals of the different component rays of a summary ray we evaluated the variance of the residuals, defined as follows.

$$\sigma^2_k = \sum_{i=1}^{n_k} (\delta t_i - \overline{\delta t_k})^2 / (n_k - 1) \qquad (1.1)$$

where, δt_i is the travel-time residual of the i^{th} component ray, δt_k is

the mean residual for the k^{th} summary ray. σ^2_k is the variance of the k^{th} summary ray; and n_k is the number of component rays in the summary ray.

We only used summary rays that had at least four component rays; i.e $n_k \geq 4$. Then we averaged the variance of all the summary rays at a given grid size, epicentral distance and source depth bin. To improve the estimate of this variance we repeated the process on the same grid but with all the cells displaced one-sixth of their cell width in longitude. This was done six times in total and the mean of all six was used as the estimate of the variance for this data subset. In the P-wave study we used 1/4 cell displacement and four rotations rather than 1/6 and six rotations.

$$\sigma^2(\theta,\Delta,Z) = \left(\sum_{q=1}^{n_r} \sum_{p=1}^{n_s} \sigma^2_{p,q} \right) / n_r n_s \qquad (1.2)$$

Where p is the summation index over the n_s summary rays that are from the same θ,Δ, and Z bins; and q is the summation index over the n_r orientation of the grids; $n_r = 6$ for S-wave study and $n_r = 4$ for P-wave study. $\sigma^2_{p,q}$ is the variance of an individual summary ray; evaluated as in equation 1.

The error in the estimate was derived from the scatter between the estimates of the 4 or 6 unrotated grids. At small-scales with grid size less than 5^o the data were only included if there were more than 25 summary rays, i.e. $n_s > 25$.

Each epicentral distance and source bin were characterized by a reference ray that was located at the mean epicentral

depth and travelled the mean epicentral distance for that data subset. These reference rays were traced through the spherically symmetric Jeffereys-Bullen reference Earth model. For the S-wave study we had 66 reference rays, 11 (epicentral distance bins) times 6 (epicentral depth bins); while we had 210 reference rays in the P-wave study (35 times 6).

All the data are presented in figures 1.4 and 1.5. All the S-wave data for all the scales are presented in figure 1.4(a). The scale axis is logarithmic; this allows the data at all scales to be reasonably viewed. Interestingly, the data approximately describe a straight line. In figure 1.4(b) we present data for two selected depth bins close in to the origin with a linear scale axis. Notice that these data also describe a reasonable straight line. In figure 1.5(a) we present a subset of the P-wave data with a linear scale axis; while in figure 1.5(b) we present two selected depth bins close in to the origin. We observe that the data show the general trend of decreasing variance at decreasing scale as would be expected since the rays are travelling ever more similar paths. Note the knee in the P-wave data occur at a distance of 5^{o}-10^{o}; for even sampling it implies that the signal averaged scale-length is of the same order. Since the sampling is clustered, the actual scale-length will be shorter. We also observe that the data show a general trend for deeper source depth data to have lower variance. This is as expected since they have shorter path lengths over which to sample heterogeneities. If the Earth had constant heterogeneity throughout then variance would be expected to increase linearly with path-length. The variance decreases slightly as a function of epicentral distance, this implies that the degree of

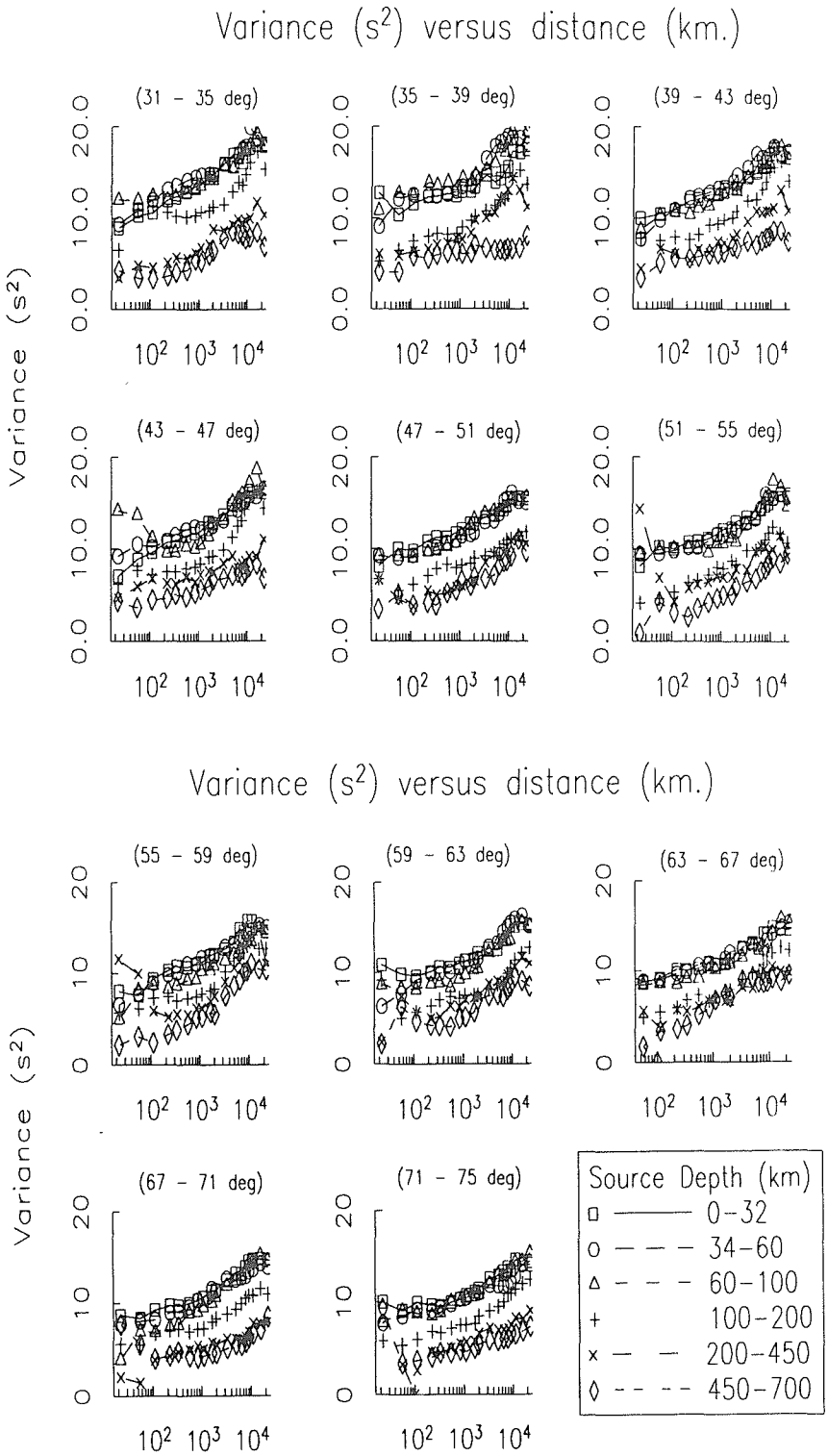


Figure 1.4 (a) Variance of all S-wave data plotted versus the logarithm of scale.

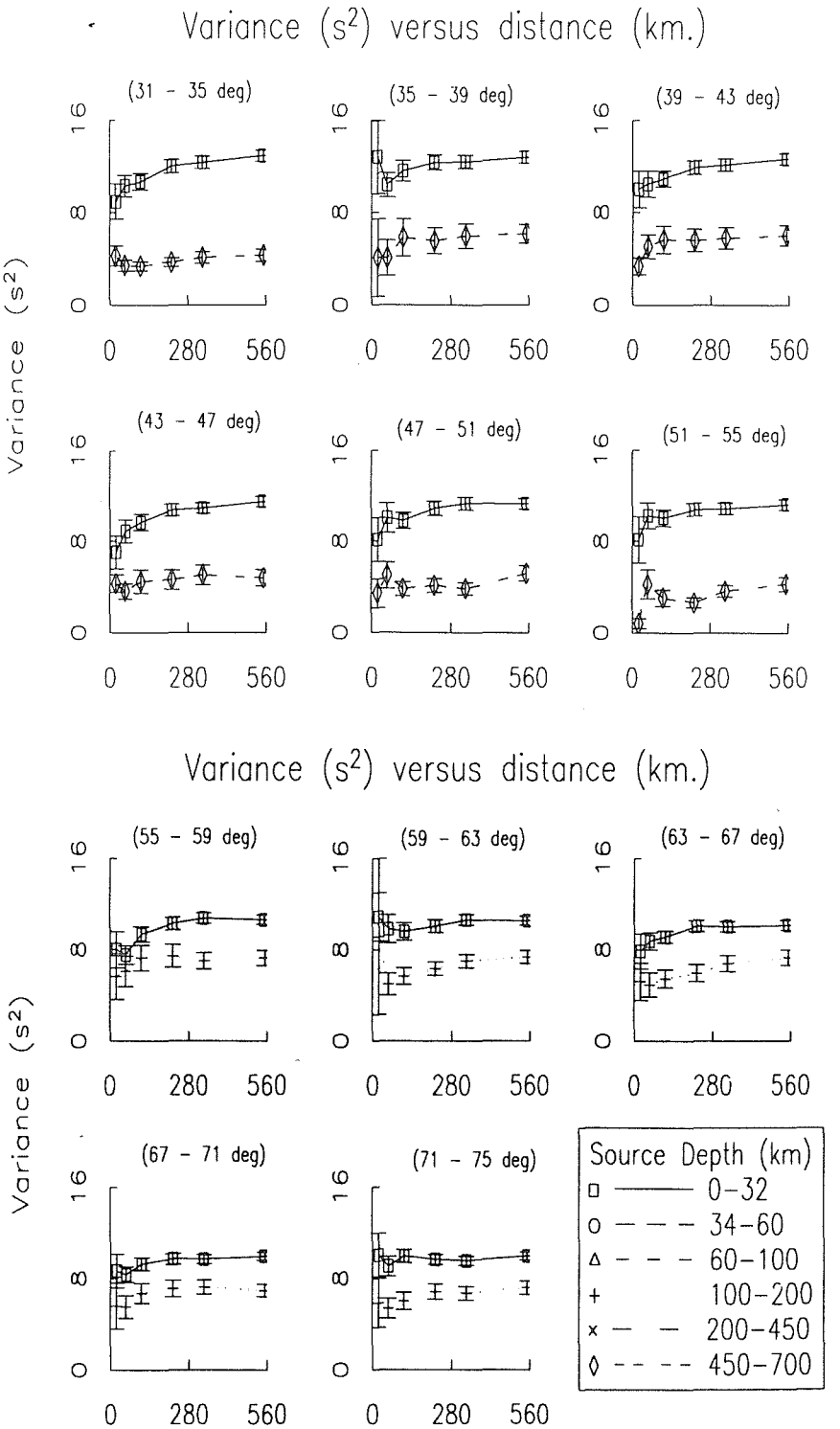


Figure 1.4(b) S-wave variance for small scales, note scale axis is linear.

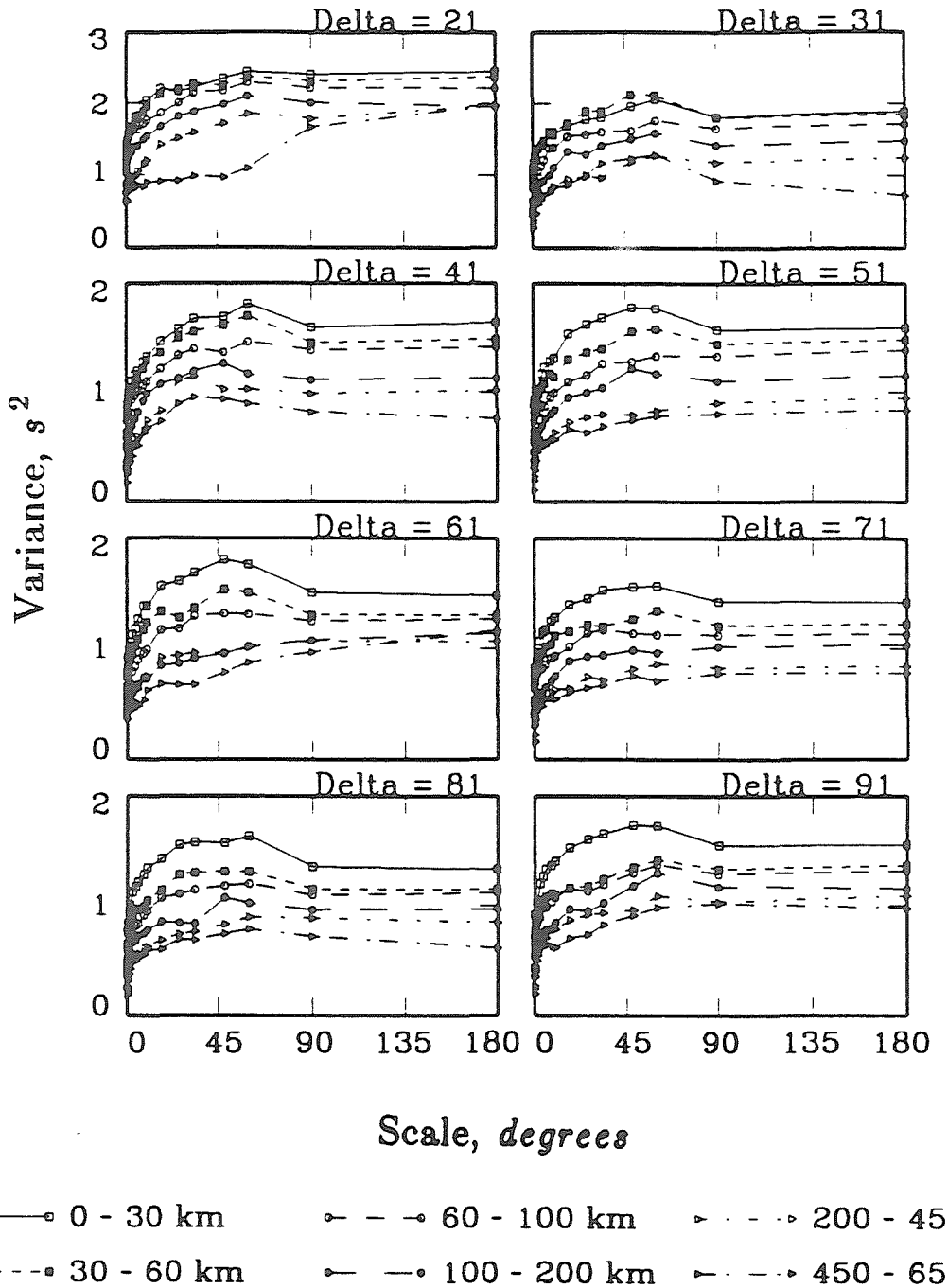


Figure 1.5.(a) Selection of P-wave data; summary ray variance as a function of grid size, scale axis is linear. From *Gudmundsson et al.* [1989].

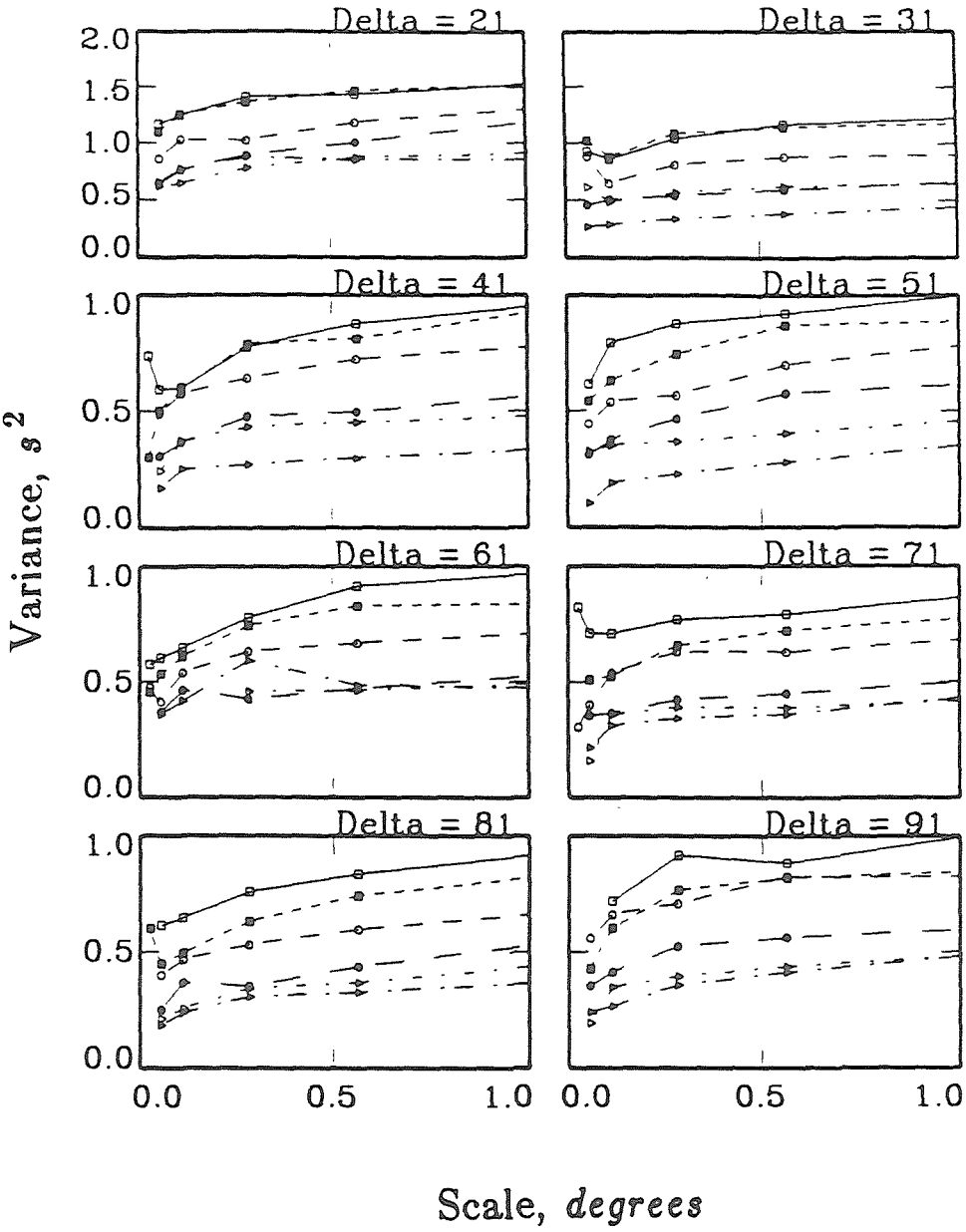


Figure 1.5(b) P-wave summary ray variance at small scale. Linear axis. From *Gudmundsson et al.* [1989].

heterogeneity decreases with depth. At the smallest scales we would predict non-zero intercepts from extrapolating the data towards the axis corresponding to scale size of 0km. We believe that this intercept is a measure of the areally incoherent signal that we equate with picking error, etc. Hence the data curves follow our intuition in their behavior and are reasonably coherent from curve to curve. This gives us hope that we can model this signal and invert for some properties of the Earth structure.

We make estimates of the intercept by extrapolating all the data for a given depth bin of a scale smaller than 5^o assuming that the decay is linear and that the slope is fixed as b=B/cos(i) where i is the angle of incidence at the receiver. The linear decay is supported by the fact that the data at the smallest scales is seemingly well described by a straight line, as seen in figures 1.4(b) and 1.5(b). We are assuming that most of the small-scale signal is located shallowly close to the receiver. Hence 1/cos(i) is proportional to the path-length near the surface. The different source depth bins seem to require different intercept estimates; this suggests that the incoherent signal varies and is not solely due to different estimates of smallest scale signal for the different depth bins but reflects a real decrease in the estimate of incoherent noise as the source depth increases.

In figures 1.6(b) and 1.7(a) we illustrate the intercept estimates for all the S and P reference rays respectively. We find that the estimates decrease with source depth and are relatively constant as a function of epicentral distance. Notice for the P-waves we also estimate the intercepts at regional epicentral distances.

These intercept estimates are much higher than those at teleseimic distances. This is not unexpected considering the triplication complications on the travel time curves due to the 400km and 670km discontinuities. The average variance for the S-wave intercepts is $\approx 7\ s^2$ as compared to a total signal of $\approx 14\ s^2$. For the P-wave analysis the intercepts are $\approx 0.8\ s^2$, while the teleseismic signal is $\approx 1.6\ s^2$. The error in estimating the intercept is quite high and hence the residual data arising from subtracting the intercepts from the original data have proportionately higher error estimates. The extrapolation of the intercept could be done logarithmically rather than linearly. The straight line that the data describe in figure 1.4(a) would support this choice. This was actually undertaken for the S-wave study; the intercepts are shown in figure 1.6(a). The inversion of the residual data set arising from assuming these estimates was also undertaken. The primary difference arises in estimates of the small-scale power; in contrast, the long-wavelength features are robust and are largely insensitive to the method of estimating the intercept. The S-wave data resulting from the subtraction of the linear intercept estimates (figure 1.6(b)) are shown in figure 1.6(c). These are the data that are inverted. Figure 1.7(b) presents a smoothed version of the raw P-wave intercept estimates in figure 1.7(a); these were used in the P-wave study, and not the raw estimates.

Generally one might prefer to invert for the intercept and structural signal simultaneously. In this way one could get a better fit to the data and also better estimates as to how errors in the data are reflected in both the estimates of intercepts and structure. Also, one could estimate what resolution the data gave for the intercept

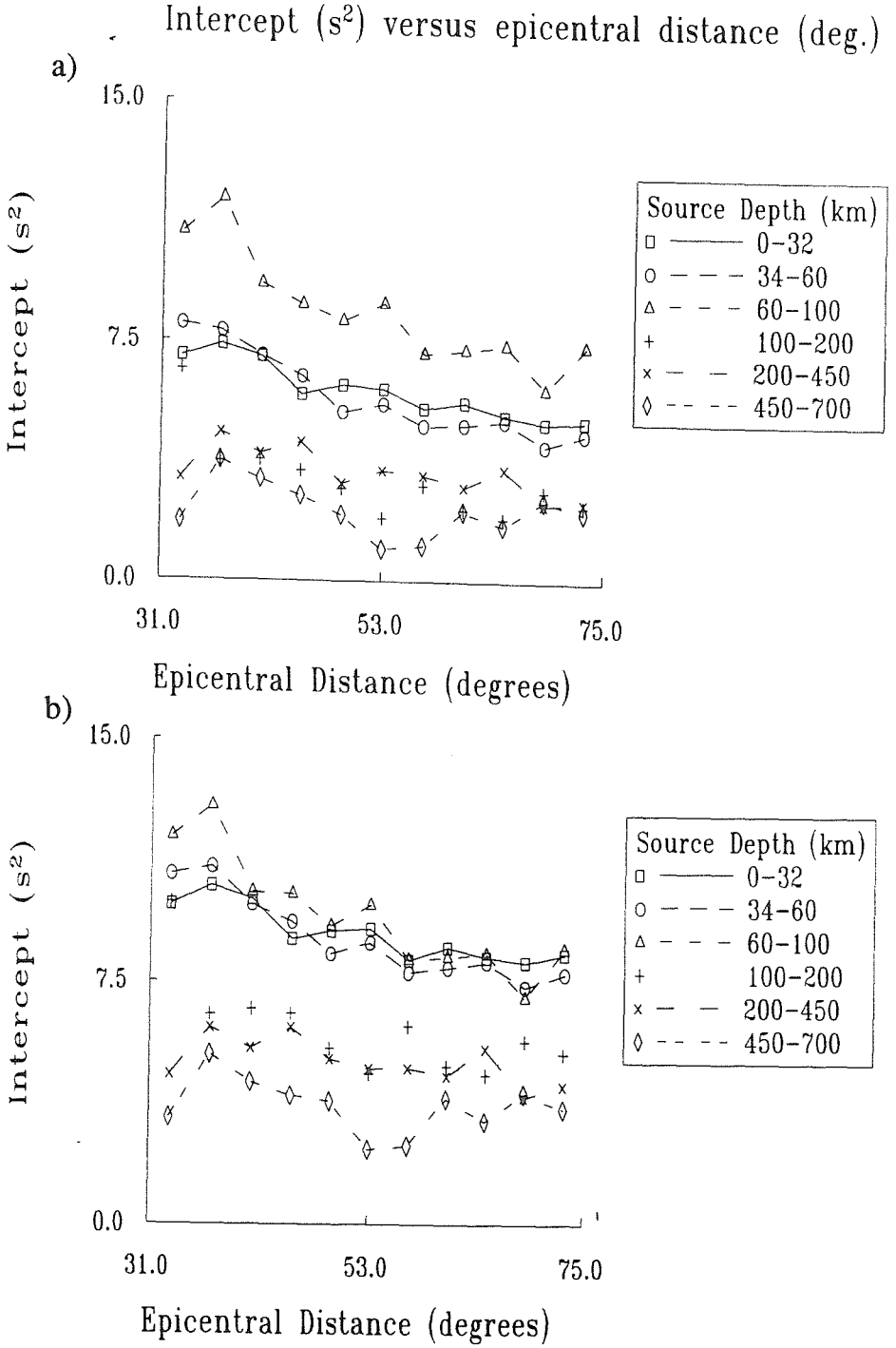


Figure 1.6 S-wave intercept estimates versus epicentral distance. (a) Logarithmic extrapolation. (b) Linear extrapolation.

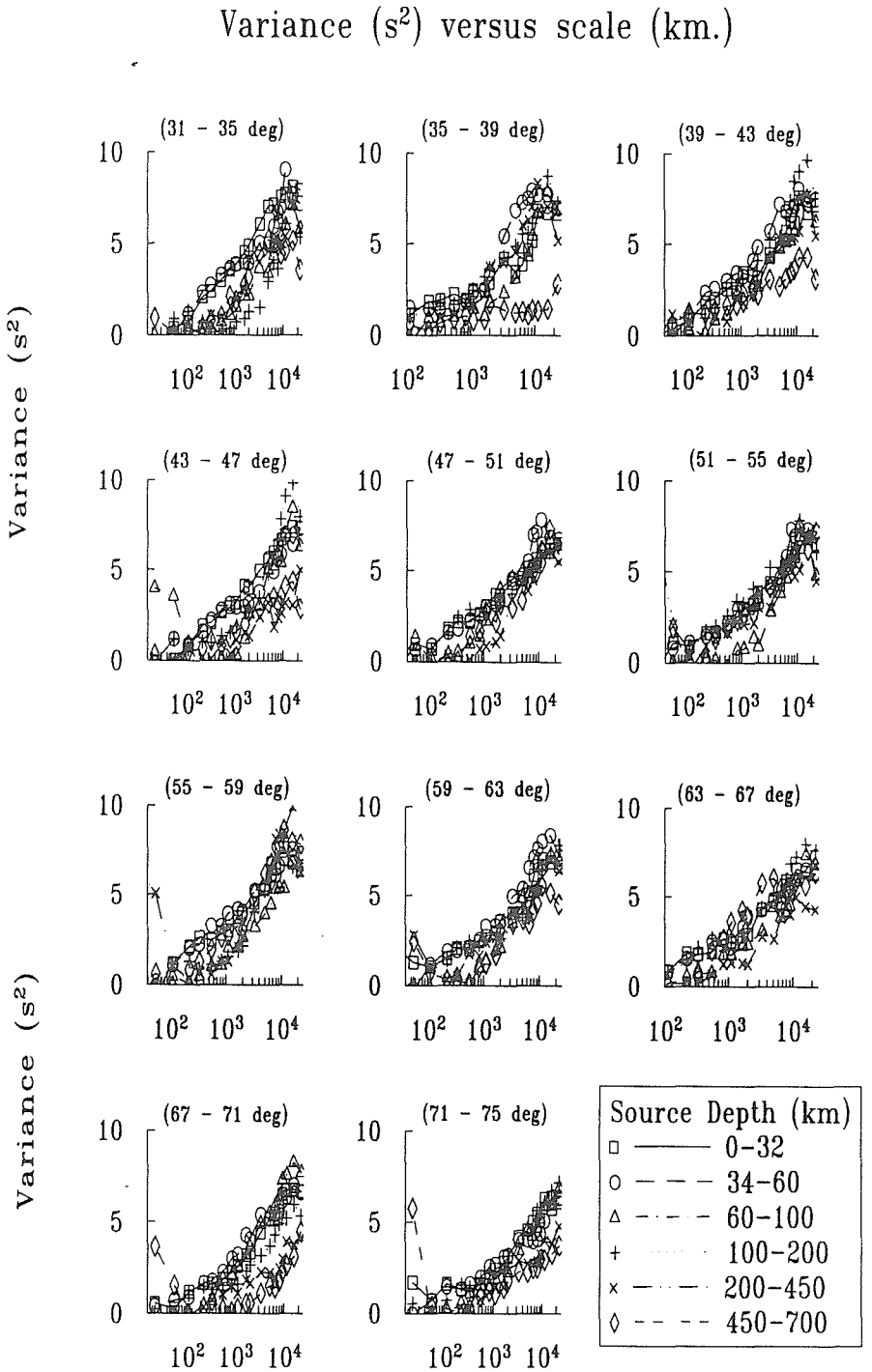


Figure 1.6(c) S-wave data after removal of intercepts in figure 1.6(b).

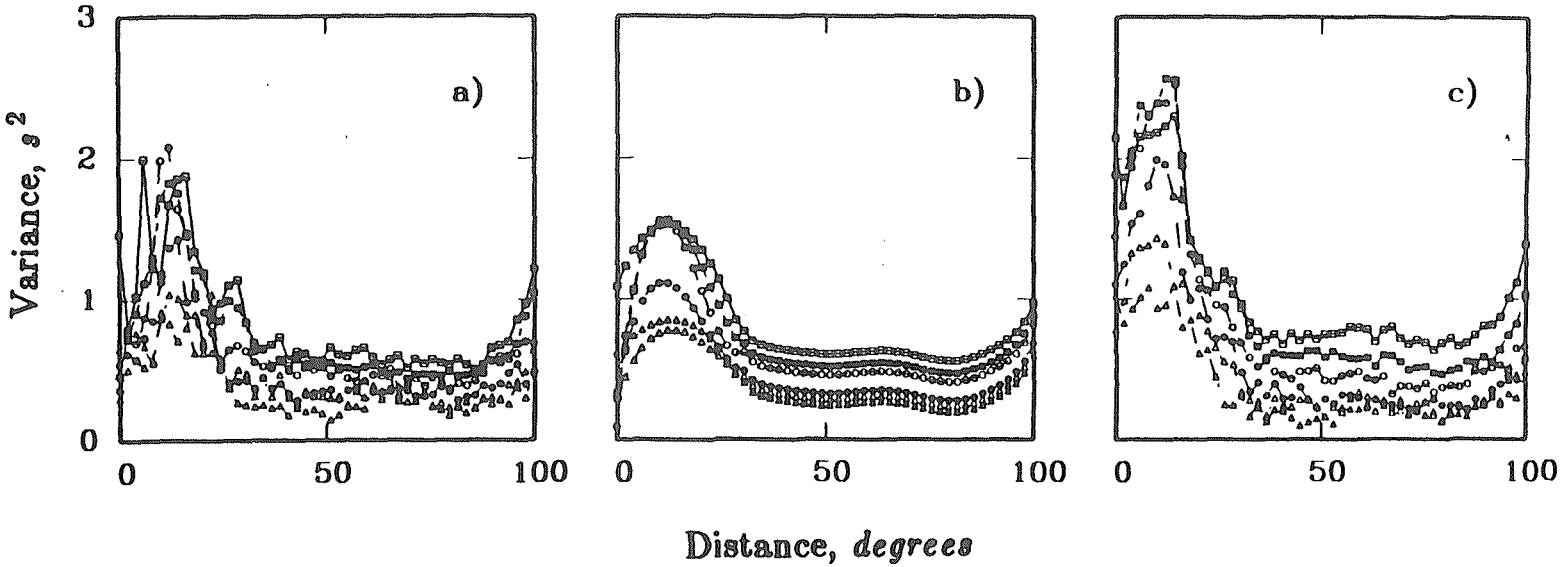


Figure 1.7 P-wave linear intercept estimates. From *Gudmundsson et al.* [1989]. (a) Raw estimates, (b) smoothed (used for inversion) (c) following adjustment to minimize variance.

and the structural parameters; i.e. how much of the derived intercept was related to the real intercept and how much to the real structure, and equally, how much of the real intercept leaked into the estimate of the structural parameters. For this present problem it was decided not to solve the structure and intercept simultaneously. This leads to a much simpler formulation resulting in an easier, more efficient and more robust inversion. The final results show that there is little room to accommodate long-wavelength structural signal in the intercept and vice-versa so for these data-sets it is felt that separating the estimation of model and intercept is warranted and possibly better. The simultaneous solution of intercept and structure is discussed further in the theory section and a method for achieving it is outlined in appendix 1.A.

TABLE 1.1

TABLE OF SYMBOLS	
symbol	meaning
$\bar{}$	Ensemble average
A	Averaging area for summary ray variance
a	Radius of A
α	Correlation distance
B	Sampling auto-covariance function
b_a	Sampling function
D	Wave parameter= 4 S/ k α^2
d	Minimum separation between two rays
δt	Travel-time residual
δU	Slowness perturbations, slowness=1/velocity
Ei()	Integral exponential function
E{} $\langle\rangle$	Expectation of variable over area A
f	Depth correlation; Arbitrary function
ϕ	Longitude
i	Angle of incidence; Imaginary number
k	Wavenumber = 2 π / λ
l	Harmonic Degree
λ	Wavelength; Angular distance between two arbitrary rays
N	Correlation function (general)
P	Probability
P_{lm}	Associated Legendre functions
Θ	Angular radius of A
θ	Colatitude
Q_l	Power spectral coefficients of structure
R	Slowness auto-covariance function
r	Distance from center of Earth
S	Length along ray
T	Travel-time auto-covariance function
W	Ray width
w	Areal weighting function
x	Position vectors
ξ,η	Six component vectors e.g., $\xi = \{\xi_1,\xi_2\}$
$X_{1/2}$	Half-width of R
y,η_1,η_2	Displacement vector
Y_{lm}	Fully normalized spherical harmonics
{ }	Temporal average

1.3 Theory

We will develop a theory to allow us to relate the travel-time variance evaluated at different scales for ray bundles travelling different distances, to the variation of the spectrum of heterogeneity as a function of depth through the mantle. The theory shall first be derived for a cartesian geometry where it is more readily understood, before being transposed to the spherical geometry required for this whole Earth application. The derivation of the relation is in two parts; the variance of summary rays is related to the auto-covariance of the travel-time residual function and then the auto-covariance of the travel-time residual function is related to the statistics of the medium.

Stochastic models are widely used to describe scattering work. Different scattering regimes are illustrated in *Wu and Aki* [1988] in their introduction to a volume dedicated to research on seismic wave scattering. Our problem: body wave propagation through the whole mantle, if considered as a formal scattering problem, potentially covers many scattering regimes. We do not treat our problem as a scattering problem, but due to the affinity of our methodology to scattering work we shall briefly consider the similarities. *Wu and Aki* [1988] present a figure illustrating the limited range of slowness perturbations (2 orders of magnitude) found over a very wide range of scales (10 orders of magnitude). This figure suggests that the spectrum of seismic heterogeneities in the Earth is

nearly white. If this is true, then to get an accurate description of the mantle spectra one will need to be able to assess the whole spectrum.

We shall begin by briefly reviewing previous uses of stochastic structural modeling in seismology so that our application and theory can be placed in context of previous work. Stochastic modeling has been primarily used to study the structure below local arrays, LASA, Montana, [*Aki*, 1973; *Berteussen et al.*, 1975; *Capon*, 1974], NORSAR, Norway, [*Berteussen et al.*, 1975; *Capon and Berteussen*, 1974; *Flatte and Wu*, 1988], Gaurribidanur seismic array, S. India, [*Berteussen et al.*, 1977] and SCARLET, S. California, [*Powell and Meltzer*, 1984]. Most of the above studies compared auto-covariance of phase and log amplitude and the cross-correlation of phase and log amplitude. They then compared their curves with the predictions of *Chernov* [1960] for a Gaussian medium. They obtained estimates of correlation lengths of order 10km and velocity perturbations of order 0.2% to 5% (with Gaurribidanur array being the only one below 2%), extending down to more than 100km from the surface. There is poor control on the trade-off between the velocity perturbations and their depth extent since the data are primarily sensitive to their product.

Frankel and Clayton [1984; 1986]; have used numerical finite-difference studies of synthetic structures to obtain a better understanding of various facets of scattering, including coda generation, scattering versus intrinsic attenuation and auto-correlation of amplitude and phase fluctuations; they were also able to consider the effect of non-Gaussian medium auto-correlation. *Aki and Chouet* [1975], and *Wu and Aki* [1985], both considered the origin of coda

waves by back-scattering from stochastic structures; while *Haddon and Cleary* [1974] have explained the precursors of PKIKP as scattering of PKP waves from stochastic structure near the core-mantle boundary (CMB).

The most widely quoted theory using stochastic structures is that of *Chernov* [1960]. He used perturbation techniques to develop general formulas for the statistics of phase and amplitude fluctuations due to propagation through media with weak fluctuations in refractive index. These included expressions for auto-covariance of phase and amplitude fluctuations, transverse and longitudinal auto-covariance of phase and amplitude; and their cross-correlation. He evaluated his expressions for a Gaussian auto-correlation of the refractive index. The assumptions underlying Chernov's acoustic theory are listed in table 1.2.

TABLE 1.2.

Primary assumptions of Chernov's theory.	
1	small velocity variations, i.e. $\Delta v << v$
2	statistically homogeneous and isotropic
3	travel path through random media > correlation distance
4	correlation length > wavelength
5	Born approx., most energy remains in primary wave, over the whole path through the random medium.
6	Gaussian correlation function

Chernov also uses the Rytov approximation, which allows him to relax the 5^{th} condition; such that most of the energy must remain in the primary wave over a wavelength and not over the whole path. Our results suggest that these conditions generally don't

apply in the Earth. Our ray theory does not require the 4^{th}, 5^{th} or 6^{th} condition.

Developments on the work of Chernov include Tatarski's application of similar theory to the effect of isotropic turbulence on wave propagation [*Tatarski*, 1961]. *Wu and Aki* [1985] extended Chernov's scalar acoustic analysis to the case of elastic wave scattering in a random media. *Aki* [1969], and *Aki and Chouet* [1975] have applied the single scattering theory of scalar waves to modeling coda generation of local earthquakes. *Gao et al.* [1983a; 1983b] has extended the isotropic single scattering to an isotropic multiple scattering model. *Haddon and Cleary* [1974] extended Chernov's theory to apply to small random variations in density and elastic parameters in a restricted volume in an otherwise spherically symmetric Earth model.

In the above applications, the statistics of the random media were uniform in depth. *Flatte and Wu* [1988] used the parabolic wave equation to obtain information on the depth extent of the variation. They realized that looking at the coherence of arrivals arising from plane waves approaching arrays at different angles of incidence allowed them to obtain this information. The parabolic wave equation model's energy in a narrow cone around the direction of propagation. *Flatte and Wu* [1988] introduced angular correlation functions, which relate the auto-covariance of phase and amplitude and the cross-correlation of both, as a function of the separation in the angle of incidence. This is related to the medium by an integral through depth, where the separation between the rays increases with depth like $Z\theta$, where Z is the depth and θ is the difference in the

angle of incidence. Using the parabolic wave equation approximation allows them to reduce Chernov's volume integrals to line integrals and hence makes it feasible to incorporate along path variations in statistics.

Our theory is fundamentally different from the above in that it is based on ray theory and not wave theory. It is not applied to wavefronts. Since ray theory is the high frequency limit of wave theory our expression for the relationship of stochastic structure to the travel-time residual auto-covariance should and does correspond to that limit of Chernov's theory. It is similar to *Flatte and Wu*'s [1988] theory in that its simplicity in modeling wave propagation gives us greater flexibility in allowing variations of the statistics along the ray path. The first order parabolic wave equation is valid in a 15^{o} cone around the propagation direction; ray theory can be considered a zeroth order parabolic wave equation valid only along the ray path. Since we are only considering the first arrival we needn't worry about multiple scattering because this energy arrives later. Equally, our data did not involve wave-fronts or fitting best plane waves and hence the residual represents signal collected all along the ray path from the source to the receiver. The model is global and not restricted to the upper mantle beneath arrays or small regions deep in the mantle. Our array can be thought of as consisting of all the stations reporting to ISC and all the major events occurring between 1964 and 1986.

Since parameterizing the medium stochastically is novel in global seismology and as illustrated by the brief review above has not been a common tool for seismologists, we would like to introduce

the reader to the philosophy and some of the terminology of this parameterization. If you are familiar with the following terms: auto-covariance, homogeneous and isotropic statistics, expectation, variance, statistical independence, and are comfortable with parameterizing the Earth as a stochastic medium, then you can skip the following section and continue after the heading, "Comparison of Wave and Ray Formulation".

1.3.a Stochastic philosophy and terminology

A stochastic (statistical) model restricts itself to obtaining a description of the statistics of the model parameters rather than the model parameters themselves. Statistics involves taking averages of quantities or evaluating the probability that certain quantities will have certain values. For clarity we shall discuss these statistical concepts for a scalar function of one variable; e.g., the Dow Jones index as a function of time. Later we shall discuss the extensions to a scalar function of many variables; e.g., isotropic velocity as a function of spatial position. When we consider a value that varies over time e.g., the Dow Jones index; this is usually done discretely by sampling the quantity over time and evaluating an average. Continuously we might define a temporal average as follows

$$\{f(t)\} = \int_{-\theta}^{+\theta} f(t+t')dt' \Big/ \int_{-\theta}^{+\theta} dt' \tag{1.3}$$

Where { } is used to express an average over the independent variable.

More generally, statisticians consider an "ensemble" of similarly prepared systems, and average over the "ensemble." Hence

$$\overline{f(t)} = \frac{1}{N}\sum_{k=1}^{N} f^k(t) \tag{1.4}$$

where the "ensemble" average is denoted by the overbar. Where $f^k(t)$ is the value of f(t) in the k^{th} system of the "ensemble" and where N is the very large number of systems in the "ensemble."

If the process is stationary with respect to f, then this means that there is no preferred origin in time for the statistical description of f. The "ensemble" average and time average are equivalent if the function $f^k(t)$ for each system of the "ensemble" will, after a finite time, pass through all the values accessible to it (the ergodic assumption).

This methodology transfers directly when the independent variable is a vector rather than a scalar. Hence the spatial average is

$$\{f(\mathrm{x})\} = \iiint f(\mathrm{x}+\mathrm{x}')dV \,/\, \iiint dV \tag{1.5}$$

While the "ensemble average" is

$$\overline{f(\mathrm{x})} = \frac{1}{N}\sum_{k=1}^{N} f^k(\mathrm{x}) \tag{1.6}$$

There is a twist in discussing the equivalent term to stationarity because of the change in the independent variable from a scalar to a vector. Firstly, a field is spatially homogeneous if its mean value is constant and if its auto-covariance is unchanged from point to point for the same displacement vector. By auto-covariance (also known as correlation function) we mean

$$N(\mathbf{x}_1,\mathbf{x}_2)=\overline{[f(\mathbf{x}_1)-\overline{f(\mathbf{x}_1)}][f(\mathbf{x}_2)-\overline{f(\mathbf{x}_2)}]} \tag{1.7}$$

Hence homogeneity implies

$$N(\mathbf{x}_1,\mathbf{x}_2)=N(\mathbf{x}_1+\mathbf{x}_0,\mathbf{x}_2+\mathbf{x}_0) \tag{1.8}$$

by setting $\mathbf{x}_0 = \mathbf{x}_2$ we get

$$N(\mathbf{x}_1,\mathbf{x}_2)=N(\mathbf{x}_1-\mathbf{x}_2,0) \tag{1.9}$$

Hence in an homogeneous field the auto-covariance depends only on $\mathbf{x}_1$-$\mathbf{x}_2$. An homogeneous field is isotropic if N(x) depends only on |x|, i.e. only on the distance and not on the direction joining two points, i.e.

$$N(\mathbf{x}_1,\mathbf{x}_2)=N(|\mathbf{x}_1-\mathbf{x}_2|) \qquad \forall \mathbf{x}_1,\mathbf{x}_2 \tag{1.10}$$

What does equation 1.6 mean? What is the averaging over? This is the crux to being comfortable with the philosophy of representing the medium by a statistical parameterization. Equation

1.6 is defined in terms of an ensemble average. These are an assembly of realizations of different possible states of the medium. Given that the Earth, seismically over the time frame we're considering, has only one realization then it would seem that we can't define an "ensemble" average. We shall make the assumption that the field is laterally homogeneous and isotropic. Radially we shall allow it to vary, so that

$$N(\mathrm{x}_1) = N(r_1, \theta, \phi) \neq N(r_2, \theta, \phi) \tag{1.11}$$

We can then assume the ergodic hypothesis and reformulate the averaging spatially rather than "ensemble" wise.

$$N(\mathrm{x}_1, \mathrm{x}_2) = \{[f(\mathrm{x}_1) - \{f(\mathrm{x}_1)\}][f(\mathrm{x}_2) - \{f(\mathrm{x}_2)\}]\} \tag{1.12}$$

where we have assumed that x_1 and x_2 have the same radial co-ordinate. Hence we are averaging over the whole volume of the layer. In this sense we are admitting that we won't know the exact value at any point, but we will know the probability of its fluctuation away from the mean. Once we accept this viewpoint we can develop the "ensemble" formulation also; i.e. all the realizations are all the values at a given radial depth in the Earth.

Hence, the fundamental tenet of this methodology is that we admit that we will get no information as to the lateral location of seismic velocity variation; only the most probable value and the probability for fluctuations away from this mean. Consequently in

the analysis of data we shall not care about absolute locations of events and stations only on their relative separation.

We have already defined the following statistical concepts, stationarity, ergodicity, homogeneous and isotropic statistics, auto-covariance, mean and "ensembles." Four further concepts that will be used are now discussed.

The auto-correlation (also known as the correlation coefficient) is the ratio of the auto-covariance to the mean square fluctuation (the auto-covariance with zero separation) N(0), i.e. a normalized auto-covariance. Hence

$$B(|\mathbf{x}|) = \frac{N(|\mathbf{x}|)}{N(0)} \tag{1.13}$$

Since N(0) is the maximum value of N(x), it implies that $B(x) \leq 1$.

The expectation of a random function f(x), which is a function of a variable x, is defined as follows

$$\langle f(x) \rangle = E\{f(x)\} = \int_{\forall f} f(x) P(f) df \tag{1.14}$$

where < > represents the expectation operator. P(f) is the probability that f(x) has the value f. For instance if f is as likely to be +f as -f then <f(x)> = 0. This is a third mode of averaging and is in the spirit of an "ensemble" average since the behavior of the different systems of the "ensembles" define the probability.

By "variance of a collection of data", we mean the L_2 scatter around the mean, defined as follows

$$\sigma^2(d) = \frac{\sum_{i=1}^{N} (d_i - \bar{d})^2}{M} \tag{1.15}$$

where M=N if we have independent information as to the value of the mean, while M=N-1 if we need to use the data to define the mean. The standard deviation is the square root of the variance. Note $\sigma^2(d) \geq 0$.

The final concept to be discussed is statistical independence. Two random variables are independent if the probability of one variable having a certain value is independent of the value of the other variable, i.e.,

$$P(x,y) = P(x)P(y) \tag{1.16a}$$

$$P(x|y) = P(x) \tag{1.16b}$$

Where P(x,y) is the probability of the first variable having a value x and the second variable having a value y. P(x|y) is the probability of the first variable having a value x conditional on the second variable having a value y.

1.3.b Comparison of ray and wave formulations

As already mentioned, our theory is developed in the ray-limit; i.e. the high-frequency, short wavelength limit. We shall now briefly discuss the implications of this approximation and associated topics. The response at a point depends upon information the wave has collected along its many paths to the point, but only energy scattered from the first Fresnel zone arrives in phase and interferes constructively. As described by *Nolet* [1987] p.10, this can be used to define an effective ray-width given by

$$\left[w(s)^2+s^2\right]^{1/2}+\left[w(s)^2+(L-s)^2\right]^{1/2}-L=\lambda/4 \qquad (1.17)$$

where w is the ray width, s is the distance along the ray from the end-point, L is the total ray length, and λ is the wavelength.

The resolution of seismic data restricts information, at a point in the Earth, to a spatial average of $\approx$ the ray width around this point. As known from seismic reflection work, it is possible to improve this resolution and to approach the theoretical resolution limit of a fraction (1/4 to 1/2) of a wavelength [*Lindsey*, 1989]. This can only be achieved by disentangling the diffraction. Migration attempts to do this. It's commonly done by running the observed wave-field backwards using an approximation to the wave equation and using an imaging condition [*Claerbout*, 1985]. Since we do not consider wave theory directly, our resolution is proportional to the ray "width" rather than wavelength.

If the scatterers are smaller than the ray-width then the scattered energy is incoherent and destructively interferes at the point. Hence the phase of the arrival is largely unchanged while the energy of the arrival is reduced since the energy, rather than the amplitudes, is additive in this case. By contrast, if the ray-width is smaller than the scatterer then the energy is scattered coherently from the scatterer and constructively interferes at the receiver. This leads to the phase being advanced or delayed depending upon whether the heterogeneity is slower or faster than average respectively. Also, the amplitudes are additive and hence this scattering leads to a prominent arrival. The net effect is that for volumes smaller than the ray-width we will smooth the velocity variations; hence our spectra will have a short-wavelength taper that will vary with depth through the mantle.

Another viewpoint from which to discuss the effects of diffraction on travel-times is that of propagating wave-fronts. Consider a spherical shell propagating away from a source. As this wavefront encounters heterogeneity, it will develop a rougher surface and will scatter energy. If sufficient energy is scattered out from the wavefront, it can lead to part of the wavefront being disconnected from its parent wavefront. If the energy is scattered ahead then the responsible heterogeneity will only be imaged if the advanced wavefront has sufficient energy to generate an arrival above the noise level. Equally, a second arrival could be observed if the parent wavefront "heals" sufficiently. If the scattered energy is delayed, then it is possible that the parent wavefront can "heal" and make an arrival above the noise level and hence hide the delayed

wavefront that sampled the heterogeneity. From this discussion it is clear that the observation of scattered energy not only depends on the scale of the heterogeneity and its magnitude, the wavelength of the energy, and the path length, but also on the ambient noise. This effect has been investigated by *Wielandt* [1987], who suggested that due to wave-front healing the medium will seem faster than it actually is. His study involved abrupt velocity contrasts and ignored noise. More realistic conditions, I believe, will lead to the conclusion that one can be biased in both directions. In fact, since high-velocity regions are anti-waveguides, while low velocity regions trap waves, it is possible that the opposite of the "Wielandt effect" applies; i.e. the Earth seems slower than reality, since more scattered energy is slowed by slower heterogeneities than is advanced by faster heterogeneities. In this scenario, the parent wavefront might heal for the faster heterogeneity but might not heal sufficiently to make an observed arrival for the slower heterogeneity. Hence, the delayed arrival due to the slow heterogeneity would be observed but the parent wavefront rather than the advanced arrival would be observed for the fast heterogeneity. This point is worthy of further investigation, but presently we shall assume that the "Wielandt effect" does not exist and that we are modeling a smoothed version of the real Earth.

How do the ray-width and wavefront viewpoints compare? They are similar. In the wavefront perspective, small features will not put much energy into the scattered waves, hence it is probable that we will see the averaged value as contributed by all the neighboring heterogeneities, as expected from the ray-width per-

spective. Since we presently don't have a detailed assessment of the wavefront viewpoint, we shall assume that the ray-width criterion is reasonable for defining "small." The wavefront discussion illustrates that this is simplistic, and for large-scale, large magnitude variations it may be inaccurate.

In another sense they seem different. The ray-width is smallest at both the receiver and event and is largest at the mid-point. The process of wavefront "healing" suggests from its terminology that it is something that progresses as the wavefront propagates, i.e. wavefront disruptions are healed by diffractions from the parent wavefront, and to fill in the whole gap requires a certain propagation distance. Hence, disruptions closer to the receiver have less time to "heal" and we are more likely to observe the scattered energy. At first glance it seems that we are more sensitive to near receiver structure in this formulation, and least sensitive to near source structure. Actually, this seeming difference is only apparent and "healing" is an unfortunate word. Small heterogeneities near the source have a larger effect (are more "damaging"!) since the wavefront is curved and has less surface area. By contrast, near the receiver the wavefront is planar and the energy is lower since the surface of the wavefront is much larger. Therefore, even though there is longer time for wavefront healing for near source disruptions, similar-sized heterogeneities produce much larger disruptions to the wavefront near the source as opposed to near the receiver. Hence, near source and near receiver structures have similar effects; as expected from the ray-width description.

In Chernov's 1960 wave theory of scattering, diffraction is included, and is parameterized by means of the wave parameter D.

$$D = 4S / k\alpha^2 \tag{1.18}$$

Where k is the wavenumber $k=2\pi/\lambda$; α is the correlation length and S is the total path length. D=0 is the ray limit while $D=\infty$ is the Fraunhofer diffraction limit. The wave parameter relates the radius of the 1st Fresnel zone $(S/k)^{1/2}$ to the radius of the scatterer $\alpha/2$. *Chernov* [1960] finds that if D=0; the phase fluctuations will be twice as large as if $D=\infty$, for a Gaussian correlation function. Hence, with increased diffraction the fluctuations are reduced, as the above discussion would have suggested. If the heterogeneities are very much smaller than the wavelength e.g., geologic grains compared to seismic waves; the very heterogeneous medium can be replaced by an effectively homogeneous medium, see *Aki and Richards* [1980] p.748.

As a result of arguments presented in the above discussion, we believe that applying ray theory to travel-time data obtained from finite-frequency seismograms leads to a fair description of the medium, provided we realize that we should smooth the short-scale features in the medium.

1.3.c Derivation of Ray Theory

This following derivation is an extended and slightly changed version of the derivation previously presented in *Gudmundsson et al.* [1989].

Consider that the Earth contains small random slowness perturbations $\delta U(x)$. Slowness is the reciprocal of velocity and $\delta U = -\delta V/V^2$. Let $R(x,y)$ be the auto-covariance of δU. x is the position within the Earth while y is the separation between the two points. If we assume that $\delta U(x)$ is laterally homogeneous and isotropic; then $R(x,y) = R(r,y)$, where r is the distance from the center of the Earth and $y = |y|$.

The travel-time residual δt, is the difference between the observed travel-time and the predicted travel-time from the spherically symmetric Earth model (in our case the Jeffereys-Bullen velocity model). The travel-time residual δt is related to the random slowness perturbations $\delta U(x)$ by a line integral along the ray-path as defined by the reference Earth model. The δt form a random function that is a function of both source and receiver location, $\delta t(\xi_1,\xi_2)$, where ξ_1 is the position vector of the receiver and ξ_2 is the position vector of the source. Let $T(\xi_1,\xi_2,\eta_1,\eta_2)$ be the auto-covariance of $\delta t(\xi_1,\xi_2)$, where η_1 is the lateral displacement vector of the receivers and η_2 is the lateral displacement vector of the sources. Since δU is laterally homogeneous and isotropic, we can assume that $\delta t(\xi_1,\xi_2) = \delta t(\xi)$ (where ξ is a six component vector $\{\xi_1,\xi_2\}$) is also laterally homogeneous and isotropic; provided the mapping is uniformly random. Therefore $T(\xi_1,\xi_2,\eta_1,\eta_2)= T(\eta_1,\eta_2)= T(\eta)$ where η is a six-component

vector $\{\eta_1,\eta_2\}$. Given these properties of the travel-time residual, we need to estimate the scatter in δt within a sample from a circular area $A = \pi a^2$; i.e. the variance of the residuals within A referenced to the mean residual for the area.

$$\begin{aligned}\sigma^2(a) &= E\{(\delta t(\xi) - \overline{\delta t(a)})^2\} \\ &= E\{(\delta t^2(\xi)\} - E\{(\overline{\delta t}^2(a)\}\end{aligned} \tag{1.19}$$

E{ } is the expectation over area A. Since δt is laterally homogeneous and isotropic, $E\{\delta t^2(\xi)\}$ over A is equivalent to $E\{\delta t^2(\xi)\}$ over all space. The first term in equation 1.19 is simply the global variance of δt, i.e.,

$$E\{\delta t^2(\xi)\} = T(0) \tag{1.20}$$

The second term in equation 1.19 is the variance of the mean

$$\begin{aligned}E\{\overline{\delta t}^2(a)\} &= E\{(\frac{1}{\pi a^2}\int_A b_a(\xi)\delta t(\xi)d\xi)^2\} \\ &= \frac{1}{\pi^2 a^4}\iint_A E\{\delta t(\xi_1)\delta t(\xi_2)\}E\{b_a(\xi_1)b_a(\xi_2)\}d\xi_1\, d\xi_2 \\ &= \frac{1}{\pi^2 a^4}\iint_A T(|\xi_1 - \xi_2|)B_a(a,|\xi_1 - \xi_2|)d\xi_1\, d\xi_2\end{aligned} \tag{1.21}$$

b_a is a scaled sampling function, an irregular 2-D pin cushion of delta functions.

$$b_a(\xi) = \frac{\pi a^2}{N} \sum_{i=1}^{N} \delta(\xi_i) \tag{1.22}$$

where N is the number of measurements within the area A and ξ_i are the sample positions. B(a,ρ) is the auto-covariance of the scaled sampling function.

To separate b_a and δt from within the same expectation operator requires that they are statistically independent. That is that sources and receivers are not correlated to structure. This assumption is suspect for subduction zone events and also for most receivers, since they are on land, and continents are generally fast compared to oceans.

If the global mean of δt is not zero, then equations 1.20 and 1.21 both have $\overline{\delta t}^2$ on the right-hand side; hence $\overline{\delta t}^2$ cancels from both sides of the expression. Of course, if the spherically symmetric reference model is correct, $\overline{\delta t}$=0.0, and can be ignored.

To evaluate the integral of equation 1.21, let us look at the geometry in figure 1.8. Let the point ξ_1 = P be a distance s from the center of A. Consider all points ξ_2 a distance ρ from P, within A. If ρ < (a-s) the points ξ_2 span a whole circle. While if (a-s)<ρ<(a+s) the angle ϕ is excluded. The cosine rule gives

$$a^2 = s^2 + \rho^2 - 2s\rho\cos\theta \tag{1.23}$$

Hence, the points ξ_2 span an arc of length θ = π.; if 0<ρ<a and 0 < s < (a-ρ), while they span an arc length

$$\theta = \cos^{-1}\left(\frac{s^2+\rho^2-a^2}{2\rho s}\right) \tag{1.24}$$

if $0<\rho<a$ and $(a-\rho)<s<a$ or if $a < \rho < 2a$ and $(\rho-a)< s< a$. Because δt is isotropic, T is a function of ρ only. If we assume that the sampling density is isotropic, then we can rewrite equation 1.21 by

$$\begin{aligned} E\{\overline{\delta t}^2(a)\} &= \frac{1}{\pi^2 a^4}\left[\int_0^a 2\pi T(\rho)B(a,\rho)d\rho \int_0^{a-\rho} 2\pi s ds\right] \\ &+ \frac{1}{\pi^2 a^4}\left[\int_0^a 2\pi\rho T(\rho)B(a,\rho)d\rho \int_{a-\rho}^{a} 2\cos^{-1}\left(\frac{s^2+\rho^2-a^2}{2\rho s}\right) s ds\right] \\ &+ \frac{1}{\pi^2 a^4}\left[\int_a^{2a} 2\pi\rho T(\rho)B(a,\rho)d\rho \int_{\rho-a}^{a} 2\cos^{-1}\left(\frac{s^2+\rho^2-a^2}{2\rho s}\right) s ds\right] \end{aligned} \tag{1.25}$$

which simplifies after tedious algebra to

$$E\left\{\overline{\delta t}^2(a)\right\} = \int_0^{2a} w(a,\rho)T(\rho)B(a,\rho)d\rho \tag{1.26}$$

where

$$w(a,\rho) = \frac{4}{\pi a^2}\left[\cos^{-1}\left(\frac{\rho}{2a}\right) - \frac{\rho}{2a}\left(1-\left(\frac{\rho}{2a}\right)^2\right)^{1/2}\right]\rho \tag{1.27}$$

Let us consider a few cases. We shall assume that the sampling is uniform, i.e. B=1. First, consider that the travel-time residuals δt, is a non-zero constant δt_0, over the area A. Then the auto-covariance, $T(\rho) = T_0$, a constant within the area A, since $\overline{\delta t} = 0.0$. The variance of the mean over this area will also be T_0, since the

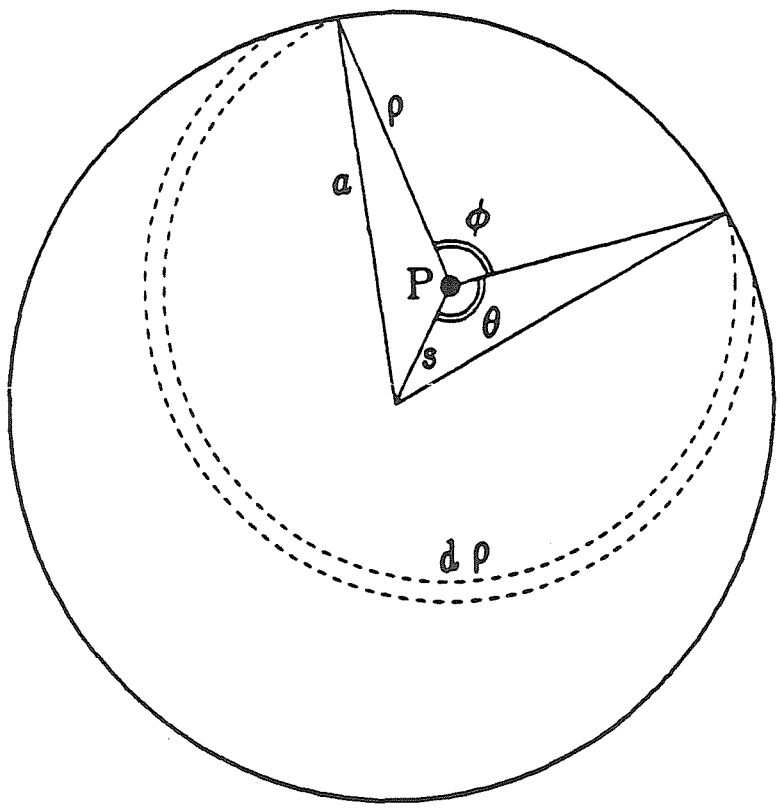


Figure 1.8 Geometry for evaluating the relationship of variance of travel-time residuals to auto-covariance of travel-time residuals in cartesian geometry. From *Gudmundsson et al.* [1989].

mean is obviously δt_0. The global variance of the travel-time T(0) will also be T_0. Hence $\sigma^2(a) = 0.0$ as expected. Consider a second example where the δt are completely uncorrelated. Then $T(\rho) = T_0\delta^2(\rho)$, where δ^2 represents the 2-D Dirac delta function. Now the variance of the averages decreases as the area increases, since they are uncorrelated. In fact the variance of the averages decays as the inverse of the area. Since the correlation function is a Dirac delta function the variance of the averages is zero for any finite area; and so can't decrease. The variance $T(0) = T_0$. Hence $\sigma^2(a) = T_0$, for any finite area. Third, consider an intermediate case where $T(\rho)= T_0f(\rho)$, where the auto-correlation function $f(\rho)$ decreases monotonically. The average of the mean decreases as area A increases; the greatest change will occur when the scale of the area, a, is similar to the correlation length, α.

Figure 1.9 shows some examples of the weighting functions of equation 1.23, $w(a,\rho)$. They are zero at their endpoints (0,2a) and their integral = 1.0. Notice that their maximum is just below the center of their domain. Hence, in averaging over a circle of radius a, the main contribution to the variance comes from a scale of length $\rho \approx a$.

By combining equations 1.19, 1.20 and 1.26 we get

$$\begin{aligned}\sigma^2(a) &= T(0) - \int_0^{2a} w(a,\rho)T(\rho)B(a,\rho)d\rho \\ &= \int_0^{2a} w(a,\rho)(T(0) - T(\rho))B(a,\rho)d\rho \end{aligned} \tag{1.28}$$

since

$$\int_0^{2a} w(a,\rho)B(a,\rho)d\rho = 1 \tag{1.29}$$

because wB is the probability density of sample spacing within a circular area of radius a and hence must equal unity by definition. Note that the integral of $w(a,\rho)B(a,\rho)T(0) = T(0)$ only because we have assumed that $\delta t(\xi)$ is laterally homogeneous and isotropic.

We now have an expression relating the variance within summary rays of scale a, $\sigma^2(a)$, to the auto-covariance function of the travel-time residuals field, $T(\rho)$. We need to express the travel-time auto-covariance in terms of the medium auto-covariance R(r,y). To do this we must idealize the summary ray geometry.

Since we bin the data finely in epicentral distance and source depth, all rays contributing to a particular datum have very similar ray parameters. Hence, we argue that the rays can be approximated as being parallel and shifted horizontally, relative to each other. This is the most questionable approximation; for the intermediate scales our idealization is good, while at the largest and very smallest scales it is obviously not so good. In general it is probable that we have rays that are diverging, converging and also crossing. At the smallest scales there is a tendency for the summary rays to correspond to a few stations and a large number of events. Hence, the resulting geometry is fan-like and we would expect these rays to be less sensitive to structure near the source. In both studies, we have only included data that does not suffer badly from this effect. At the largest scales of course, e.g., the whole Earth, it is very difficult to think of approximating the geometry of all those rays by two

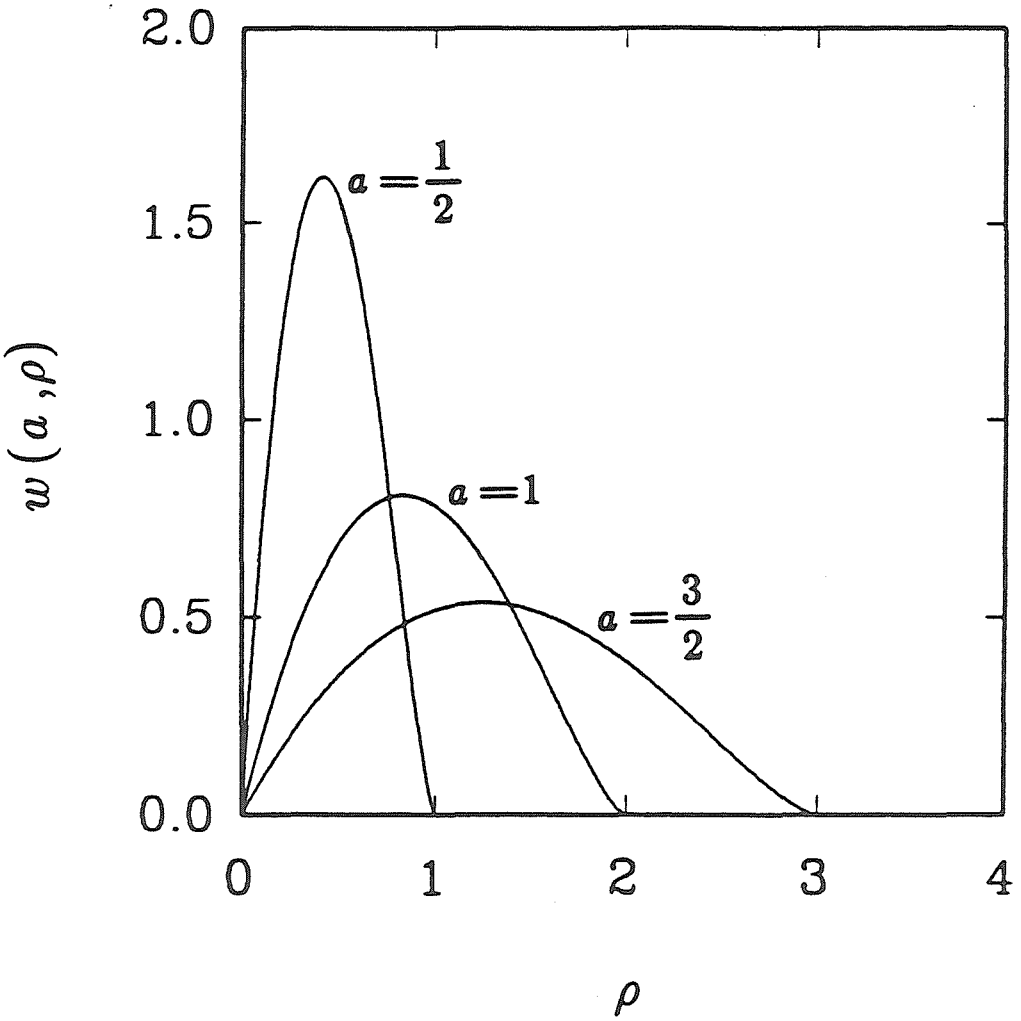


Figure 1.9 Examples of the weighting functions w(a,ρ). Cartesian geometry. Notice peak of kernel is at ≈ a. From *Gudmundsson et al.* [1989].

parallel shifted rays. Fortunately it is likely that this result is also close to the result for a random distribution of rays when all the ray-paths are averaged together. The exact effect of this idealization needs to be investigated but it is clear that we are making something close to the most reasonable approximation.

From Fermat's principle we have that

$$\delta t(\xi) = \int_{ray} \delta U(x) dS \tag{1.30}$$

Hence,

$$\begin{aligned} T(\rho) &= E\left\{\delta t(\xi_1)\delta t(\xi_2)\right\} \\ &= \int_{ray1}\int_{ray2} E\left\{\delta U(x_1)\delta U(x_2)\right\} dS_1 dS_2 \\ &= \int_{ray1}\int_{ray2} R(|x_1 - x_2|) dS_1 dS_2 \end{aligned} \tag{1.31}$$

If the correlation length is small compared to the radius of curvature of the rays, then we can think of the rays as being locally straight and the auto-covariance function as locally constant. If ρ is the horizontal distance that the rays are displaced relative to each other, i is the angle of incidence and ϕ is the angle that the ray displacement vector makes with the ray azimuth, then the minimum distance between the two rays is (see figure 1.10)

$$d = \sqrt{\cos^2(\phi)\cos^2(i) + \sin^2(\phi)} \tag{1.32}$$

Consider point P_1 on ray 1 at Z, then we can relate τ (the distance from P_1 to a point on ray 2) to d (the minimum distance between the

rays) and to S [the path length along ray 2 as measured from P_2 (the point on ray 2 closest to ray 1)] as follows

$$\tau = \sqrt{d^2 + S^2} \tag{1.33}$$

and hence

$$|S| = \sqrt{\tau^2 - d^2} \tag{1.34}$$

and

$$|dS| = |dS_2| = \frac{\tau d\tau}{\sqrt{\tau^2 - d^2}} \tag{1.35}$$

The integral over dS_2 (ray 2) in equation 1.31 can now be translated into an integral over τ, taken from d to ∞., which gives us

$$T(\rho) = 2\int_{ray1} \int_{d}^{\infty} \frac{\tau R(\tau) d\tau}{\sqrt{\tau^2 - d^2}} \tag{1.36}$$

From equation 1.32 we see that d depends on the azimuth of the ray displacement vector. If i = 90^o, we have $d \approx \rho$. The deviation of d from ρ is greatest when i=0^o, (the turning point); where $d=|\sin\phi|\rho$, which on average is $2\rho/\pi$. Hence, d does not vary greatly and we shall approximate d by ρ.

If we assume that the correlation function of the medium is Gaussian i.e.

$$R(\tau) = R_0 e^{-r^2/a^2} \tag{1.37}$$

and assume that the correlation function is isotropic i.e. it's the same

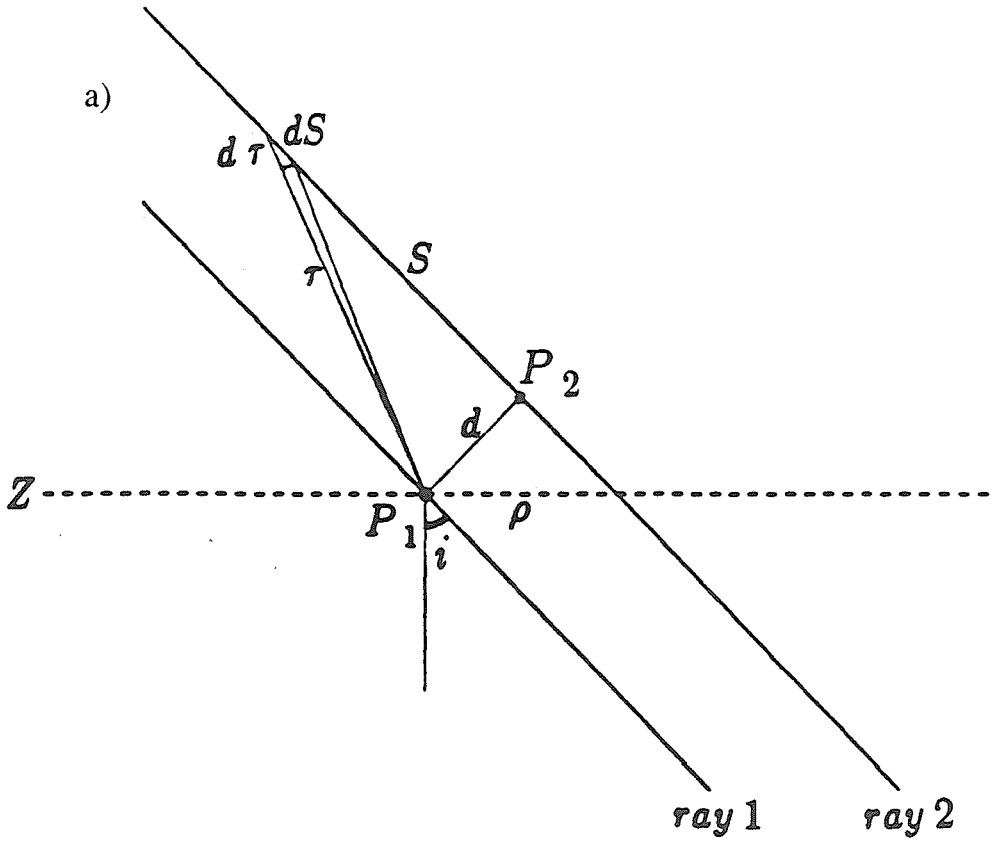


Figure 1.10 (a) Cross-section of ray geometry used in relating travel-time auto-covariance to slowness auto-covariance.

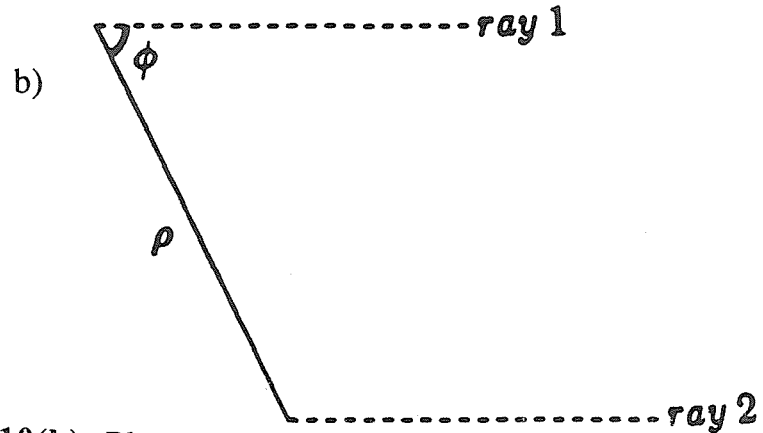


Figure 1.10(b) Plan view of ray geometry. From *Gudmundsson et al.* [1989].

along the ray as it is horizontally, at least for a few correlation lengths, then equation 1.31 becomes

$$T(\rho) = 2\int_{ray}\int_{\rho}^{\infty} R_o \frac{\tau e^{-\tau^2/\alpha^2} d\tau}{\sqrt{\tau^2-\rho^2}}\, dS$$
$$= \sqrt{\pi}\int_{ray} R_0 \alpha e^{-\tau^2/\alpha^2} dS \qquad (1.38)$$

This result is a good approximation for many choices of auto-correlation functions, provided they do not have strong side lobes; i.e., the medium must not have strong periodicity. If the medium was periodic, then one would expect the variance to oscillate as the scale decreased, rather than to decrease monotonically. The variance is not strictly monotonic, but within the error estimates (estimated from the scatter between the variance estimates from the individual summary rays contributing to a given reference ray) the data do not suggest periodicity. Hence, if present it is weak in comparison to other patterns of heterogeneity. One exception occurs at the very largest scale where variance at the very largest scale is actually lower than the scale of the next one or two scales. This could be a reflection of a periodic system with a very long wavelength, order $l = 1$ or 2. We suggest that a more likely explanation of this feature is the poor global sampling given by the present-day distribution of stations and sources. This needs to be confirmed. Equation 1.38 can be rewritten to give

$$T(\rho) \approx C\int_{ray} x_{1/2} R(\rho) dS \qquad (1.39)$$

where $x_{1/2}$ is the half-width of $R(\tau)$, ($R(x_{1/2})=0.5R(0)$), and C is a constant $\approx$ 2. Equation 1.39 holds exactly for the Gaussian auto-correlation function and $C=\sqrt{\pi/\ln 2}$ = 2.13.

If R_o and α are constant along the ray-path we get

$$T(\tau)=\sqrt{\pi}R_o\alpha S e^{-\rho^2/\alpha^2} \tag{1.40}$$

This can be compared with the transverse auto-covariance of phase as derived by *Chernov* [1960]. The phase correlations can be converted to travel-time correlations assuming $\Delta\phi = \omega\Delta t$; which is true provided the medium is non-dispersive (which holds well for body-waves). Hence, the equivalent expression from *Chernov* [1960] is

$$T(\rho)=\frac{\sqrt{\pi}}{2}\left(e^{-\rho^2/\alpha^2}+\left[\frac{1}{2iD}\left[Ei(\xi_2)-Ei(\xi_1)\right]\right]\right)R_o\alpha S$$
$$\text{where } Ei(x)=\int_0^x\frac{e^z}{z}dz\ ;\ \xi_2=\frac{i\rho^2/2\alpha^2}{D+i}\ ;\ \xi_1=\frac{i\rho^2/2\alpha^2}{D-i}\ . \tag{1.41}$$

which reduces to our result in the high-frequency limit of D=0 and ka>>1, since

$$Ei(\xi_2)-Ei(\xi_1)=2iDe^{-\rho^2/\alpha^2}\left[1-D^2/3(\ )+...\right] \tag{1.42}$$

In evaluating the transverse correlation he needs to specify a constant Gaussian auto-correlation for the medium. Hence

his wave theoretical result cannot be applied in the case of a variable along-path non-Gaussian auto-correlation. His result shows though that given a non-negligible D, we need to consider diffraction effects and wavefront healing. I repeat that I believe we can apply our theory, provided we realize that we have intrinsic smoothing over a scale, up to a volume similar to the ray-width. If the medium was constant and Gaussian, Chernov's wave theoretical result can be approximated to

$$T(\rho)=\frac{\sqrt{\pi}}{2}R_o\alpha S\left[e^{-\rho^2/\alpha^2}+\left[\frac{\alpha^2}{\rho^2 D}\sin\left(\frac{\rho^2 D}{\alpha^2(1+D^2)}\right)+\frac{\alpha^2}{\rho^2}\cos\left(\frac{\rho^2 D}{\alpha^2(1+D^2)}\right)e^{-\rho^2/\alpha^2(1+D^2)}\right]\right] \quad (1.43)$$

for large D we get $T(\rho) = T(\rho)_{ray}\ C$ where

$$C=\left[\frac{1+\frac{\alpha^2}{\rho^2}\cos\left(\frac{\rho^2}{\alpha^2 D}\right)e^{-\frac{1}{D^2}}}{2}\right] \quad (1.44)$$

Hence the transverse auto-covariance of the travel-time function decreases compared to the corresponding ray limit. Note the factor of 2 in the denominator and the condition that $\rho/\alpha > 3$ for the expansion to hold. Therefore, since the data are from finite frequency seismograms, our evaluation of $T(\rho)$ will be less than if we actually had infinite frequency seismograms. Hence, we will underestimate the structure, which is what we have already covered in our discussion of ray-width and wavefront healing. This illustrates that for a Gaussian structure we can be missing no more than 1/2 the signal as

measured by T(ρ). *Frankel and Clayton* [1986] suggest that the crust is better represented by a "self-similar" correlation function. This has much more power at shorter scales, and one would be missing more than 1/2 the infinite frequency structure on a finite frequency seismogram in this case. A close viewing of equation 1.43 shows that as α decreases, the discrepancy becomes larger, as expected.

We can combine equations 1.28 and 1.39 to give

$$\begin{aligned}\sigma^2(a) &= C\int_{ray} x_{1/2}(r)\left[R(r,0)-\int_0^{2a} w(a,\rho)B(a,\rho)R(r,\rho)d\rho\right]dS \\ &= C\int_{ray} x_{1/2}(r)\int_0^{2a} w(a,\rho)B(a,\rho)[R(r,0)-R(r,\rho)]d\rho dS\end{aligned} \tag{1.45}$$

Of interest is a simplification for the case of uniform sampling. That is, B(a,ρ)=constant. One gets

$$\sigma^2(a) = C\int_{ray} R(r,0)x_{1/2}\left(1-\left(\left(\frac{\alpha^2}{a^2\pi^{1/2}}\right)\left[1-\Phi\left\{\frac{1}{2};2;-4\left(\frac{a}{\alpha}\right)^2\right\}\right]\right)\right)dS \tag{1.46a}$$

where Φ is the degenerate or confluent hypergeometric function [*Gradshteyn and Ryzhig*, 1980] p.1057. A potentially useful practical form is the following,

$$\sigma^2(a) = C \int_{ray} R(r,0) x_{1/2} (1 - f((a/\alpha)^2)) dS$$

where

$$f(w) = \frac{1}{w}\left[1 - \frac{4}{\pi}\int_0^1 \sqrt{1-x^2}(e^{-4wx^2})dx\right]$$

Approximation for w << 1

$$f(w) = 1 - w + \frac{5}{6}w^2 - \frac{7}{12}w^3 + \frac{7}{20}w^4 - \frac{11}{60}w^5$$

and for w >> 1

$$f(w) = \frac{1}{w}\left[1 - \frac{1}{\sqrt{\pi w}}\right] \qquad (1.46b)$$

In equation 1.45 we have related the observables, the variance of the summary rays to the model parameters, R(r,ρ). Hence, we have completed the derivation for a cartesian, flat, Earth. It is suitable for localized studies, but for the present study we need to account for the finiteness and sphericity of the Earth.

Let the fluctuations in the medium be parameterized as follows

$$\delta U(r,\theta,\phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} A_{lm}(r) Y_{lm}(\theta,\phi) \qquad (1.47)$$

where $Y_{lm}(\theta,\phi)$ are fully normalized spherical harmonics, θ is the co-latitude, and ϕ is the longitude. $A_{lm}(r)$ are the coefficients of the appropriate spherical harmonics; note that they are a function of radial position in the Earth.

$$Y_{lm}(\theta,\phi)=\left(\frac{(2l+1)(l-m)!}{4\pi(l+m)!}\right)^{1/2}P_l^m(\cos\theta)e^{im\phi} \tag{1.48}$$

where P_l^m are the associated Legendre functions. We take the harmonic coefficients to be independent random functions of depth. We can then define an auto-correlation function in terms of the power spectral coefficients of the structure, Q_l. By definition

$$Q_l=\frac{1}{2l+1}\sum_{m=-l}^{l}A_{lm}A_{lm}^{*}=E\left\{A_{lm}A_{lm}^{*}\right\} \tag{1.49}$$

for the set of surface spherical harmonics of the same degree l. Here * represents the complex conjugate. If f(r) is the coherency of the harmonic pattern with depth then

$$E\left\{A_{lm}(r_1)A_{pq}^{*}(r_2)\right\}=Q_l\left(\frac{r_1+r_2}{2}\right)f(|r_1-r_2|)\delta_{lp}\delta_{mp} \tag{1.50}$$

where δ is the Kroenecker delta operator. Thus

$$\begin{aligned}E\{\delta t_1\delta t_2\}&=E\left\{\int_{ray1}\int_{ray2}\sum_{l=0}^{\infty}\sum_{m=-l}^{l}\sum_{p=0}^{\infty}\sum_{q=-p}^{p}A_{lm}(r_1)A_{pq}^{*}(r_2)\left(\frac{(2l+1)(l-m)!}{4\pi(l+m)!}\right)^{1/2}\left(\frac{(2p+1)(p-q)!}{4\pi(p+q)!}\right)^{1/2}\right.\\&\left.P_l^m(\cos\theta_1)P_p^q(\cos\theta_2)e^{im\phi_1}e^{-iq\phi_2}dS_1dS_2\right\}\\&=\int_{ray1}\int_{ray2}\sum_{l=0}^{\infty}\sum_{m=-l}^{l}Q_l\left(\frac{r_1+r_2}{2}\right)f(|r_1-r_2|)\frac{(2l+1)(l-m)!}{4\pi(l+m)!}P_l^m(\cos\theta_1)P_l^m(\cos\theta_2)e^{im(\phi_1-\phi_2)}dS_1dS_2\end{aligned} \tag{1.51}$$

The second step arises from substituting in equation 1.50. Using the addition theorem for associated Legendre functions we get

$$E\{\delta t_1 \delta t_2\} = \frac{1}{4\pi}\int_{ray1}\int_{ray2}\sum_{l=0}^{\infty}(2l+1)Q_l(r)f(\delta r)P_l(\cos\lambda)dS_1 dS_2 \tag{1.52}$$

where

$$\cos\lambda = \cos\theta_1\cos\theta_2\cos(\phi_1-\phi_2)+\sin\theta_1\sin\theta_2$$
$$r = (r_1+r_2)/2$$
$$\delta r = r_1 - r_2 \tag{1.53}$$

where λ is the angular distance between two arbitrary points on the two rays. Rather than defining the fluctuations, we could have defined the material properties immediately in terms of a correlation function, and the obvious choice would be the integrand of equation 1.52, i.e.

$$\frac{1}{4\pi}\sum_{l=0}^{\infty}(2l+1)Q_l(r)f(\delta r)P_l(\cos\lambda) \tag{1.54}$$

Hence, equation 1.52 is analogous to equation 1.32. As before, we assume that the two rays stay parallel along their path and that the statistics of the medium are isotropic, hence, f(δr) is defined by what the horizontal correlation is, since by isotropy they are equal. The propagation distance and the radius of curvature of the ray must far exceed the correlation length. Hence, following the same steps as in the cartesian case we get this result

$$T(\lambda) = \frac{C}{4\pi}\int_{ray} x_{1/2}(r)\sum_{l=0}^{\infty}(2l+1)Q_l(r)P_l(\cos\lambda)dS \tag{1.55}$$

where T is the autocovariance of the travel time residuals as mapped

onto the Earth's surface and λ is the horizontal angular distance between two rays.

To derive the areal averaging equations we proceed in a similar manner to the cartesian case. If Θ represents the angular radius of the area at hand (if the area is centered at the pole then Θ is the colatitude of its margin), and λ represents the angular distance between two arbitrary points within the area, then the area is

$$A = 2\pi(1-\cos\Theta) \tag{1.56}$$

while provided $\Theta < \pi/2$, $w(\Theta,\lambda)$ the areal weighting function is

$$w(\Theta,\lambda) = \frac{\pi - 4\cos\Theta\cos^{-1}\alpha + \cos^{-1}\beta_1 + \cos^{-1}\beta_2}{2\pi(1-\cos\Theta)^2}\sin\lambda \tag{1.57}$$

if $0<\lambda<\Theta$ and

$$w(\Theta,\lambda) = \frac{\pi - 4\cos\Theta\cos^{-1}\alpha + \sin^{-1}\beta_1 - \sin^{-1}\beta_2}{2\pi(1-\cos\Theta)^2}\sin\lambda$$

if $\Theta<\lambda<2\Theta$, where

$$\alpha = \frac{\cos\Theta(1-\cos\lambda)}{\sin\Theta\sin\lambda}$$

$$\beta_1 = \frac{(1-\cos\lambda)(1+\cos\lambda-\cos\Theta(1+\cos\Theta))}{(1-\cos\Theta)\sin\Theta\sin\lambda}$$

$$\beta_2 = \frac{(1-\cos\lambda)(1+\cos\lambda+\cos\Theta(1-\cos\Theta))}{(1+\cos\Theta)\sin\Theta\sin\lambda} \tag{1.58}$$

There is no closed form solution for $\Theta > \pi/2$, except at $\Theta = \pi$ the solution is $w(\pi,\lambda) = (\sin\lambda)/2$. In this case the areal averaging kernel must be evaluated numerically. Examples are shown in figure 1.11.

We have evaluated the theory for areal averaging assuming that the averaging area is a spherical cap. As discussed in the data analysis section the areal averaging is done over square areas on a spherical surface. Hence applying this theory to our data involves a small approximation. For the smallest scales the difference is small, while for the largest scales the boxes actually are spherical caps. Hence, the biggest difference is at intermediate scales. Compared to other approximations, this one is relatively benign.

Methods developed in the mining industry for extending limited measurements to interpolate values spatially could potentially be adapted to estimate the travel-time auto-covariance function. We could use parametric geostatics as derived from regionalized variable theory, such as kriging, to develop estimates of the whole travel-time autocorrelation surface from the individual points, taking into account their distribution. This method would also allow an estimate of the quality of the surface derived [*Clark*, 1979]. More robustly we could use non-parametric geostatics [*Henley*, 1981]. Unfortunately this is not a trivial task for our spatial field since it is a function of a six-component vector, the source, and receiver location. For our application we see no advantage, but for other applications these above methods could allow cheaper, more robust and more informative methods to be applied.

We can again express the observables in terms of the structural parameters representing the Earth's heterogeneity. We combine together equations 1.29 and 1.55 to get

$$\sigma^2(\Theta) = \frac{C}{4\pi}\sum_{l=0}^{\infty}\int_0^{2\Theta}\int_{ray} w(\Theta,\lambda)B(\Theta,\lambda)(2l+1)x_{1/2}(r)Q_l(r)(1-P_l(\cos\lambda))dSd\lambda \tag{1.59}$$

This can be cast in matrix form as

$$\mathbf{D} = \mathbf{GXF}^{\mathrm{T}}$$

where

$$D_{ij} = \sigma_i^{\,2}(\Theta_j)$$

$$G_{ik} = \frac{dS_i}{dr}\Delta r_k$$

$$F_{jl} = \frac{C}{4\pi}\int_0^{2\Theta ij} w(\Theta_j,\lambda)B(\Theta_j,\lambda)(1-P_l(\cos\lambda))d\lambda$$

$$X_{kl} = (2l+1)x_{1/2}(r_k)Q_l(r_k) \tag{1.60}$$

where we have transformed the ray integral to an integral over Earth radius and then discretized that integral. The Δ-Z bins are indexed by i. The scale sampling is indexed by j. Earth radius and harmonic degree are indexed k and l respectively. Equation 1.60 describes a linear doubly-decoupled inverse problem in depth and harmonic degree (l). The matrix $\mathbb{D}$ contains the data; each row represents the summary ray variance for a given Δ-Z bin, as a function of scale. $\mathbb{G}$ is a matrix of depth kernels and each row represents the path length of the i^{th} ray within the different depth bins. $\mathbb{F}$ is a matrix of spectral kernels, where the j^{th} row relates the variance measurement at the j^{th} scale bin to the structure at the different har

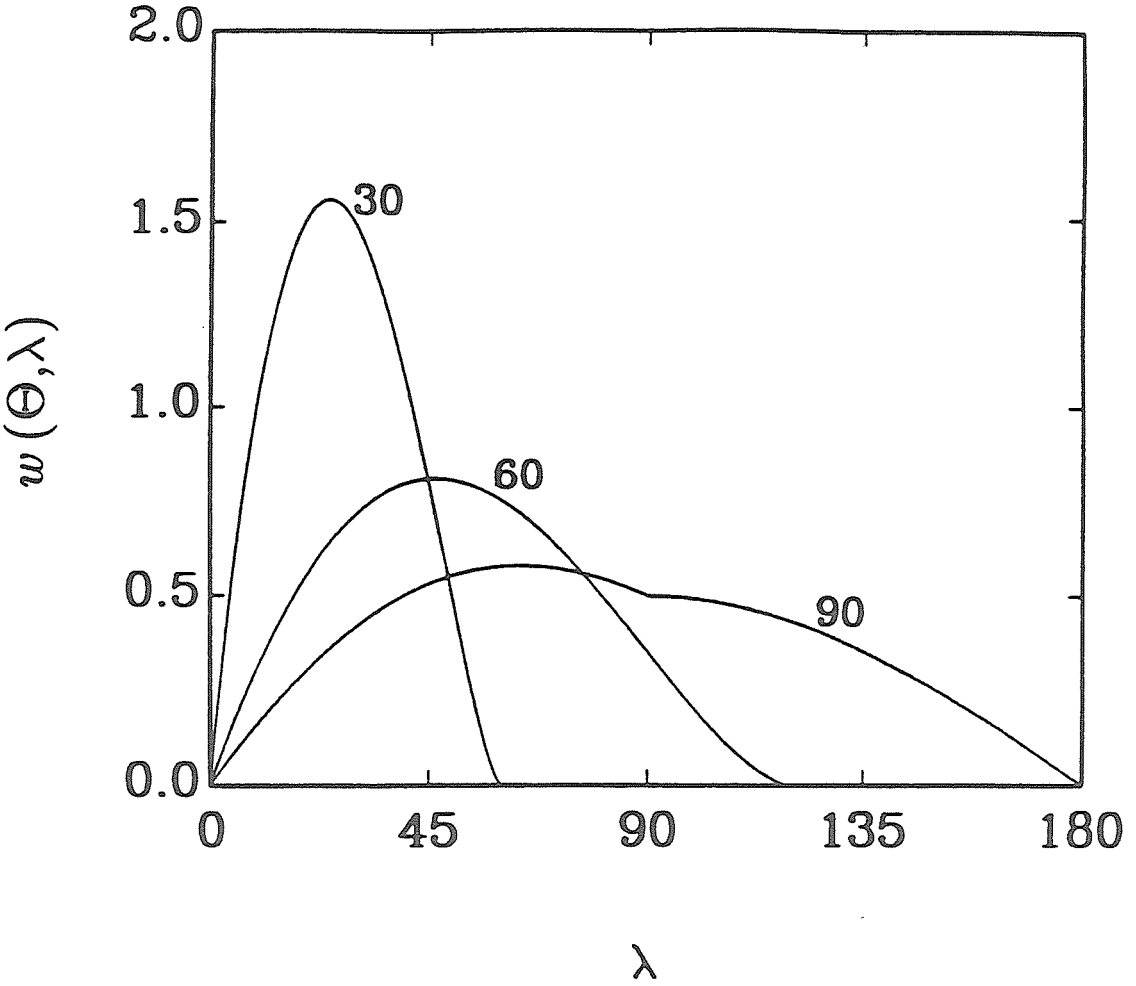


Figure 1.11. Examples of weighting functions $w(\Theta,\lambda)$. Spherical geometry. Notice that the peak of the kernel occurs for separation $\lambda \approx \Theta$. From *Gudmundsson et al.* [1989].

monic degrees. X is the unknown model matrix, where each row represents the product of the half-width and power spectrum of the Earth's slowness field at a given depth bin.

1.4 Inversion

In equation 1.60 we have a forward model describing the effect of lateral heterogeneity as described by X_{kl} (the product of the half-width of the correlation function and the power in the spherical harmonic spectrum of the heterogeneities) on the variance of summary rays as a function of depth. Since the forward model is linear we can invert it using any one of the many linear inversion techniques that have been developed, provided we can adapt it to apply for two kernel matrices.

In inverting the above doubly-decoupled linear equation, we chose to use a damped least squares analogue. In the sense that $\mathbf{x}=(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{d}$ is the least squares solution of $\mathbf{d}=\mathbf{A}\mathbf{x}$ and $\mathbf{x}=(\mathbf{A}^T\mathbf{A}+\alpha^2\mathbf{I})\mathbf{A}^T\mathbf{d}$ is the damped least squares solution, (where $\mathbf{I}$ is the identity matrix and α is the damping parameter) so

$$\mathbf{X}=\mathbf{G}^T(\mathbf{G}\mathbf{G}^T+\alpha^2\mathbf{I})^{-1}\mathbf{D}(\mathbf{F}\mathbf{F}^T+\beta^2\mathbf{I})^{-1}\mathbf{F} \tag{1.61}$$

is the damped least squares analogue for equation 1.60, a matrix equation where we have two kernel matrices and the left-hand side is a matrix rather than a vector. By setting $\alpha = \beta = 0.0$, and substituting equation 1.60 into equation 1.61, it can be demonstrated that it is a solution of equation 1.60. While by analogy with the simple exam-

ple above, it can be seen that the damping parameters are in the correct positions.

In least squares inversions the model obtained is the one that minimizes the sum of the squares of the differences between the data and the data as predicted by the model; i.e. it minimizes the prediction error. Damped least squares minimize a linear sum of the prediction error and the L2 norm of the model. The damping parameter e.g., α^2 above; determines the relative importance given to data prediction and minimizing the model. The damping parameters were selected to minimize the model error, and to maximize the depth and spectral resolution. A selection of normalizations for the three terms were investigated and it was found that the tradeoff surface changed only slightly with changes in the normalizations. An example of a tradeoff surface found in the S-wave study is shown in figure 1.12. Solutions from within its well (defined by the smallest close loop in figure 1.12) were investigated and it was found that close to the minima the solution was insensitive to the exact choice. The small cross with vertical and horizontal bars shows the damping parameters chosen for the actual solution presented, while the cross with diagonal bars shows the damping parameters used in an alternate solution that will also be presented to illustrate which features were robust to variations in damping parameters and which were not. The resulting damping parameters limited the S-wave solution effectively to 9 spectral degrees of freedom and 5-6 radial degrees of freedom.

For the inversion of the P-wave data it was unfeasible to examine the tradeoff by repeating the inversion for a range of

damping parameters and plotting the tradeoff surface. Instead we simplified the problem using Singular Value Decomposition. Both G and F were decomposed as follows

$$\mathbf{G} = \mathbf{U}\Lambda\mathbf{V}^T \text{ and } \mathbf{F} = \mathbf{R}\Omega\mathbf{S}^T \tag{1.62}$$

Hence, equation 1.61 can be reduced to

$$\mathbf{X} = \mathbf{V}\Lambda(\Lambda^2 + \alpha^2\mathbf{I})^{-1}\mathbf{U}^T\mathbf{D}\mathbf{R}(\Omega^2 + \beta^2\mathbf{I})^{-1}\Omega\mathbf{S}^T \tag{1.63}$$

Notice that the only matrices that need to be inverted are diagonal and hence can be inverted trivially. This makes this process inexpensive.

Notice that damped least-squares modeling leads to global damping. In the P-wave study we have undertaken both a fine and a coarse parameterization in depth. In the fine model the layers were 30km thick in the upper-mantle, increasing to 100km in the lower mantle, and reduced again to 30km near the CMB. In this parameterization, the damping would limit the degrees of freedom and allow a smooth solution free of discretization artifacts. For our problem we find that we have poor depth resolution in the upper mantle, (since no rays bottom there) while we have large error variance in the lower mantle (since the model is small there). The boundaries in the coarse S and P models are noted in table 1.3. A coarse parameterization was undertaken so that the global damping led to lower errors than in the fine parameterization of the lower mantle. In the S-wave study only a coarse parameterization was un-

dertaken, and since we had no data in the lowermost mantle (to avoid contaminating the shear wave data with SKS data at and beyond the crossover, as discussed above) we only had 7 bins, i.e., no bin near the core-mantle boundary.

As discussed above the Backus-Gilbert method might be a preferred method of inversion. This method allows local damping to maximize resolution and to minimize model error variance at each point investigated in the model. Unfortunately it is unfeasible to evaluate the tradeoff for this double matrix problem for each point in the model. The method has been applied to the solution of the problem where the scale component has been eliminated by taking only the asymptotic variance at large-scale [*Davies et al.*, 1988]. In this case the $\mathbf{F}$ (spectral) matrix reduces to the identity matrix and the number of data are reduced. This makes the method feasible since there is one fewer kernel matrix and the data matrix is reduced. (Remember that the matrix to be inverted in the Backus-Gilbert method is of dimension data by data, as opposed to model by model in least-squares inversions.) In the long-wavelength asymptotic limit we cannot estimate the spectrum and can only solve for the product of half-width and square of slowness amplitude without any hope of separating the two factors, since we have thrown away the sensitivity of the data to different scales.

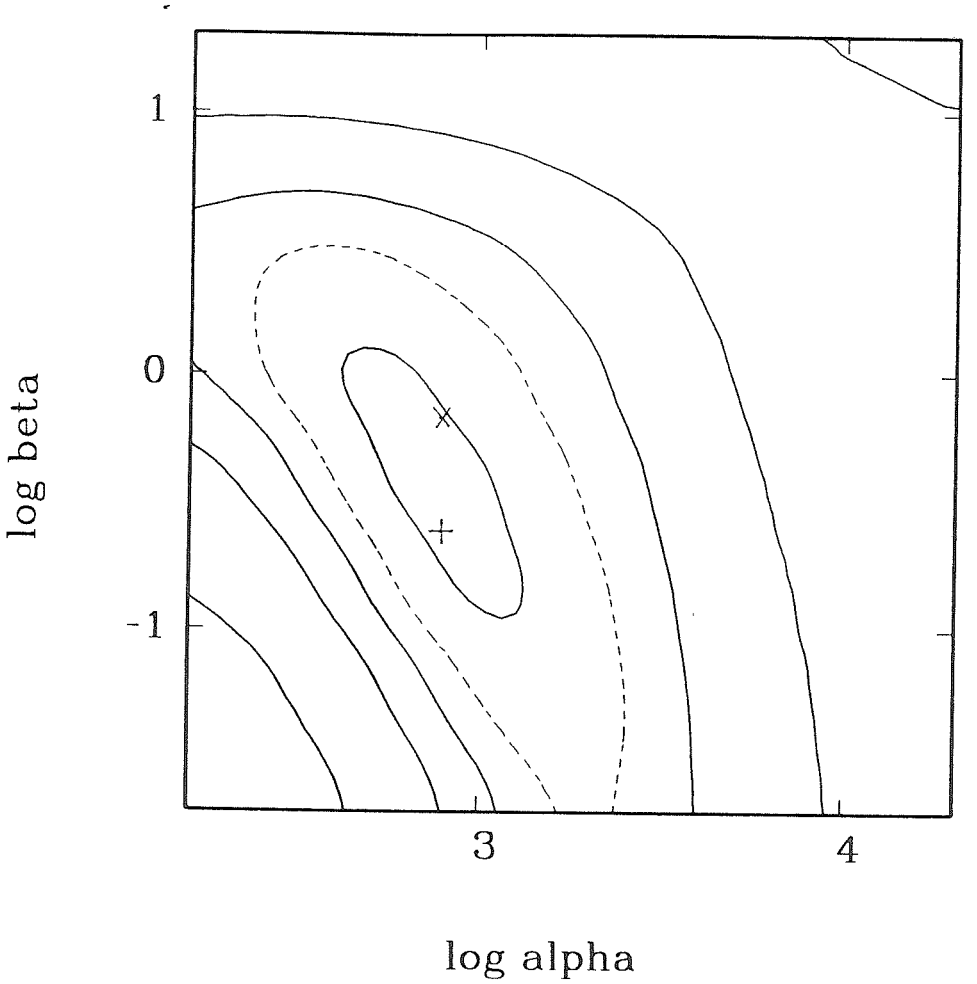


Figure 1.12 Trade-off surface for S-wave inversion. The contoured surface illustrates how a sum of the model error and the inverse of the resolution lengths of both the depth and spectral matrices varies with variation in both the depth (alpha) and spectral (beta) damping parameters. The cross with the horizontal and vertical bars shows the position of the favoured solution (figure 18) while the diagonal cross shows the damping parameters for the alternate solution (figure 26).

TABLE 1.3
Boundaries of depth bins in coarse models (km)

P	S
0	0
61	61
300	300
540	540
670	670
970	970
1470	1470
2700	2450
2900	

The singular value decomposition trade-off for the S-waves was developed considering all the singular values. Given that the selected damping parameters are in the mid-range of the singular values of the $\mathbf{F}$ and $\mathbf{G}$ matrices, the contributions corresponding to the smallest singular values are heavily damped. This was investigated by setting a finite number of the smaller singular values to zero. It was confirmed to have little effect. In the P-wave inversion the trade-off was investigated with singular values less than one thousandth of the largest singular value set to zero. The final S-wave results have the smallest singular values set to zero.

From the resolution matrices (see figures 1.13(b) and 1.14(b)) it is seen that the spectral resolution varies nearly linearly with harmonic degree. This is due to the uneven sampling of the scale axis. The spectral parameterization was chosen to be approximately parabolic; i.e. we combined together many harmonic degrees to give one model parameter, with the number of harmonic degrees

combined increasing parabolically with the harmonic degree of the mid-point of the model parameter.

The depth resolution (see figure 1.13(a) and figure 1.14(b)) of both P- and S-wave studies are nearly identical to the grid-size in the lower mantle, i.e., very good. In the upper mantle, the poorest resolution in both studies is just above the 670km discontinuity (more noticeable in the P-wave study) and the shallowest layer (more striking in the S-wave study). The resolution relative to the bin size is poor in the upper mantle; but due to the coarse parameterization of the lower mantle, the absolute measure of resolution (see the figures on the right-hand sides of figure 1.13(a) and figure 1.14(a)) can be smaller. Due to the poor resolution of the layers in the upper mantle, they are preferentially damped, especially the two regions discussed above.

In evaluating the **F** matrix, the clustering of stations and events was taken into account, i.e. $B(\theta,\lambda)$ was evaluated. This was done by evaluating the frequency of event pairs and station pairs of a given separation. This is proportional to $B(\theta,\lambda)w(\theta,\lambda)$; hence, we can derive B by dividing the histograms by the known functions w. This was repeated over different scales. It was found that the logarithm of the frequency varies nearly linearly with the logarithm of the separation, as seen in figures1.15 (a) and (b). i.e., the sampling distribution is a power law function of separation. The exponent of this power law was found to be more negative at smaller scales; i.e., as viewed at smaller scales, events and stations seem more clustered. The variation in the exponent with scale was found to be reasonably well fit by a linear relationship. For both data-sets the distribution

was evaluated for the data of a single reference ray (P 50°-52° and 1-32km; S 51°-55° and 1-32km), and was assumed to hold for all the other reference rays. The distribution for the shear waves was found to be a much weaker function of scale. This is not unexpected since the P-wave data-set is approximately 5 to 6 times larger, while the pattern of seismicity and station distribution is little changed between the two data-sets. Both do show clustering though.

In figure 1.16 the kernels for a uniform distribution (dashed line) are compared to the P-wave kernels as derived following our estimation of the magnitude of the clustering effect (solid line). As can be seen from figure 1.16, using the kernels based upon a uniform distribution would lead to one overestimating the correlation length of the medium. The shear waves show a marked deviation from a power law relationship at the largest scales, see the curve for scale 90° in figure 1.15(a). It is seen that the curve has a large deficiency of pairs at large separations, even allowing for the expected trend. Therefore, we have very little data at large-scale that are related to large separations. Hence, in the S-wave data we have little sensitivity to the largest wavelength heterogeneities. This follows from the fact that no major zones of seismicity are antipodal, equally there are no antipodal continents (on which nearly all seismic stations are located).

We checked our kernels derived with the P-wave data distribution. We did it by evaluating the variance arising from the following synthetic structure

$$(2l+1)x_{1/2}(r)Q_l(r) = \delta_{lk}\ \delta(r-r_o) \qquad (1.64)$$

The resulting variance is a direct measure of the kernel since

$$\sigma^2(\theta_j) = \frac{C}{4\pi}\int_0^{2\pi} w(\theta_j,\lambda)B(\theta_j,\lambda)(1-P_k(\cos\lambda))d\lambda = F_{j,k} \tag{1.65}$$

From figure 1.17 it can be seen that our evaluated kernels (solid lines) compare well with the kernels predicted by the synthetic structure test (boxes and circles). Notice though that the misfit is largest at the large-scales and low degrees. Hence, as with the shear wave data, but less so, we shall have difficulty in imaging the longest wavelength structures, i.e. the low harmonic degrees.

1.5 Results

The depth and spectral resolution matrices are presented in figures 1.13(a) and (b) and 1.14(a) and (b). We notice that the model has depth resolution of the order of the bin size in the lower mantle, while in the upper mantle the relative resolution is poorer with it being worst near the surface and also bad just above the upper-lower mantle boundary. The spectral resolution is found to deteriorate with harmonic degree, until order l =100-200, beyond which it remains approximately constant (P-wave study) or decreases slightly (S-wave study). Poor resolution means that the estimated value at a point is an average over the "true" values around that point. Hence, this tends to make models smoother than the Earth. Poor resolution for a parameter not only means that its value

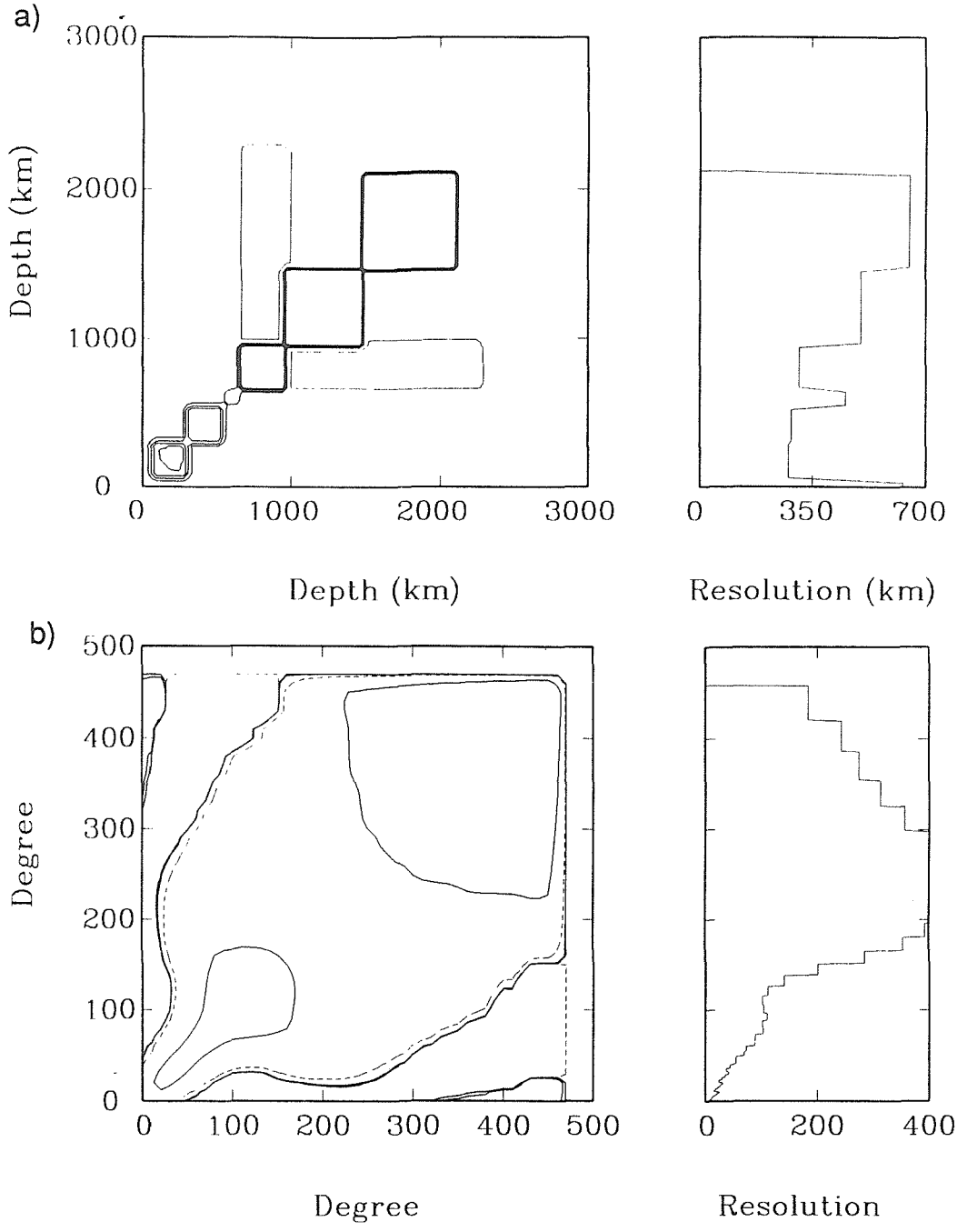


Figure 1.13 Depth (a) and spectral (b) resolution matrices of the S-wave study.

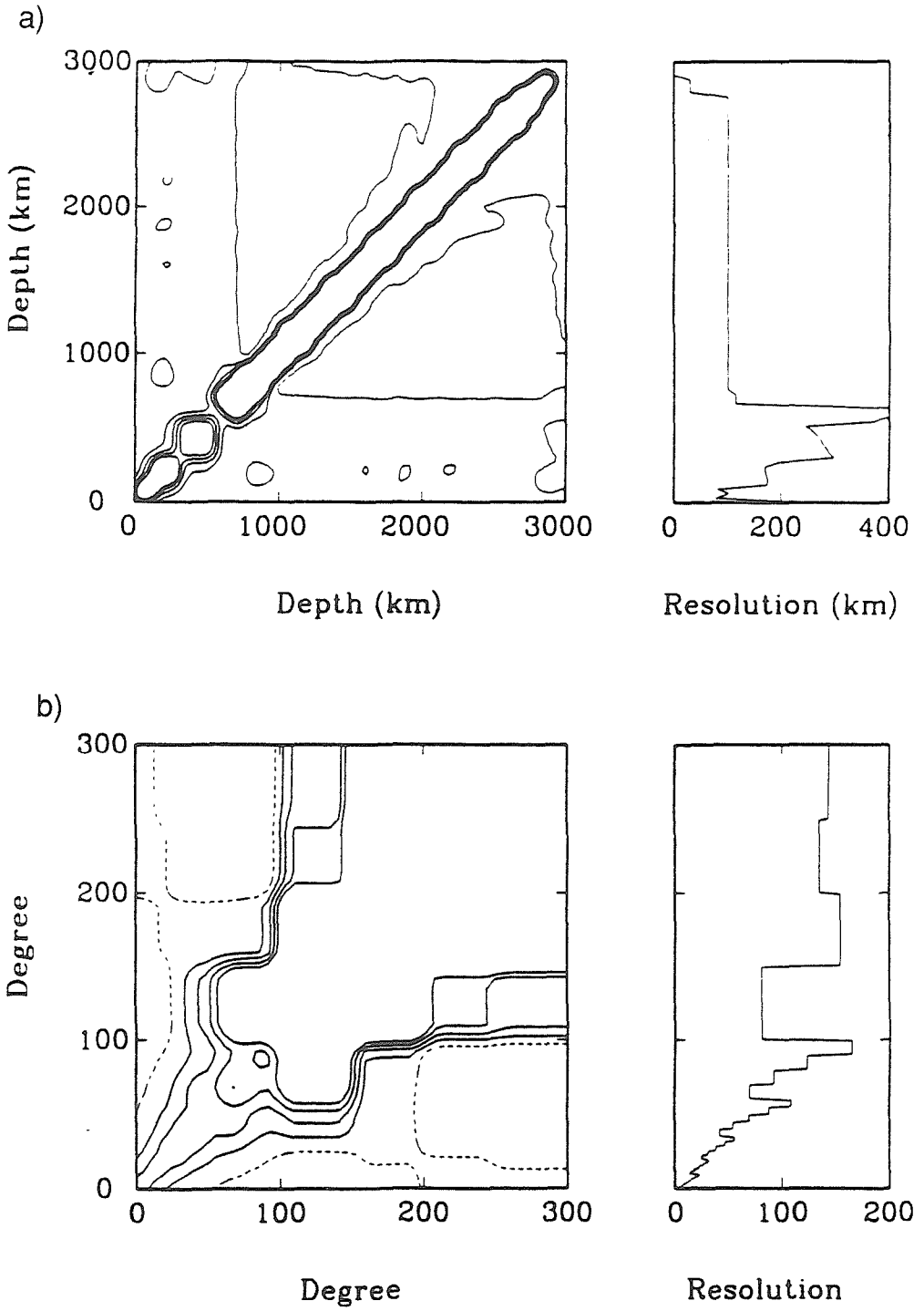


Figure 1.14 Depth (a) and spectral (b) resolution matrices of the P-wave study. From *Gudmundsson et al.* [1989].

is an average of "true" values in its neighborhood but also that its "true" power is distributed among all those parameters in its neighborhood. This is a problem in our case since our poor resolution at shallow depths and high harmonic degrees means that we will suffer some leakage into the deeper depths and into the low harmonic degrees. If the poorly resolved parameter's "true" value is large compared to the rest of the model, it will contaminate the model by increasing the power in other regions of the model, while if the "true" power is low it will contaminate the model by decreasing power in other regions of the model. Since this and previous work suggest that the shallow layers of the Earth are more heterogeneous, the poor resolution in the upper mantle will potentially lead to more power deeper in the Earth. On the other hand, it is possible that there is more large-scale power than small-scale power, and the lower harmonic degrees might be decreased since their power is leaking out into the poorly resolved higher harmonics just as the higher harmonics are leaking into the lower harmonics. Note that the resolution matrices are symmetric by definition.

The results for the S-wave study are presented in figures 1.18(a), (b), (c), (d), (e), and (f). In figure 1.18(a) we present the total power (the product of half-width and the square of the mean slowness). This is seen to be highest in the upper mantle and decreases as we go deeper. There are two exceptions, first, the surface value is lower while second, there is an increase as we cross the 670km discontinuity. We have poor resolution of the shallowest layer as shown in figure 1.13(a). Since there are errors in the data, we must damp the solution to prevent poorly constrained parameters from destroy

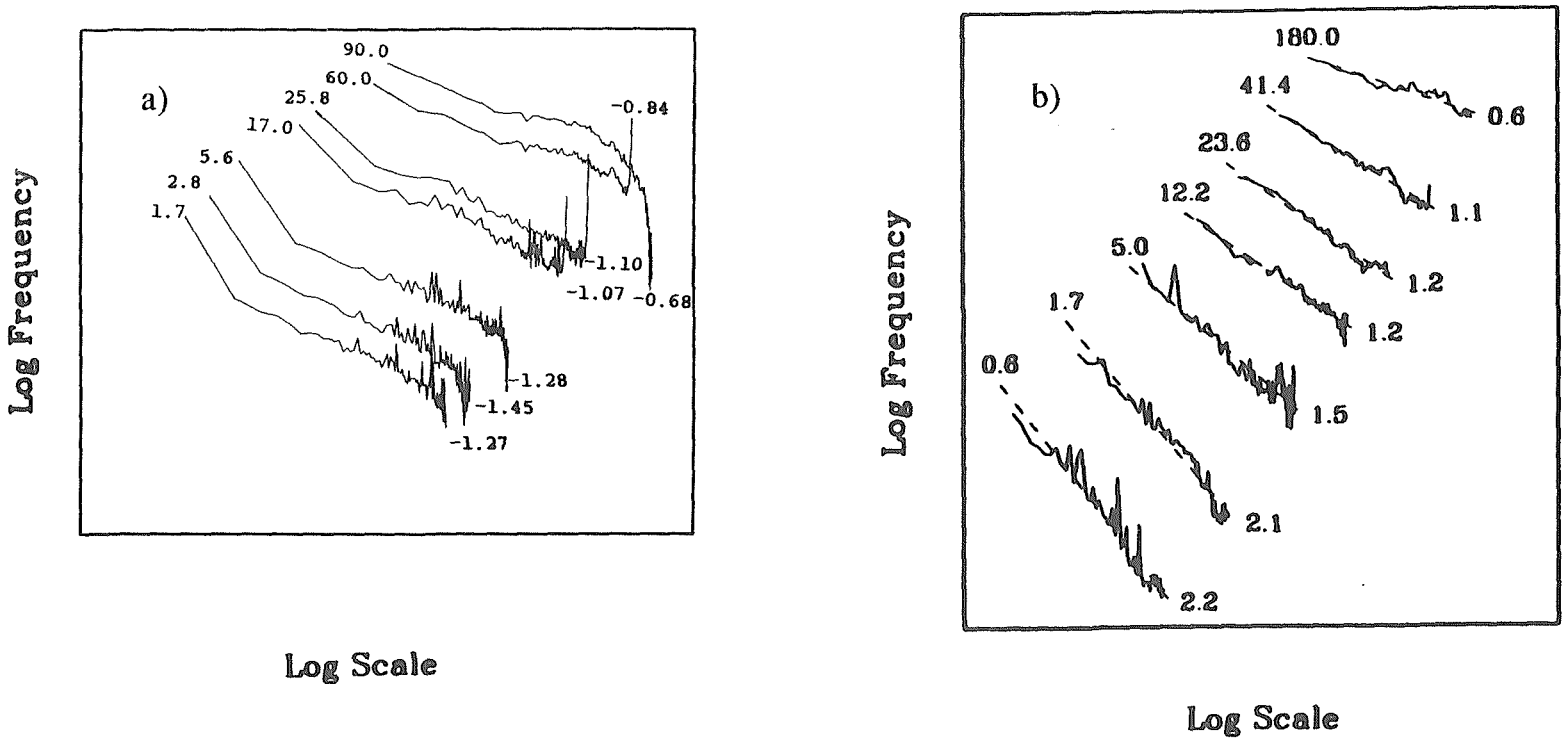


Figure 1.15 Plot of the normalized log pair frequency as a function of separation for different scales. (a) S-wave study. (b) P-wave study, from *Gudmundsson et al.* [1989].

ing the solution. The shallowest layer is heavily damped since it is poorly resolved, therefore its actual value is only weakly constrained by the data and in reality its value could be larger.

In figure 1.18(b) we present the spectra for the first 25 harmonic degrees for all seven depth bins. The spectra are normalized to unit height. The dashed line signifies negative power, an unphysical result. The negative values are very small, hence they are not very different from zero. Their magnitude suggests the maximum accuracy of our solution, i.e., we cannot claim our result is any more accurate than these unphysical terms. Their small size gives us hope that these results are not hopelessly contaminated by systematic errors.

Since the spectral resolution deteriorates with increasing harmonic degree, it was decided to split the spectrum into two halves; a long-wavelength low-harmonic degree half and a short-wavelength high-harmonic degree half. The low harmonic degree has better spectral resolution and it was felt that a reasonable estimate might be made of its corresponding correlation function. Hence, its half-width might be evaluated with some hope of being meaningful. The cut-off between high and low degree was arbitrarily taken to be l=50.

In figure 1.18(c) we present the short-scale power, the part of the total power due to structure with harmonic degree $l > 50$. Nearly all this power is concentrated in the second bin (from 60-300km depth). There is very little small-scale power in the lower mantle; while some of the values (e.g., the third depth bin) go negative. These negative values again suggest systematic errors and

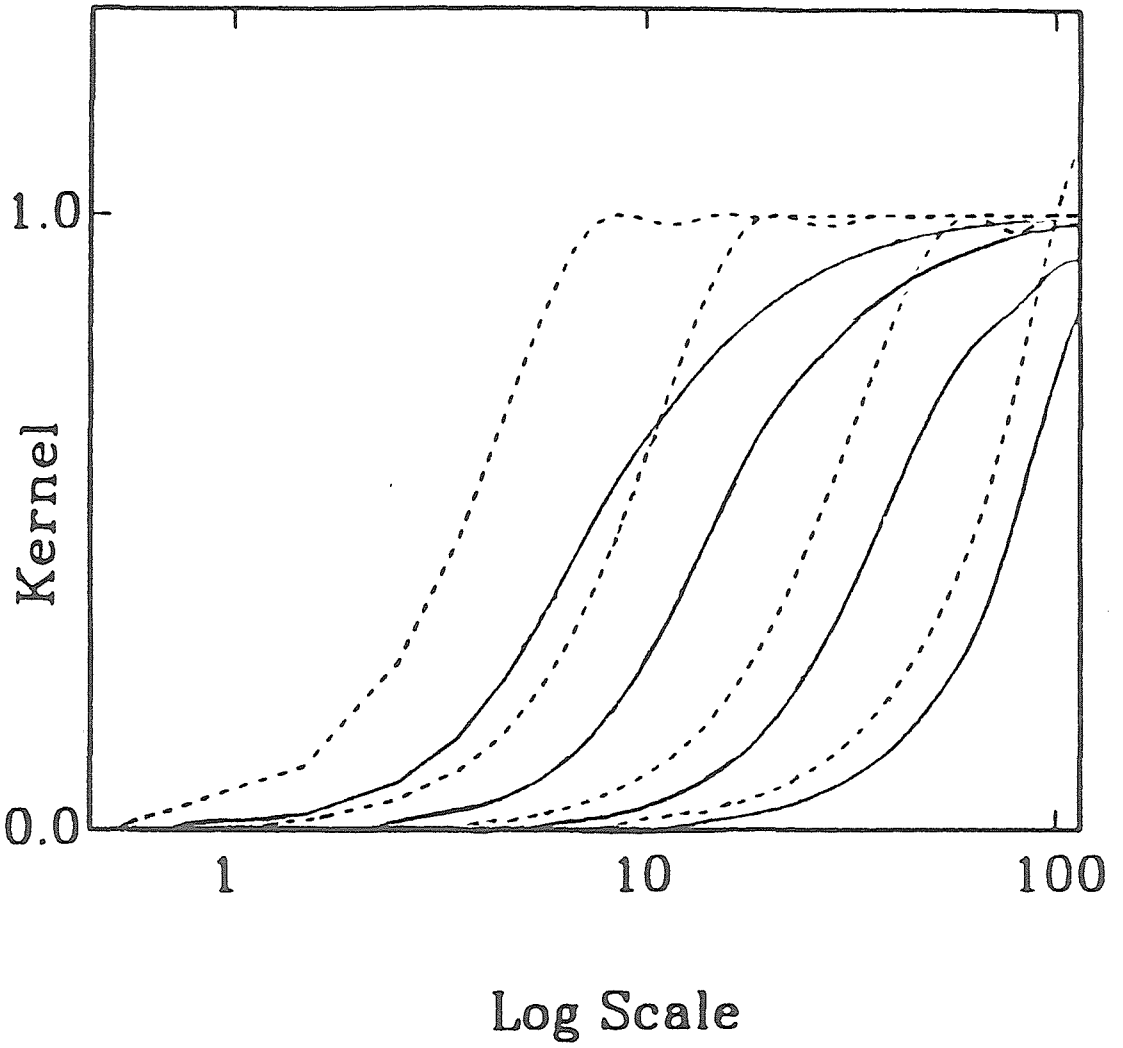


Figure 1.16 Comparison of kernels for a uniform distribution (dashed line) and the actual P-wave distribution evaluted in figure 1.15 (solid line). Harmonic degrees l =1,3,10 and 15 are shown. Adapted from *Gudmundsson et al.* [1989].

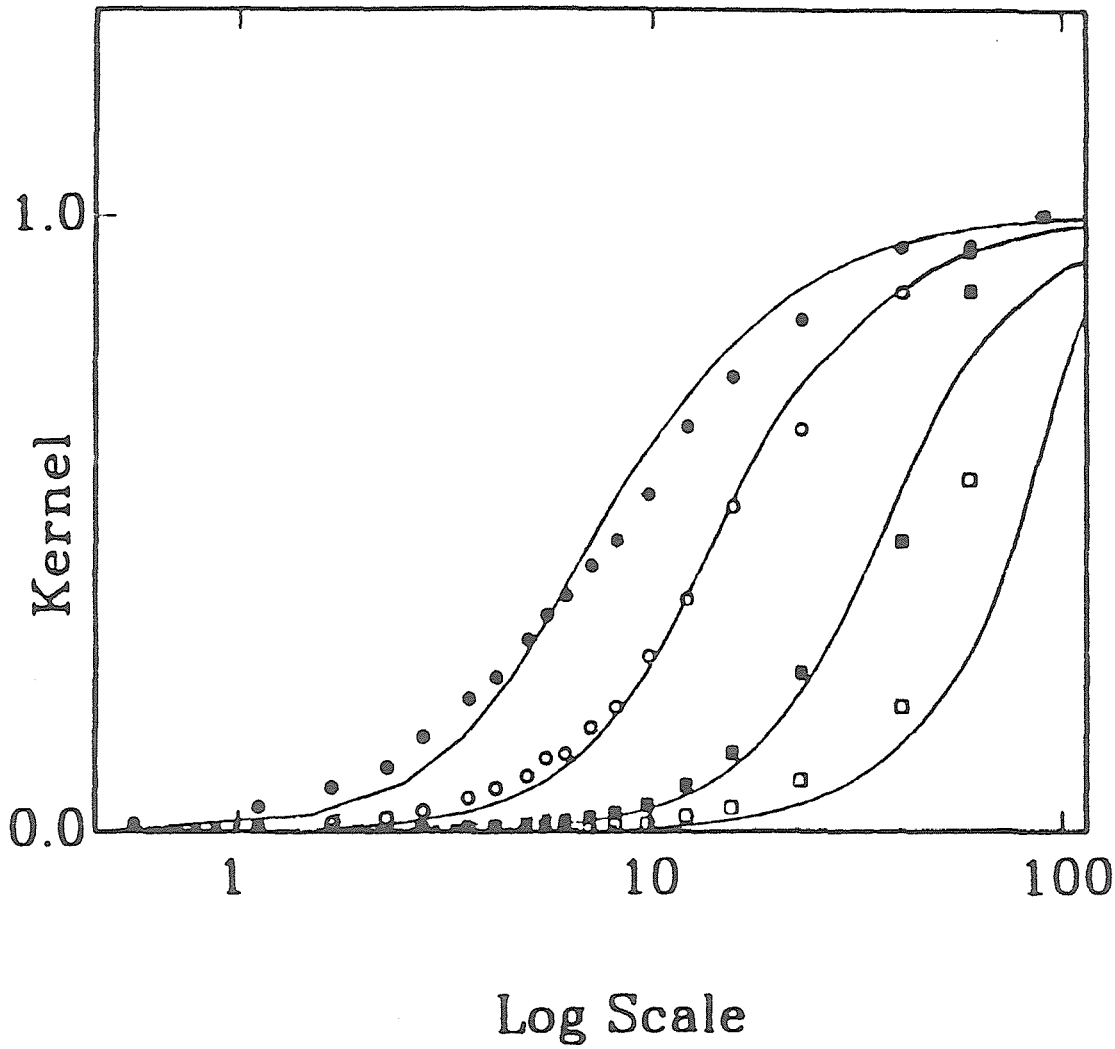


Figure 1.17.Comparison of kernels evaluated assuming P-wave distribution of figure 1.17 (solid line) and empirical kernels (boxes and circles) evaluated as a resposnse to a synthetic anomaly eqn. 1.65; for harmonic degrees 1,3,10,15. Adapted from *Gudmundsson et al.* [1989].

problems with the inversion. In figure 1.18(d) we present the large-scale power, due to structure with $l < 50$. This power is largest in the third layer (i.e., from a depth of 300-540km), and decays through the lower mantle. The layer just below the 670km discontinuity is larger than the value above the 670km discontinuity. The negative small-scale power and the huge large-scale power for the third layer suggests that the spectrum for this layer is extremely focussed in the low harmonic degrees. In comparison with the other layers of both P- and S-wave studies, we could say that this is aberrant and it suggests a problem with the spectral resolution of the power in the S-wave study for this depth.

In figure 1.18(e) we present the slowness estimate as derived from the large-scale power. In figure 1.18(f) we present the half-width that is estimated by evaluating the correlation function using solely the large scale-length power. By dividing this estimate of the half width into the large-scale power (figure 1.18(d)), we obtain our estimates of the slowness in figure 1.18(e). Estimating a half-width is pushing the data given the poor spectral resolution; but by restricting ourselves to the best resolved low harmonic degrees, we hopefully obtain estimates with some relation to reality! The same consideration applies to the slowness estimates, which also depend upon the half-width estimation. Having said that, we do believe that the results for both measures are actually reasonable. We suggest this for two reasons. First, both P- and S-wave studies show the same pattern, and second, the values agree with values previously evaluated by other techniques; e.g., tomography etc.

In figure 1.19 (a) we present the half-width for the fine depth parameterization in the P-wave inversion. We find that the error estimates are very large. The slowness estimates for the same model parameterization shown in figure 1.19(b) also have large error estimates. It was to obtain model estimates with lower errors that a coarser parameterization was undertaken in the P-wave study, and why only a coarse parameterization was undertaken in the S-wave study.

In figure 1.20 we present the results for the coarsely parameterized P-wave study. In figure 1.20 (a) we show the short scale-length power; while in figure 1.20 (b) we show the long wavelength power. We again observe the dominance of power in the upper mantle with the only exception again occurring at the 670 km discontinuity. This study also shows an increase in power in the bin nearest the CMB. In figure 1.20 (c) we present the slowness and in figure 1.20 (d) the half-width estimate. The half-width estimate is uniform through the upper mantle ($\approx$ 500 km), increasing through the lower mantle to around 1200km, before decreasing again near the CMB to a value near 400km. The slowness similarly decreases through the depth of the Earth, with three exceptions. At the surface where the value is low, at the 670 km discontinuity again, and at the CMB where it increases.

In figure 1.21 we present total power in the fine P-wave model and the coarse S-wave model. The power is concentrated in the upper 300km, and decreases as we go deeper into the lower mantle. Notice that both studies show the same increase in crossing

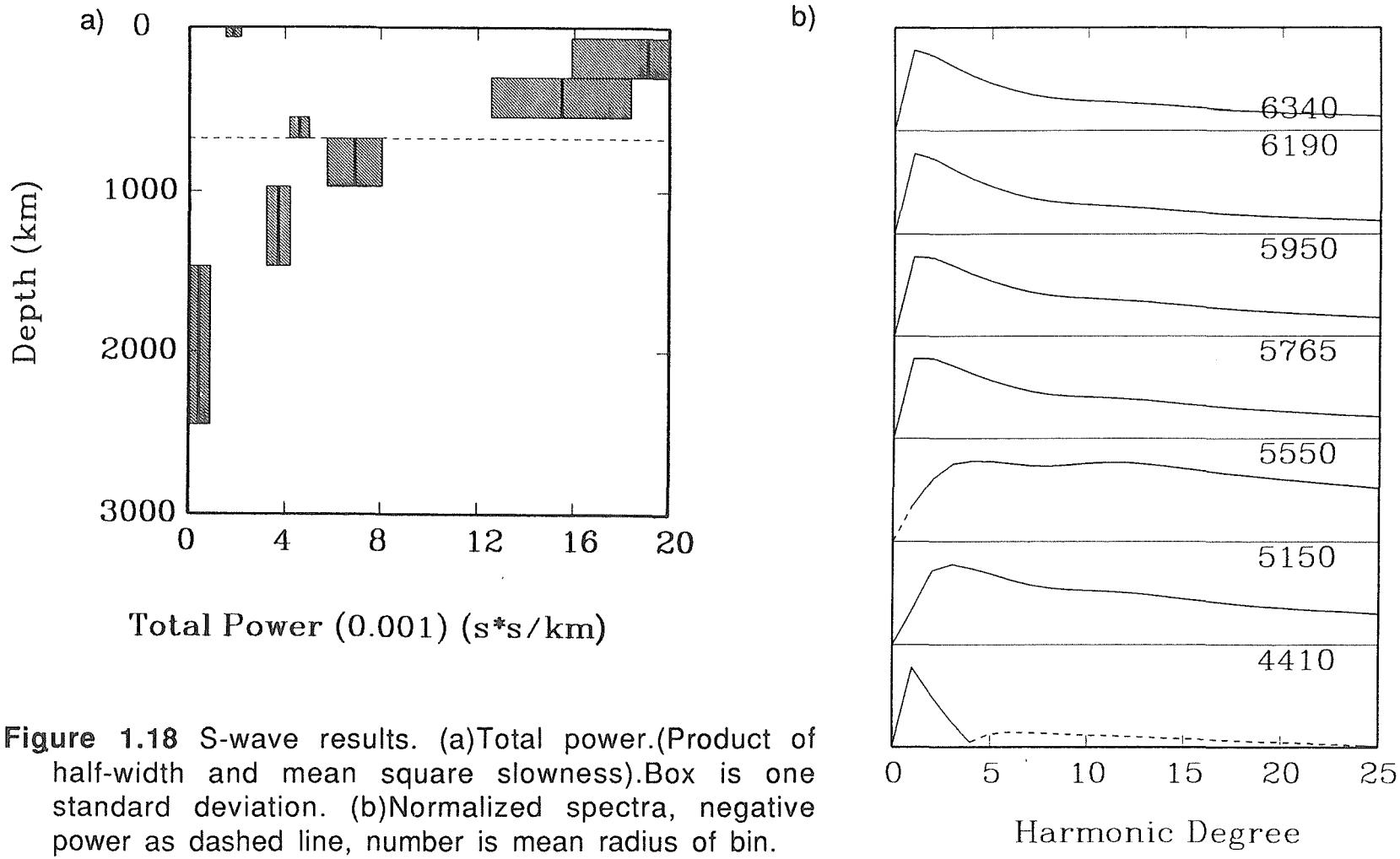


Figure 1.18 S-wave results. (a)Total power.(Product of half-width and mean square slowness).Box is one standard deviation. (b)Normalized spectra, negative power as dashed line, number is mean radius of bin.

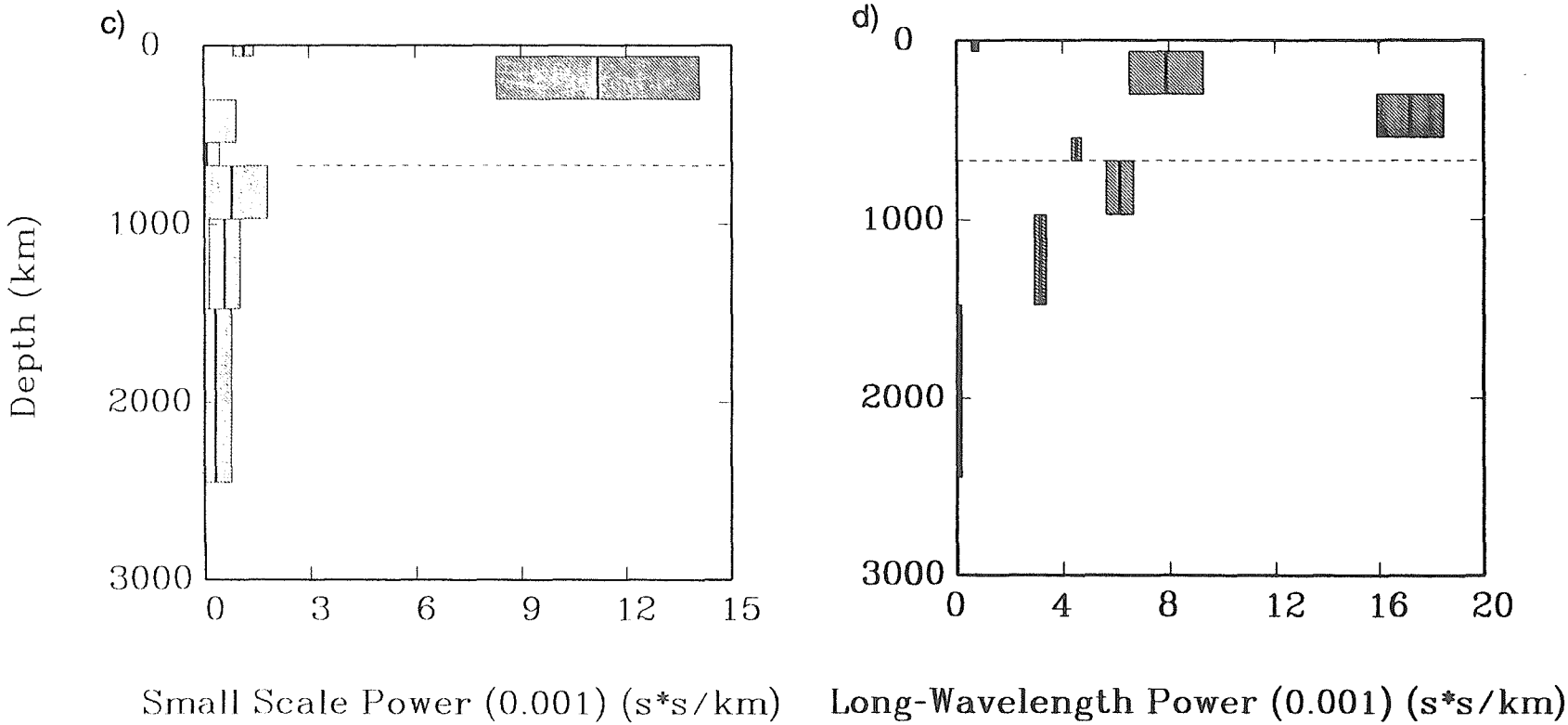


Figure 1.18 S-wave results. (c) Short scale power (l >50), notice power concentrated in uppermost mantle. (d) Large Scale Power, (l <50). Power is concentrated in the upper mantle and decays with depth in the lower mantle. Notice more power below than above 670km discontinuity.

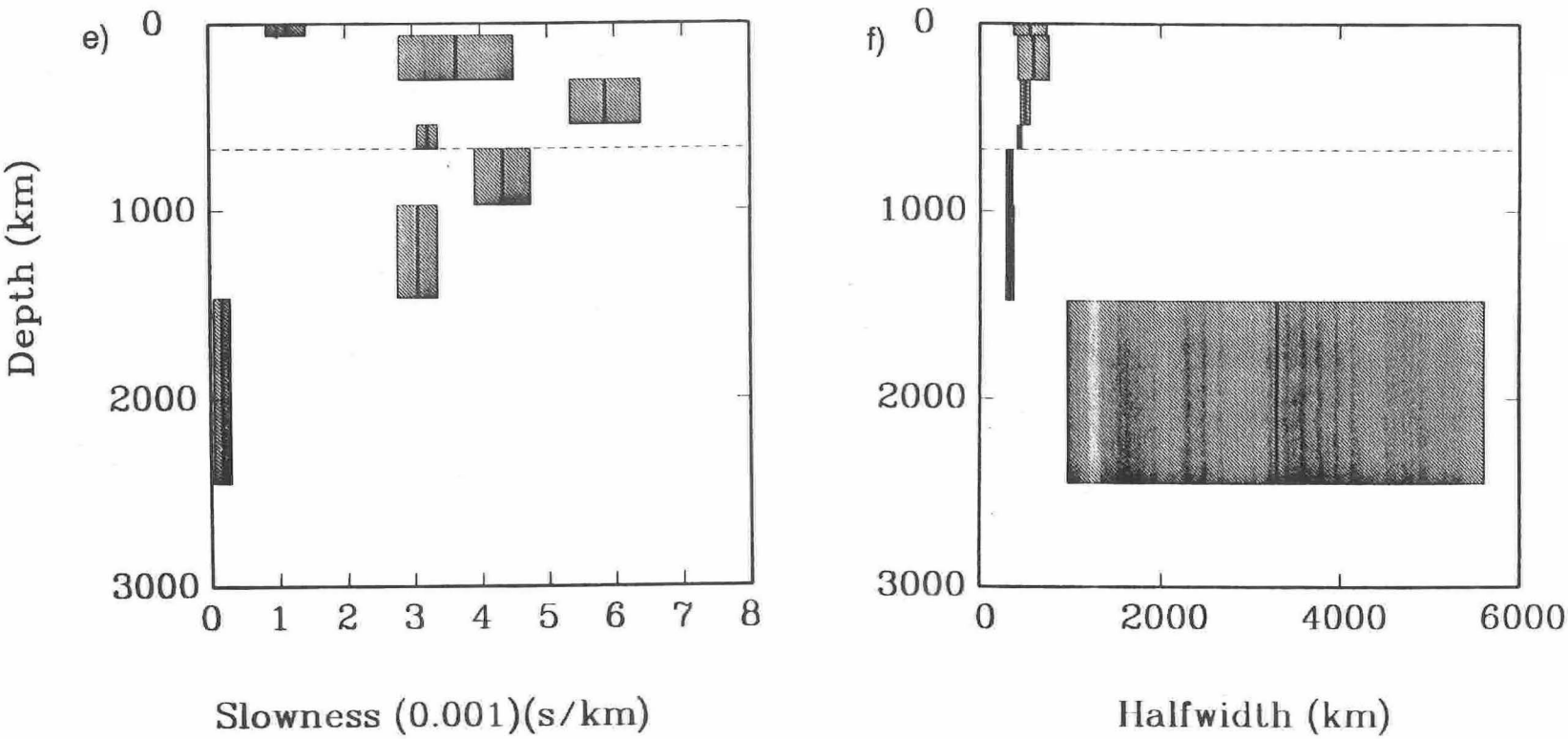


Figure 1.18 S-wave results. (e) Slowness perturbations away from reference model. Notice pattern similar to large scale power. (f) Half-width estimates of large scale correlation function. Around 300-600km for uppermost mantle. Poorly constrained in lower mantle.

the 670km discontinuity. Notice that the scales are very different; the S-wave power is approximately 60 times larger.

In figure 1.22 we present the short-wavelength power for the fine P-wave model and the S-wave model. In the P-wave study there is appreciable power in the uppermost mantle decreasing through to the 670km boundary below which there is a slight increase. This rapidly decays and for the most part the lowermost mantle is largely devoid of small-scale structure. The S-wave results are similar with a major difference being the very low level in the upper mantle below the depth of 300km. As mentioned above, we believe the difference could be the result of poor spectral resolution in the depth bin 300 - 540 km, making the short-scale power nearly zero while exaggerating the large-scale power.

The long-wavelength P and S results are also similar, with the power concentrated in the upper mantle. The S-and P-waves differ in the depth range at which the power peaks. The P-wave results show a maximum between 100 and 300km while the S-wave results show a peak between 300 and 600km. As discussed previously, due to the poor spectral resolution this difference might not be significant. Again we get the small increase just below the 670km discontinuity. The layer below the 670 km discontinuity is well resolved but the layer above isn't, hence this deviation from a smooth trend of decreasing power through the upper mantle is uncertain, and could be a result of the damping in depth. Its appearance in both data-sets though is very interesting.

The estimates of half-widths for the S-wave study are similar to the P-wave study within the error bounds at around 350-

550km. In the deepest mantle, the S-wave half-width increases to around 3000km but with very large error variance. This last estimate of half-width isn't very robust to the choice of damping parameters. The results of a slightly different damping parameter are presented in figure 1.26. This is also within a short distance of the minima in the trade-off surface (see figure 1.12) but notice here the half-width in the deepest mantle is only 500-600km. Given this variation in half-width, the slowness estimates also vary by a factor of $\sqrt{5}$ or $\sqrt{6}$ for this depth bin. Hence, the shear wave slowness estimates for the deepest mantle are also uncertain. It might not be unreasonable to assume that the half-width of long-wavelength heterogeneity in P and S velocities are equal and we would assume that the half-width is 1200km, this would lead to an estimate of 0.0005s/km for the mean amplitude in slowness estimates for the S-waves. Notice that the peak in slowness is deeper for the S-waves as opposed to the P-waves. This is as expected, given the constant half-width and the presence of the peak at the same deeper level in the S-wave long-wavelength power as discussed above.

In figure 1.23 we present the spectra for the P and S coarse models. Given the very poor spectral resolution, these should be viewed with at least a little reservation. They are suggestive though that the spectra are relatively broad in the uppermost mantle, but become sharper at depth. Interestingly, the spectra for the P-waves seem sharp at 600-670km, but the half-width estimate is still short (500km). This illustrates that the tails of the spectra at high l are significant in deciding the correlation function and its half-width. The shallow spectra of the P-wave study are negative at low har

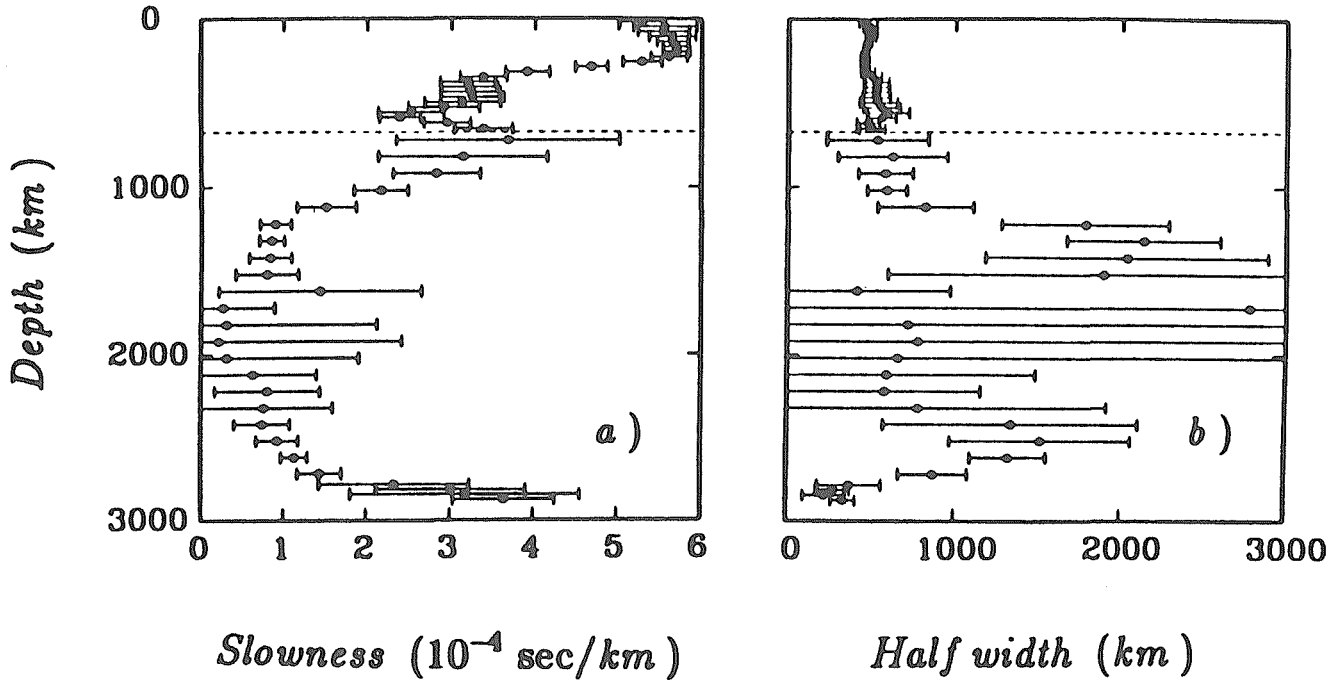


Figure 1.19 P-wave results. Fine parameterization. (a) Slowness (b) Halfwidth.From *Gudmundsson et al.* [1989].

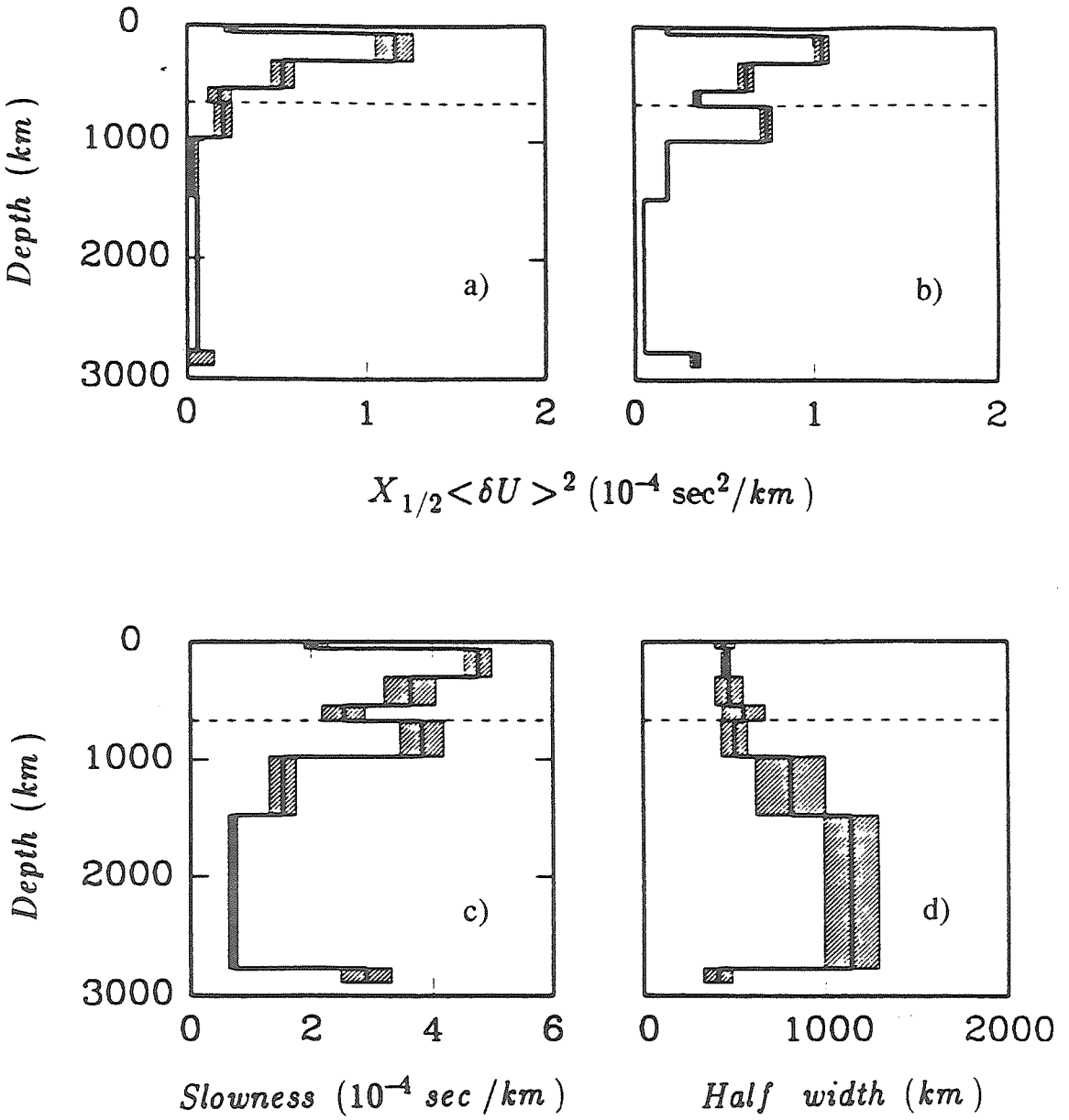


Figure 1.20 P-wave results. Coarse parameterization. (a) Short scale ($l > 50$) power, (b) Long scale ($l < 50$) power, (c) Slowness, and (d) Half-width. From *Gudmundsson et al.* [1989].

monic degree, this is due to the fact that the variance drops at large-scales for shallow source depths. Equally for the S-wave study, the power in the shallowest layer is not robust, and with slightly less damping in depth becomes negative (as seen in figure 26(b)). This arises from the fact that after removing the intercept estimates the residual data for the shallowest source depth bins consistently have smaller variance than for data from the next deeper source depth bin (see figure 1.6c).

In figures 1.24 and 1.25 we present the data predictions of the model and the original data (minus their intercept estimates) together with their error bars. It is seen that the model predicted S-wave data lie within the error bars of the actual data, but frequently the model predicted curves and data are offset a constant amount; i.e. a different intercept estimate could lead to a much better data-fit. For the P-wave data, the error estimates did not include the error arising from subtracting the intercept estimates from the data. This error is as great or greater than the maximum error in the original data. After considering the total error in the data used in the inversion, it is seen that the data predicted by the P-wave model will also be within the error bars of the original data. The variance reduction of the fine P-wave study was 98% and was 97% for the coarse study, while for the coarse S-wave study it was 94%. The chi-squared value of the S-wave study was $\approx N$, where N was the number of data (66x15=990). The number of model parameters is effectively 14 (5 degrees of freedom in depth and 9 spectral degrees of freedom), hence the number of degrees of freedom for the chi-squared distribution is 976, and hence this model gives a moderately good fit since

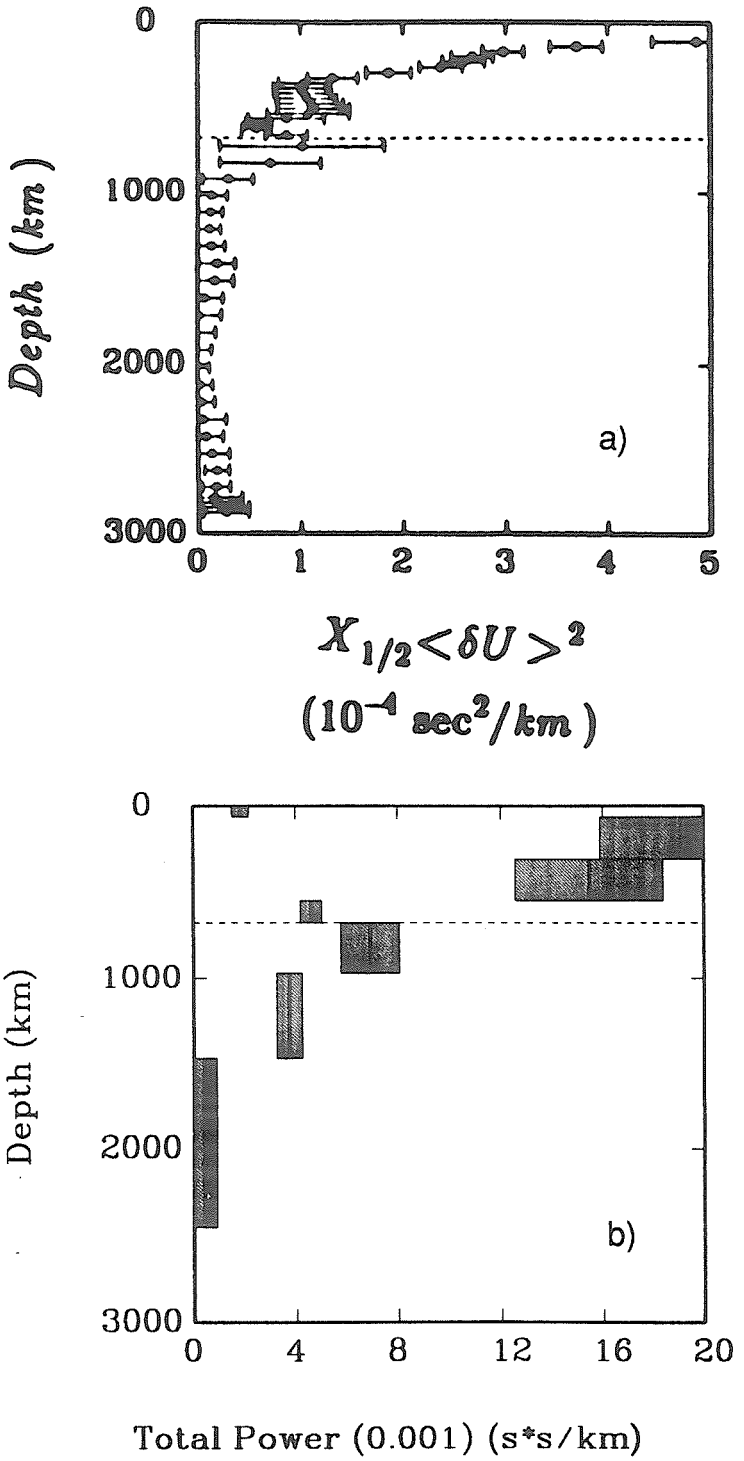


Figure 1.21 Comparison of Total Power (a) fine P-wave model (b) S-wave model.

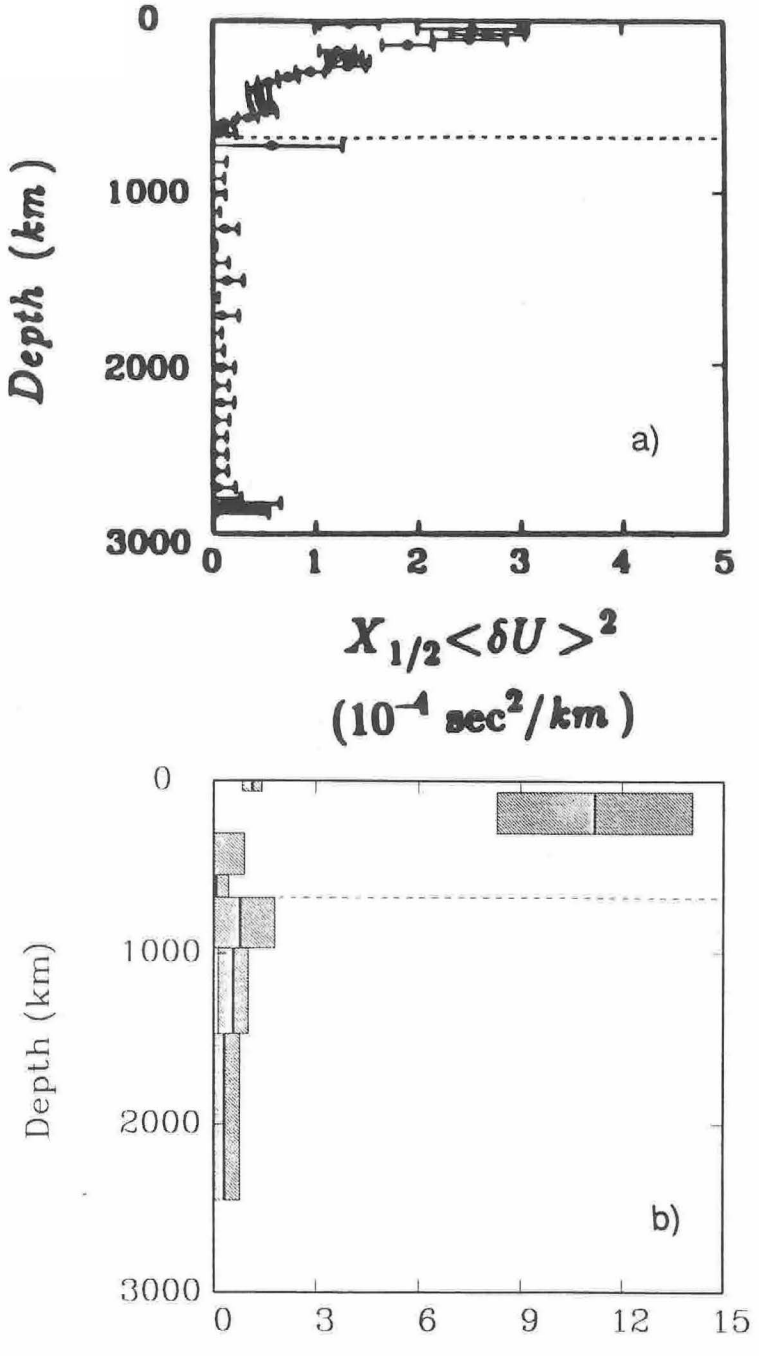


Figure 1.22 Comparison of Short Wavelength Power (a) Fine P-wave model (b) S-wave model.

$X^2 \approx$ number of degrees of freedom. Given the number of degrees of freedom, the model fit is significant and is highly unlikely to be the result of a random distribution of data.

Quite a significant part of the original signal is taken up in the intercept estimate and does not enter into the structural model. As mentioned previously, the intercept and model are not evaluated simultaneously but sequentially. This leads to a much simpler and more stable inversion scheme. Different schemes of estimating the intercept can lead to large differences in intercepts estimated. To see how these different changes might affect the final structural model, it was decided to invert a data-set where the intercept had been evaluated using logarithmic rather than linear extrapolation to the variance axis. The results are shown in figure 1.27. These intercept estimates are lower and the variance left to be inverted is much higher, $42s^2$ compared to $19.3s^2$ for the data with the linear intercept estimate removed. The variance reduction and chi-squared value are very similar to the model derived from the linear extrapolation for intercept estimation; 95% and 1.06N. The primary difference in the model results as shown in figure 1.27(c) is that there is a large increase in short-scale power at the depth range of 50-300km; and a slight increase in short-scale power throughout most of the deeper mantle from levels not significantly different from zero to levels that are slightly but significantly larger than zero. The large-scale power (d) is virtually unchanged while the half-width(f) estimated is changed for the poorly resolved shallowest mantle but otherwise the half-width and slowness estimates are unchanged. Even though these intercept estimates are substantially

different, it is shown that the long-wavelength features are unaltered. The short-scale power is increased. This is expected since we are increasing the variance change that occurs between the smallest scale and the variance axis, while the variance difference observed between two scales is unchanged. Solving for the model and intercept, simultaneously, would give us some indication of how poorly resolved the intercept and smallest scale power are. We expect there to be a nearly unlimited trade-off between the two. If we don't want to differentiate between the intercept and the smallest scale structure (S-wave < 100km; P wave < 50km) then this ambiguity is unimportant.

The residual data unexplained by the model for a single reference ray are frequently not distributed evenly about zero but are usually of the same sign for all scales, hence the variance reduction could be improved if we solved for the intercepts and model simultaneously. In the P-wave study we re-evaluated the intercept estimates to give a better fit to the data after evaluating the model; i.e. we evaluated the constant shifts of the data curves required to minimize the difference between the data and the data predicted by the model. These intercept estimates are compared with the original P-wave intercept estimates in figure 1.7. The original intercepts are in figure 1.7(a) while the intercepts arising from the adjustment are in figure 1.7(c). In this case the variance reduction was increased from 98% to 99.5% for the fine model, and from 97% to 99.3% for the coarse model. This also increased the number of model degrees of freedom substantially. It was not checked whether the increased variance reduction was significant. This suggests that the simultane

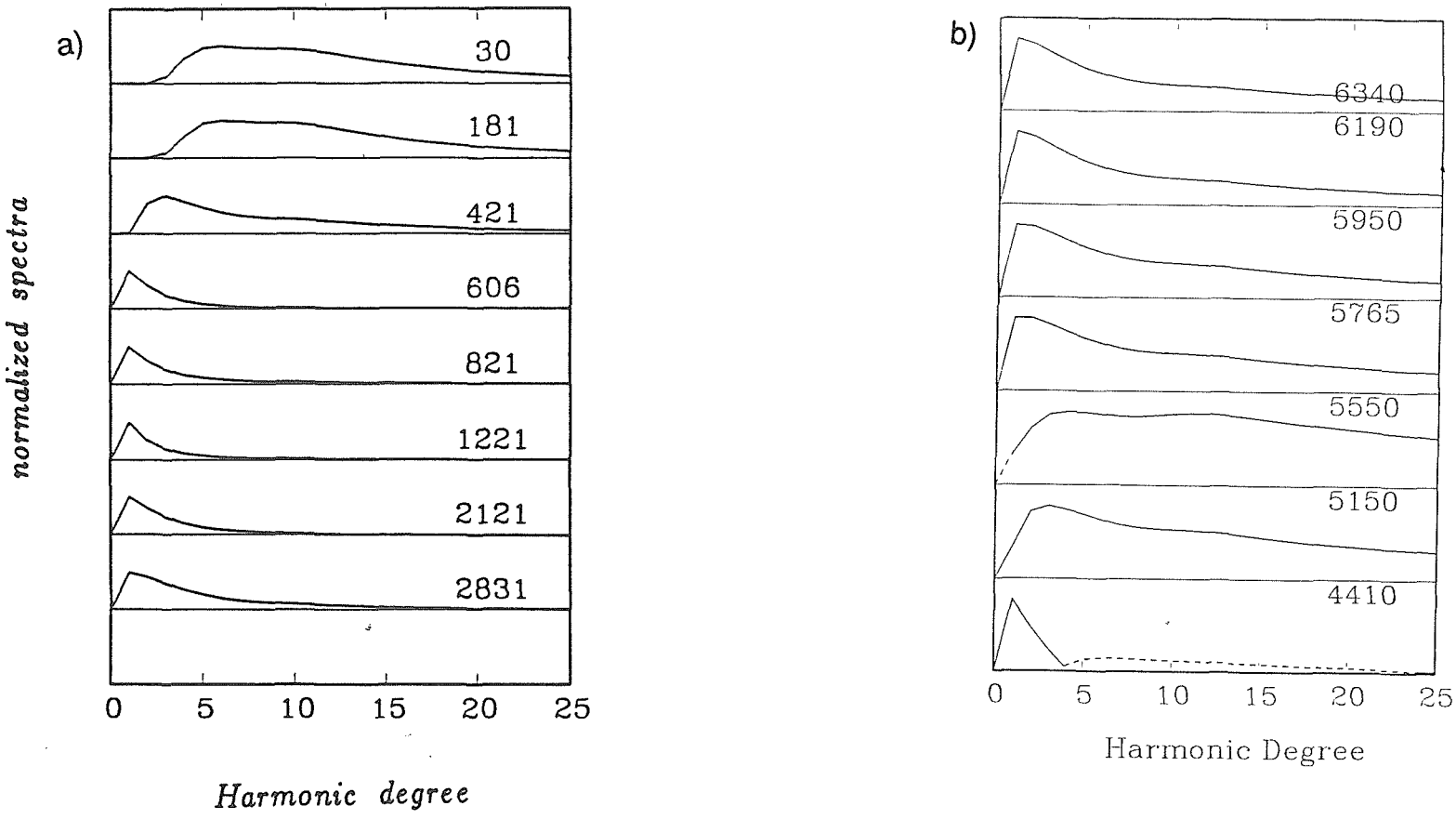


Figure 1.23 Comparison of spectra of (a) coarse P and (b) S-wave models. Numbers in P-wave study represent the depth of bin; while they represent radius of bin in S-wave study.

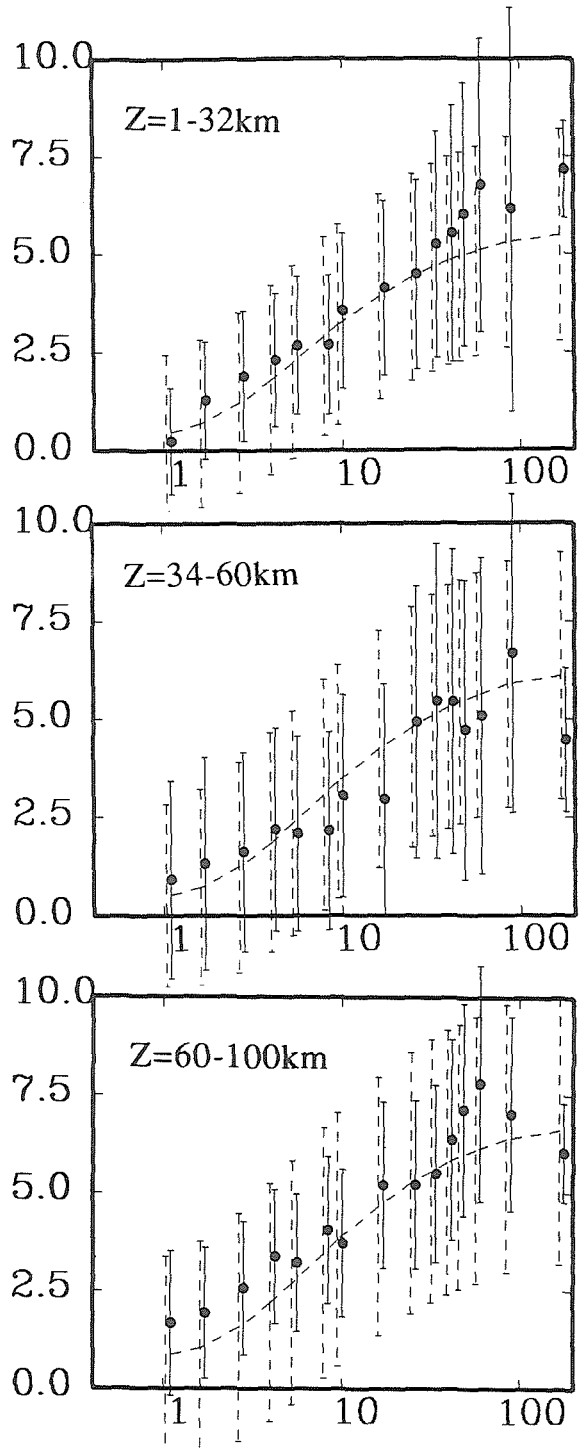


Figure 1.24 (a) Comparison of S-wave data predicted by model (dashed lines) to actual S-wave data (solid error bars and filled in circles). The vertical axis is variance in s^2. The horizontal axis is the scale axis in radians. Epicentral distance is 51^o to 55^o. (All with intercepts removed.) Depth bins are stated on figures. Shallowest bins.

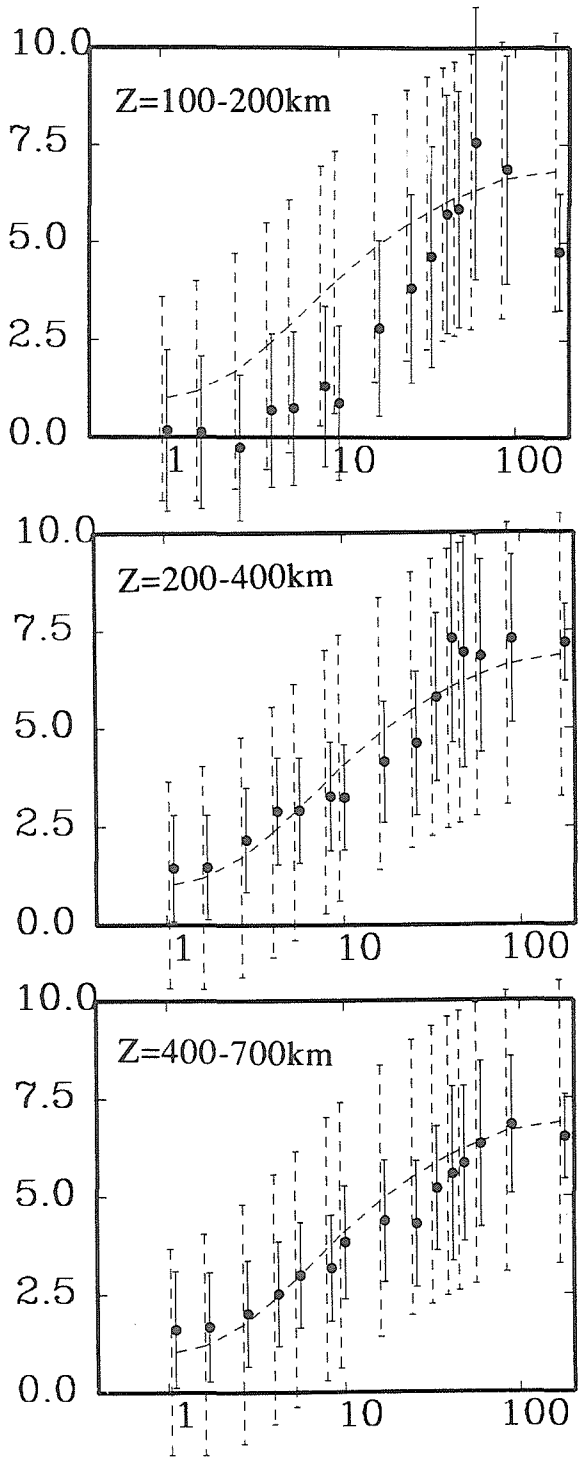


Figure 1.24 (b) Comparison of S-wave data predicted by model (dashed lines) to actual S-wave data (solid error bars and filled in circles). The vertical axis is variance in s^2. The horizontal axis is the scale axis in radians. Epicentral distance is 51° to 55°. (All with intercepts removed.) Depth bins stated on figures. Deepest bins.

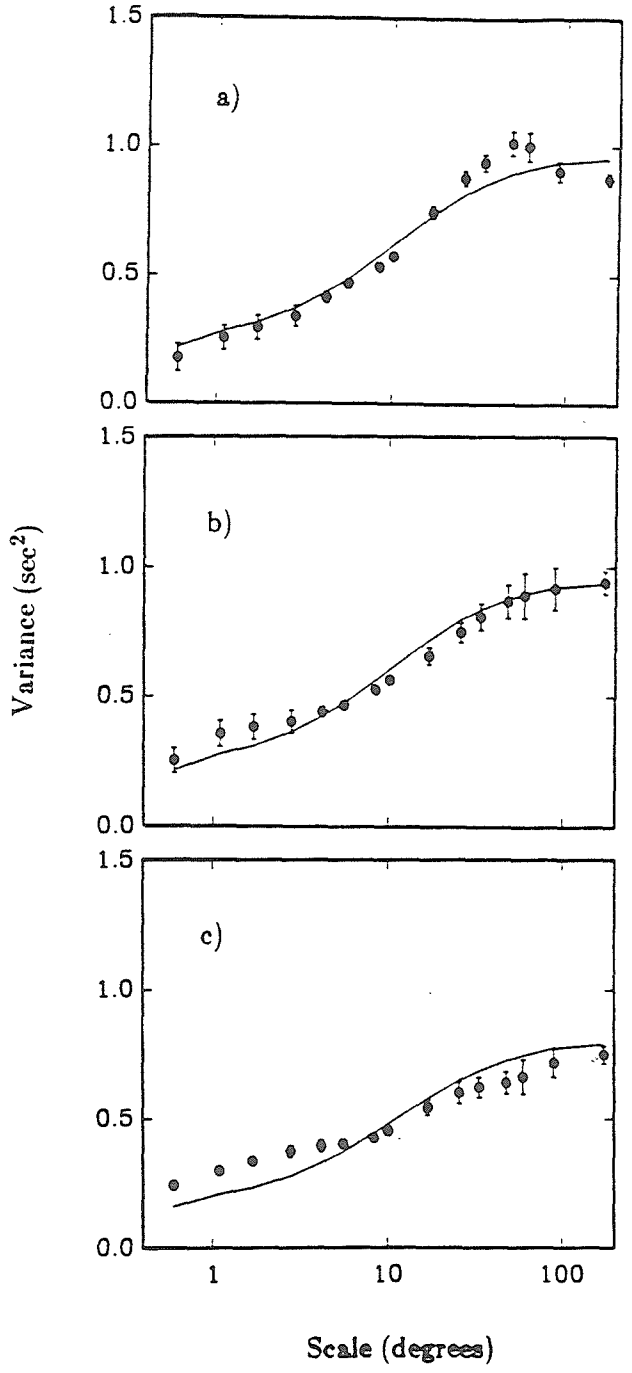


Figure 1.25 Comparison of P-wave data predicted by coarse model to actual P-wave data. (All with intercepts removed) (a) $\Delta=51^{o}$, first depth bin; (b) $\Delta=51^{0}$, third depth bin and (c) $\Delta=51^{o}$, fifth depth bin. Error bars are two standard deviations but do not include the uncertainty in estimating the intercepts. From *Gudmundsson et al.* [1989].

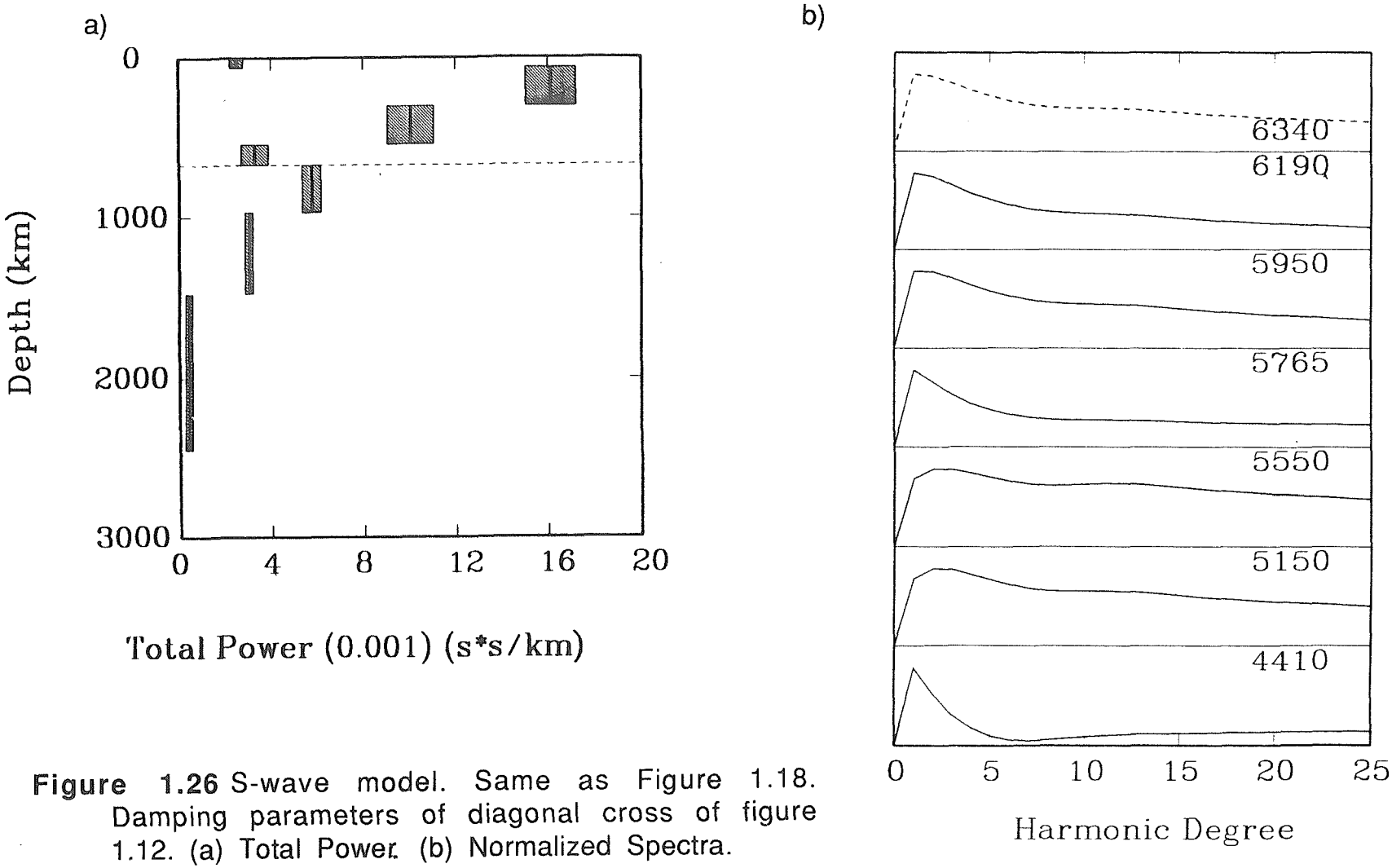


Figure 1.26 S-wave model. Same as Figure 1.18. Damping parameters of diagonal cross of figure 1.12. (a) Total Power. (b) Normalized Spectra.

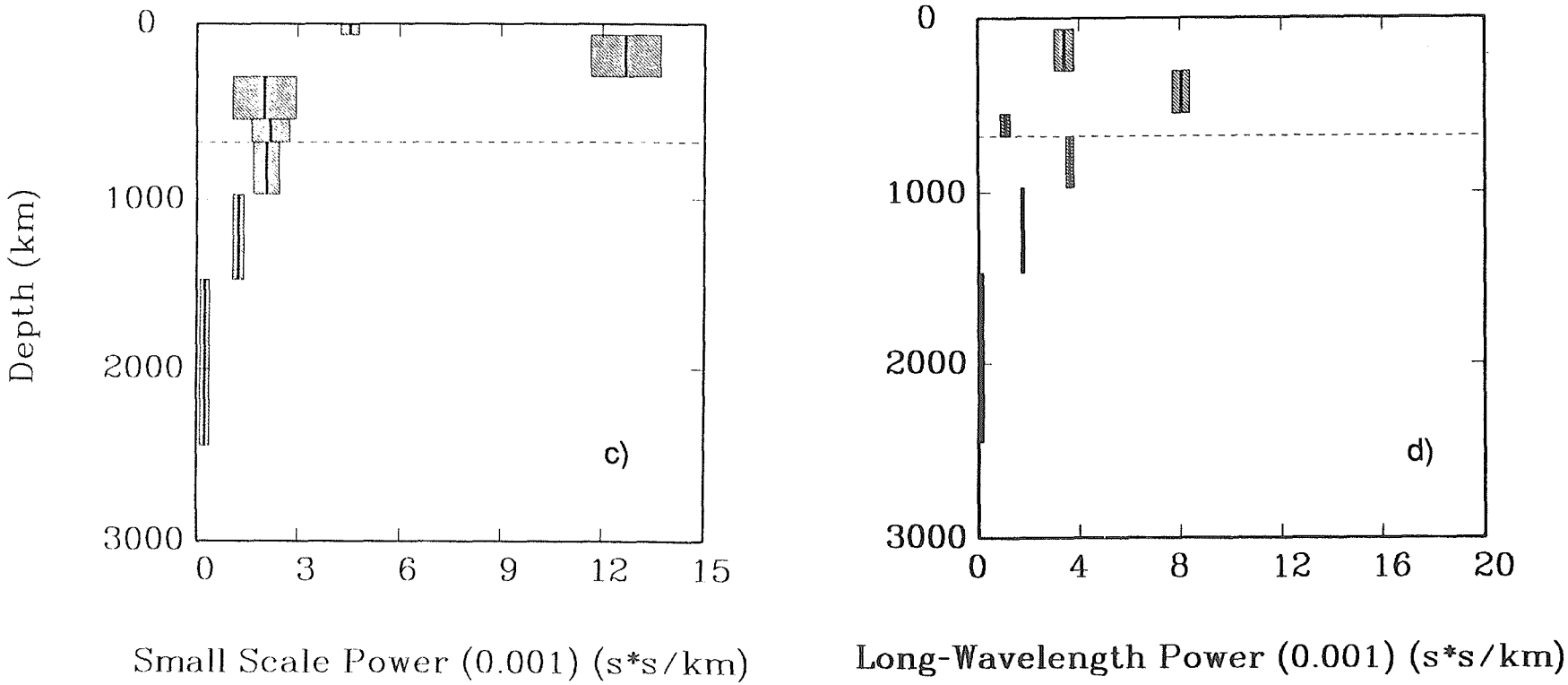


Figure 1.26 S-wave model with different choice of damping parameters (cont.). (c) Short wavelength power. (d) Long wavelength power.

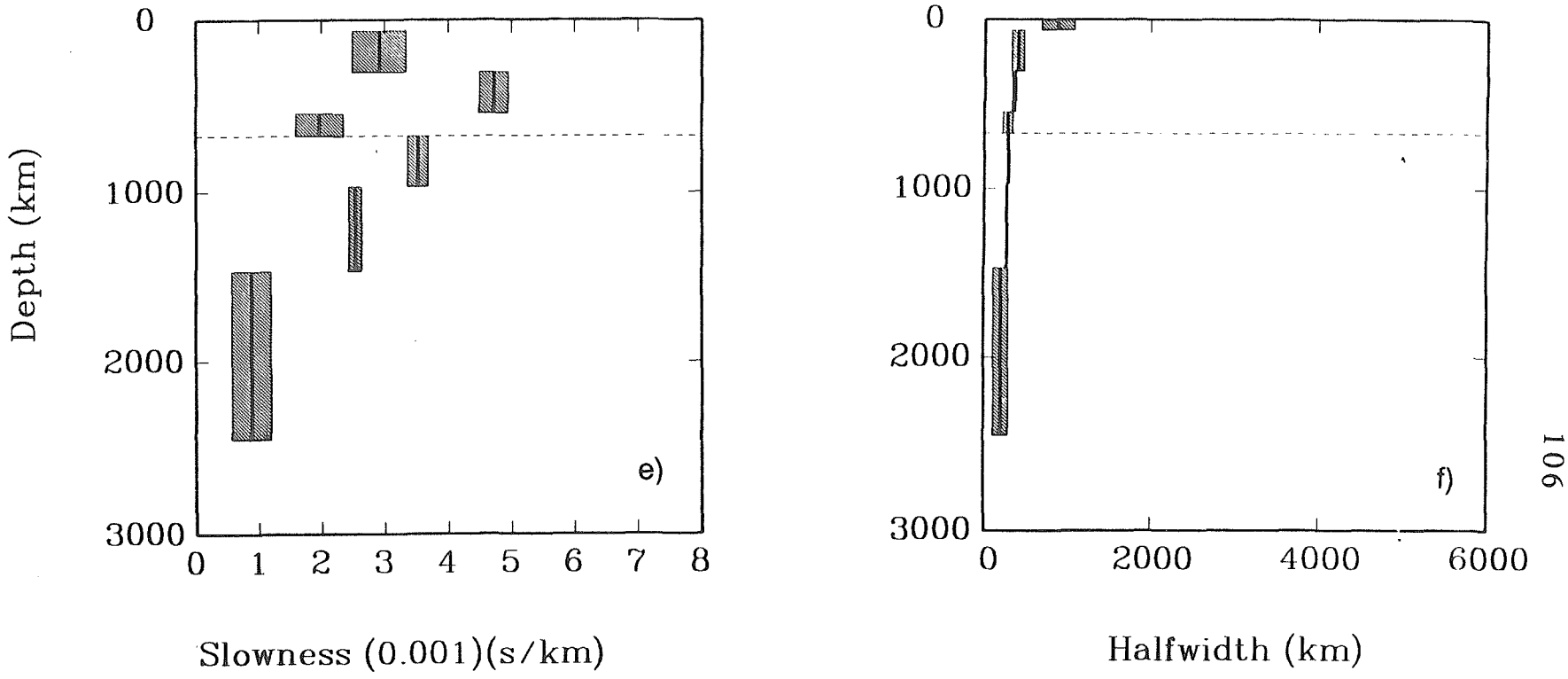


Figure 1.26 S-wave model from different choice of damping parameters (cont.). (e) Slowness. (f) Half-width.

ous solution of model and intercept might give not only a better indication of the trade-off but also a better (potentially more significant) fit to the data.

The model error variance is quite low in the whole model, i.e. the model has been sufficiently damped radially and spectrally so that the model parameters sensitive to noise in the data (low singular values) make little or no contribution to the model. The intercept estimates are a measure of the incoherent signal. This is our measure of noise in the data. It is noted that it is also possible to have scale-independent coherent noise, but this will not be estimated and will introduce systematic errors in the data, and will lead to systematic errors in the model, which will be indistinguishable from signal. The only way of confirming the absence of appreciable coherent noise (systematic error) is to obtain an equivalent model using an independent data-set. These results will be shown to be consistent with previous work suggesting that systematic errors are not a big worry.

Before discussing implications arising from these results we would like to remind the reader of limitations regarding these results. First, the model is in the high frequency limit, hence we have an intrinsic averaging, and a minimum resolution. A measure of how close the Earth is to the high-frequency limit is the wave-parameter, D, as discussed in the theory section. Given our results, it is around D=40 in the upper mantle and D=30 in the lower mantle. Obviously we would have to consider diffraction to obtain the whole structure, hence in the ray limit we only image a smoothed version of the real Earth. Second, the forward model includes assumptions that are not satisfied by the results.

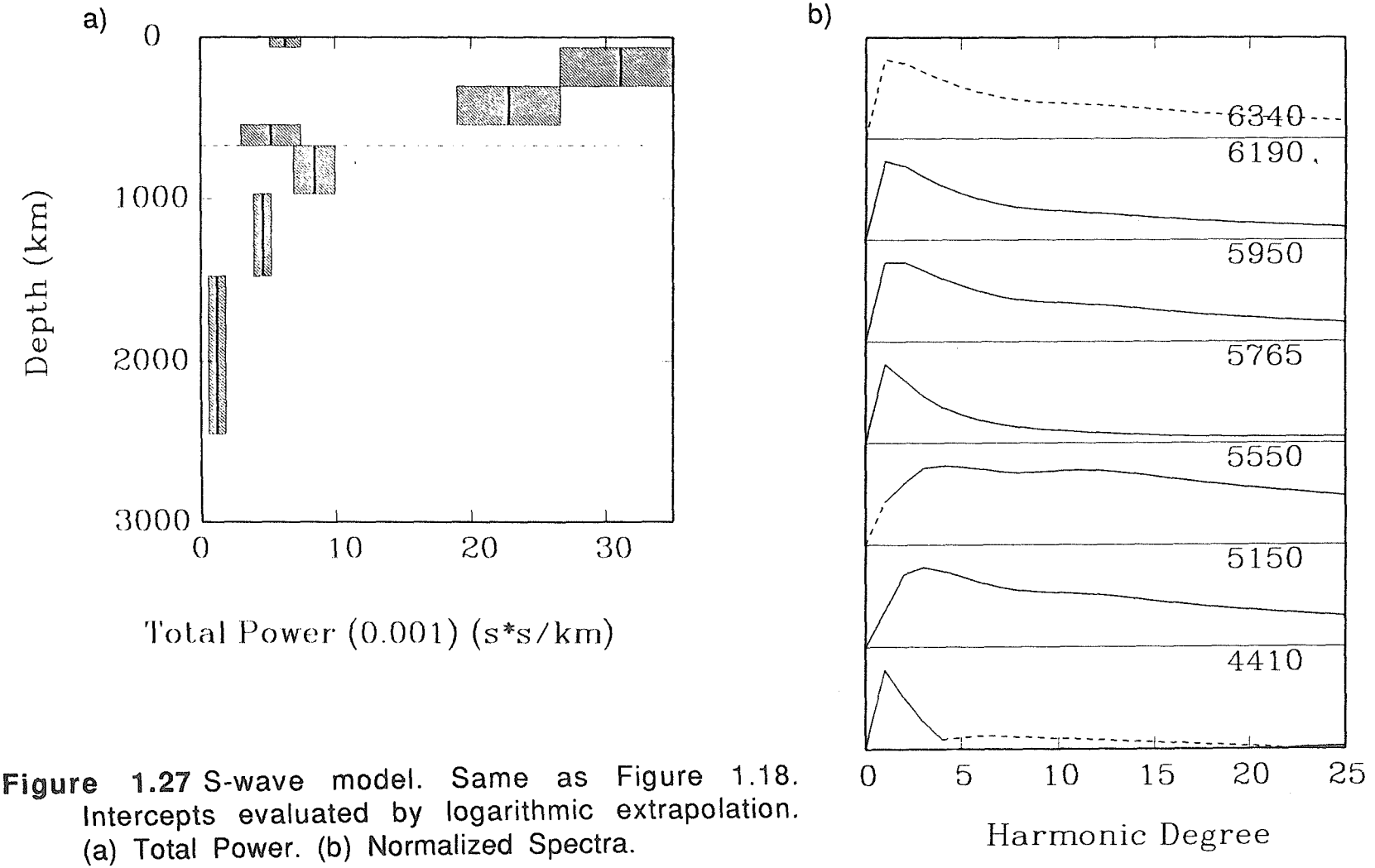


Figure 1.27 S-wave model. Same as Figure 1.18. Intercepts evaluated by logarithmic extrapolation. (a) Total Power. (b) Normalized Spectra.

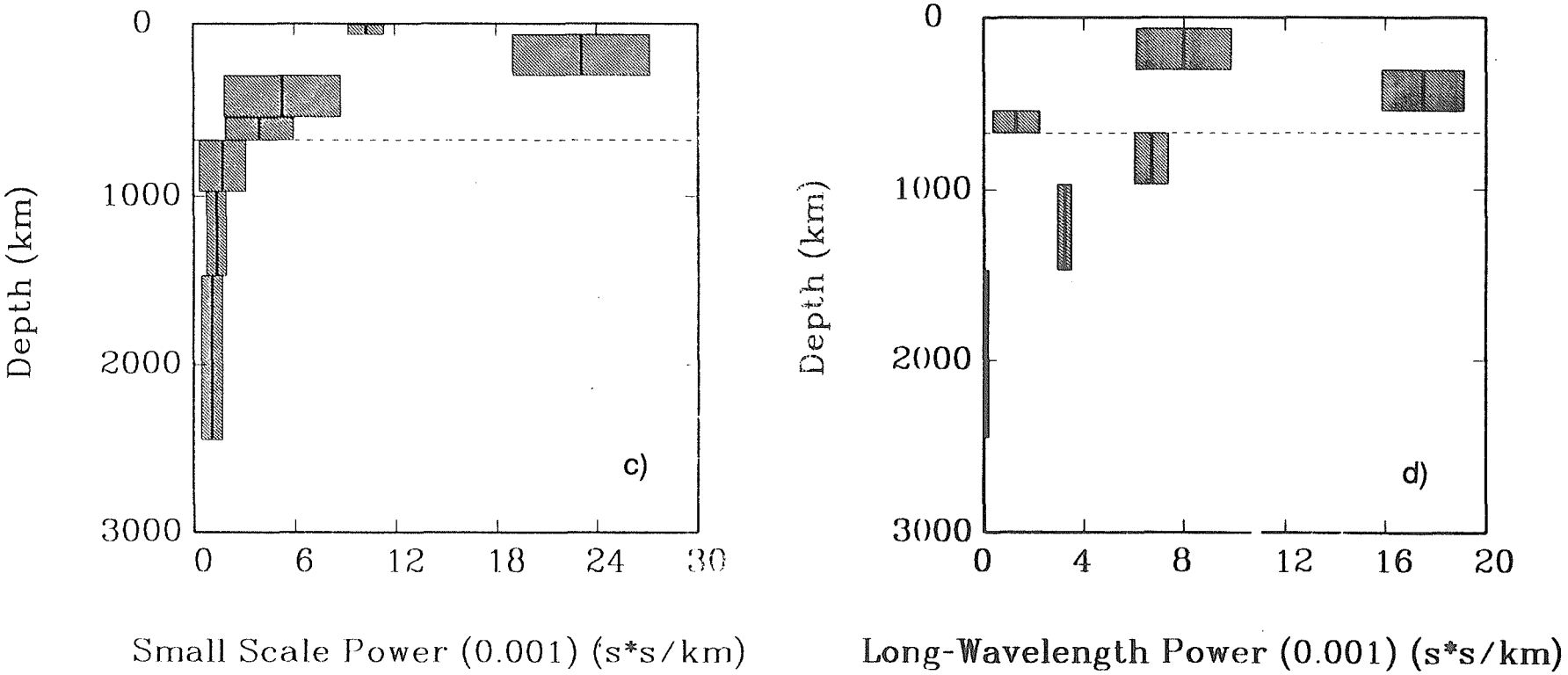


Figure 1.27 S-wave model with intercepts derived by logarithmic extrapolation (cont.) (c) Short scale Power. (d) Large scale Power.

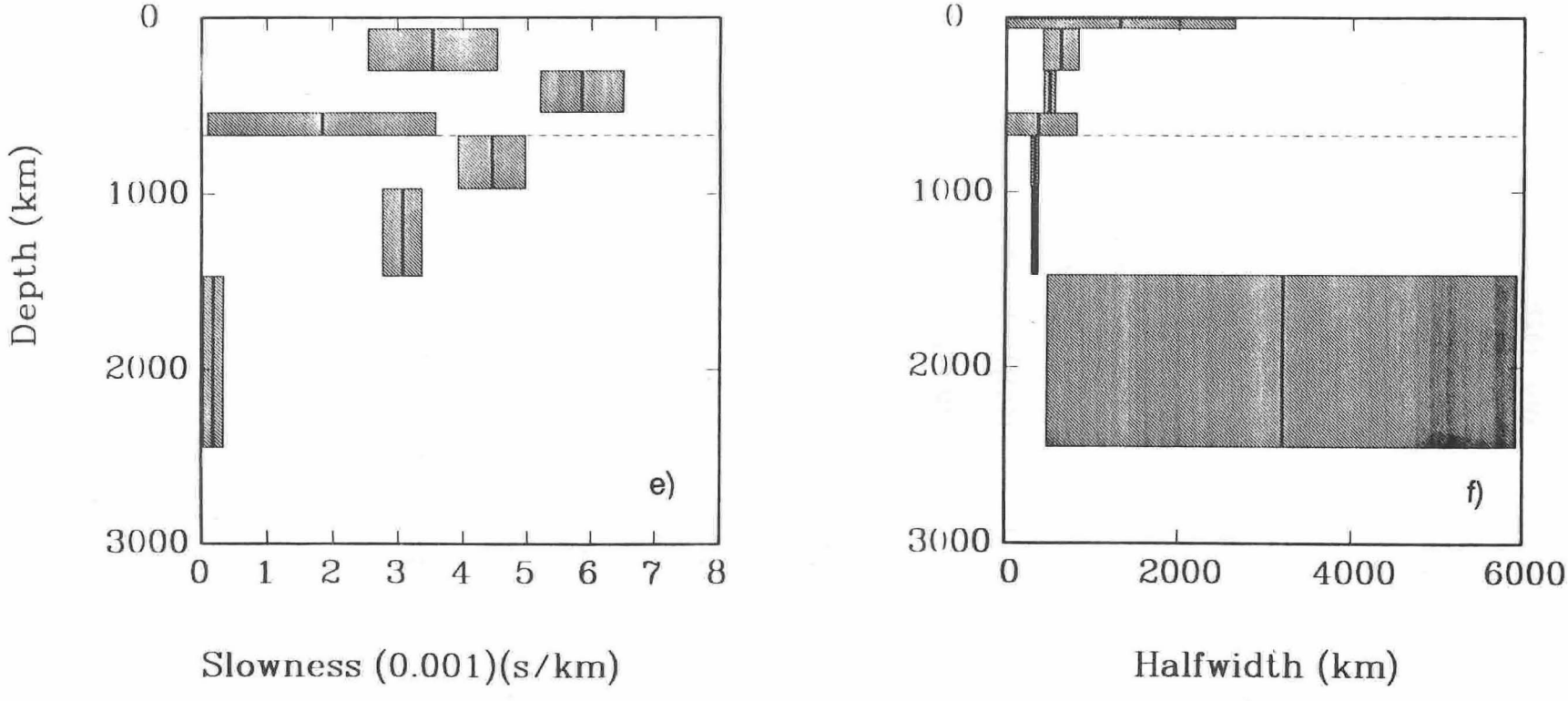


Figure 1.27 S-wave model, with logarithmic intercept estimates (cont.). (e) Slowness. (f) Half-width.

. The size of the heterogeneities in the lower mantle are approximately 1200km or so and this is slightly smaller than the radius of curvature of the rays. This assumption is satisfied, but if the size of the heterogeneities approaches 4000km, then the radius of curvature of the rays is no longer larger than the heterogeneities. With half-widths of 500km defined for most layers but appreciable differences in power attributed to layers 500km and less apart, the assumption of isotropy does not hold; i.e. the Earth seems more heterogeneous vertically than it does horizontally. This is most probably not a large problem since there is less reason to question lateral isotropy, while the theory only requires local isotropy and for small separations the vertical and horizontal variations will seem isotropic. At the largest scales the lack of isotropy potentially becomes significant since here the ray separations become very large, and "local" will need to mean a substantial distance. We restrict ourselves to teleseismic rays in both the S- and P-wave study, all the rays therefore travel nearly vertically through the upper mantle. Hence we are primarily sensitive to lateral slowness fluctuations and vertical correlation lengths in the upper mantle. Third, the assumption regarding ray geometry is questionable at large-scale and only a synthetic study of a wide range of models, using the actual ray distribution, would answer how good it is [*Gudmundsson*, 1989]. Therefore we should be wary of the lowest harmonic degrees.

Caution is also suggested by certain features in the data that are unexplainable by the forward model. In both data-sets the variance frequently decreases at the largest scale. This small but

consistent drop cannot be explained by our model where we require the variance to increase monotonically as a function of scale. From the drop in sampling of large separations at large-scale in the shear wave study (see figure 1.15(a) and the discussion above in the Inversion section), we suggest that this feature might be explained by the fact that the sampling is very different to that assumed by the model. A second possible explanation is that the sampling is biased since the sampling and structure might not be statistically independent. In the S-wave study the effect of this dip was investigated on the final result. Data with the largest scales were lifted so that there was no dip, and then inverted. The resulting model is shown in figure 1.28 and we can see that it is very similar to the preferred model, but there are some differences, these include more long-wavelength power (d); especially near the 670km discontinuity. The deepest radial bin has very low total power (a) and the long-wavelength power is slightly negative. At the very shallowest depth there is positive short-wavelength power (c) where before there was positive long wavelength power; given the small power involved and the poor resolution in this depth bin, this change is not significant. The major significant change is the increase in long-wavelength power, as expected. It is not a major change, but it reinforces the robustness of the pattern of heterogeneity and the lack of robustness in the shallowest layer and the value of the deepest layer. Also it gives us an indication of what magnitude changes possible biased sampling could be introducing; $\approx$ 20% for the depths where the resolution is good and the model appreciable, i.e. other than the top and bottom radial bins. A third feasible but, in our opinion, unlikely ex

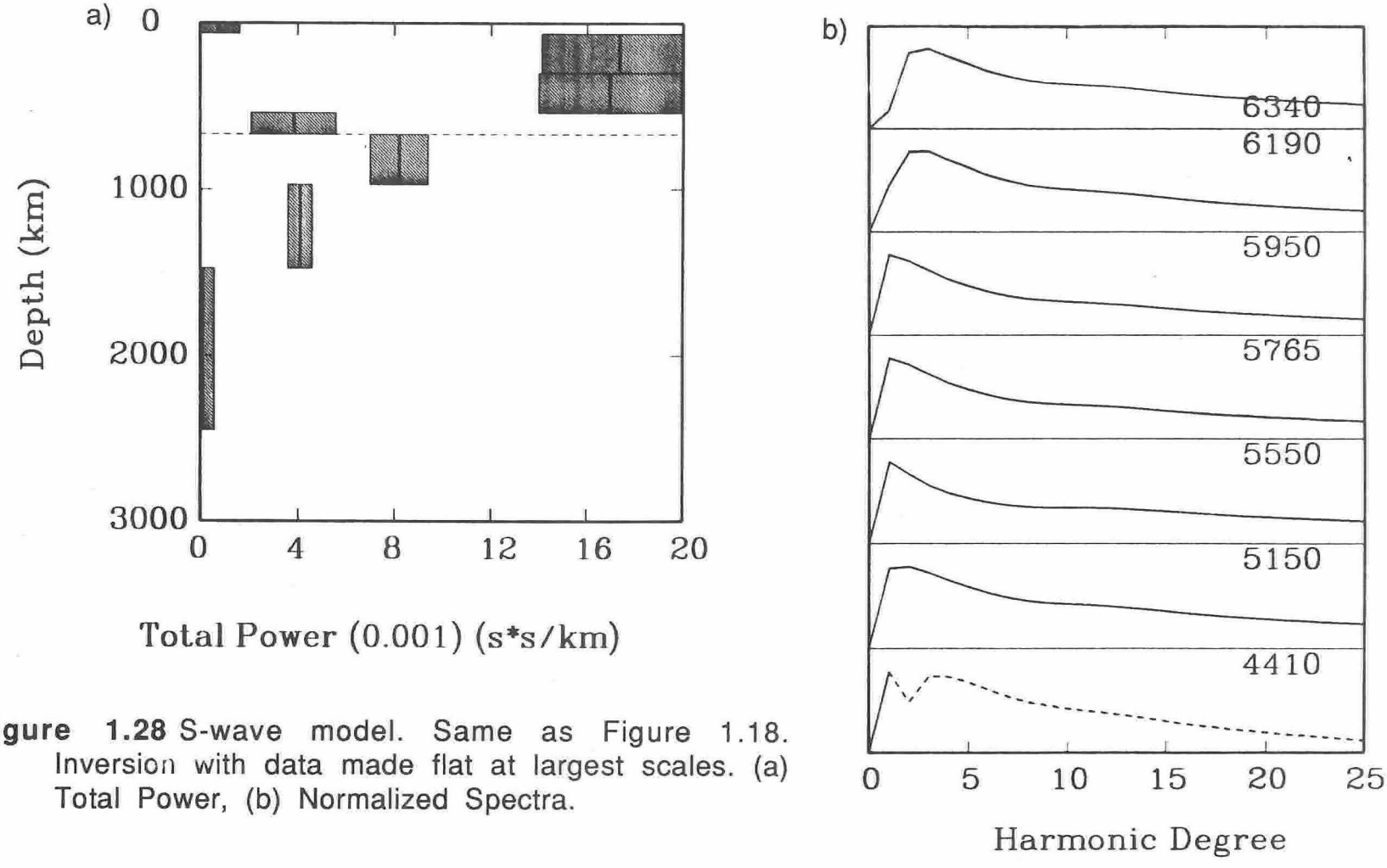


Figure 1.28 S-wave model. Same as Figure 1.18. Inversion with data made flat at largest scales. (a) Total Power, (b) Normalized Spectra.

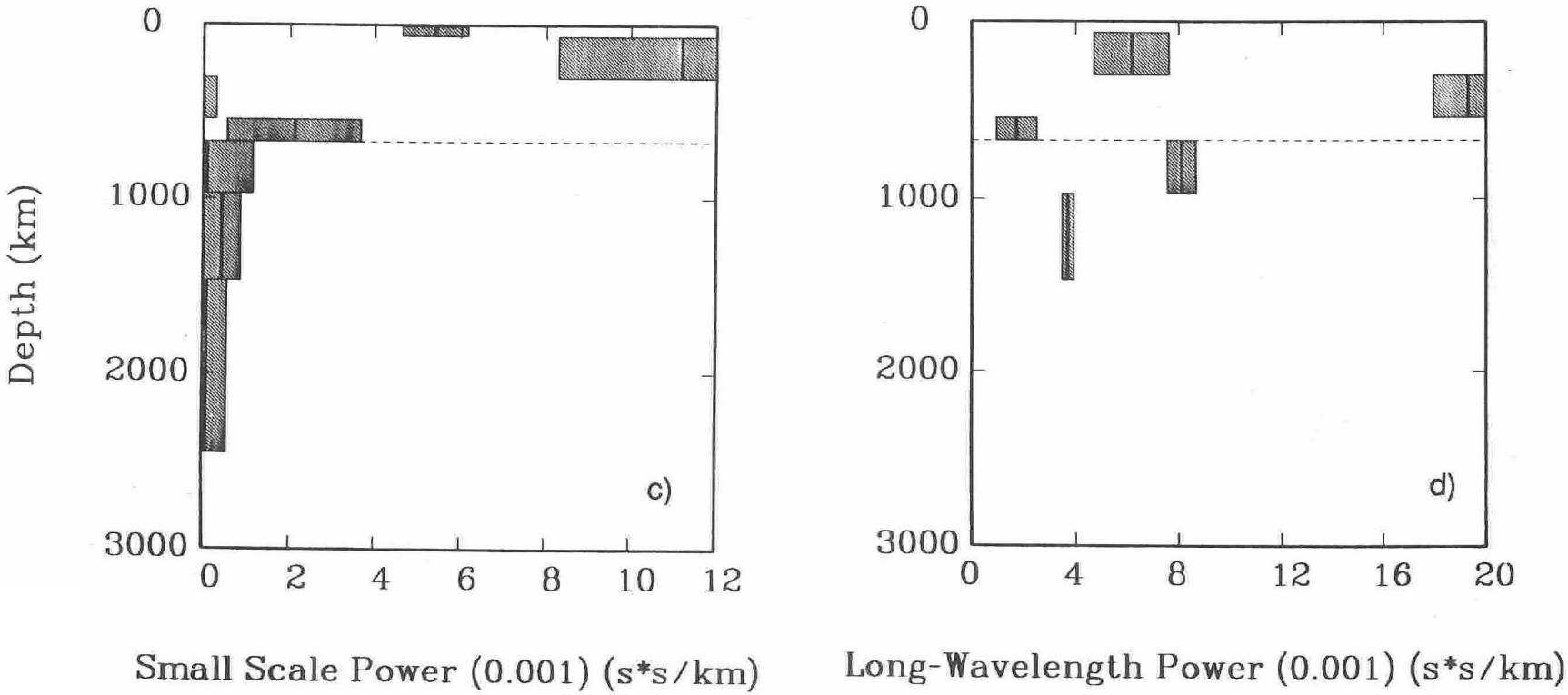


Figure 1.28 S-wave model with lifted asymptotes (cont.). (c) Short scale power (d) Large scale power.

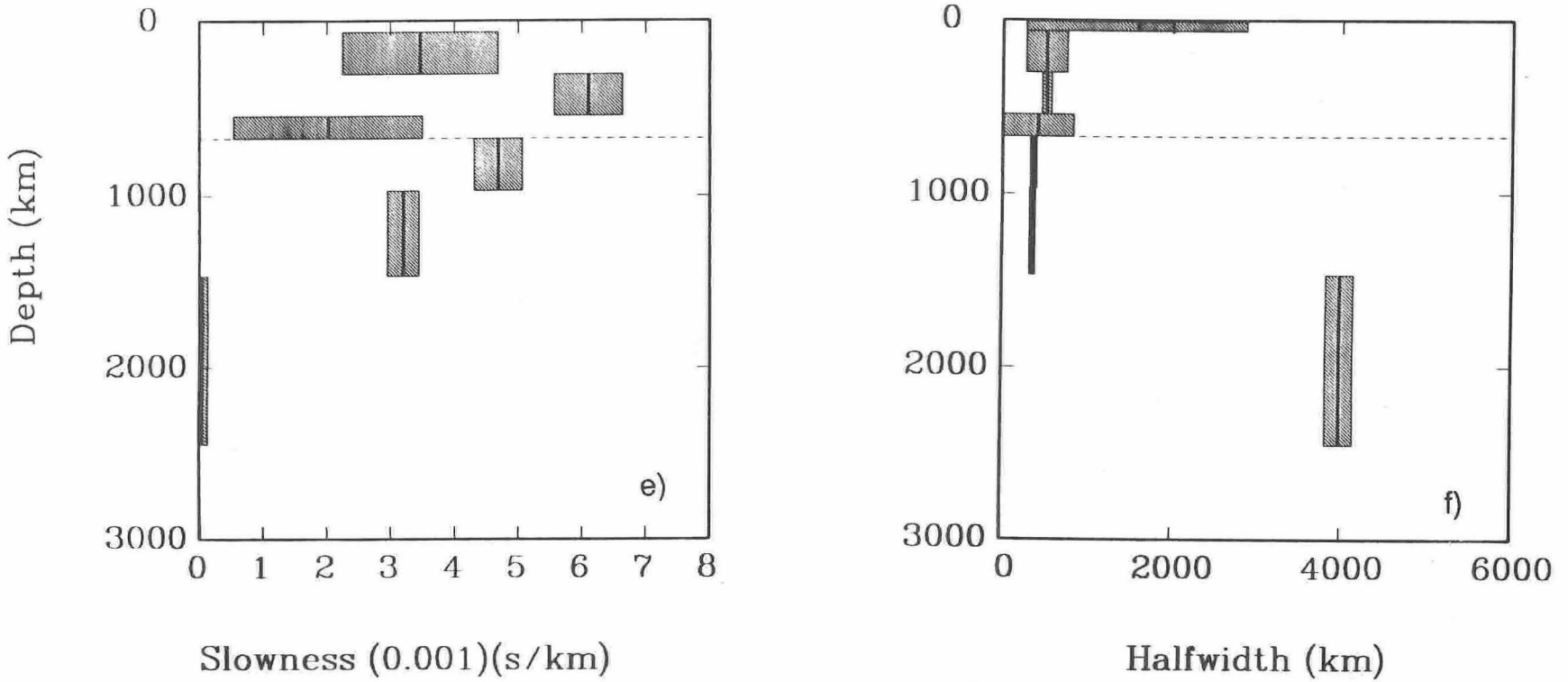


Figure 1.28 S-wave model with lifted asymptotes (cont.). (e) Slowness. (f) Half-width.

planation is that it is a reflection of a very large-scale ($l \approx 2$ or 3) periodic convection pattern.

Another feature of the data unexplainable by the forward model was discovered in the S-wave study, where it was found consistently after removal of the intercept estimate, that the data variance for the shallowest depth bin was less than the next depth bin. Again our theory insists that the variance decrease monotonically as the source depth bin gets deeper. Possibly the intercept estimates are too large for the shallowest depth bin; this might be reasonable if the very shallowest depth has a large amount of very small-scale structure that leads to an overestimate of the intercept. This is feasible. Another explanation is that shallow earthquakes are distributed globally while earthquakes deeper than 34km are restricted primarily to collision zones. Collision zones might be expected to be more heterogeneous than other provinces of the Earth.

Rather than leave one with a pessimistic view of these results, I should like to remind the reader that the simple model which we have applied to two independent data sets has given excellent fits to both data-sets.

In contrast, other global body-wave models usually explain less than 40% of their data variance [*Dziewonski*, 1984; *Hager and Clayton*, 1989]. This is usually ascribed to be the result of a noisy data-set and a limited model. We suggest that the intrinsic signal-to-noise ratio of the ISC data-set is greater than 1:1. At the coarser scales of these models, much of this signal is noise due to structure smaller than the scale of the model. This analysis would allow one to judge whether these previous models are significant relative to their

data errors. Surface wave models have also had low variance reduction [*Tanimoto*, 1989a; *Zhang and Tanimoto*, 1989].

The models derived from both S and P data-sets are very similar in pattern, with only a few minor differences. The S-wave model has more power than the P-wave model, as expected.

All the problems of this method are problems in other deterministic methods. They all suffer from uneven sampling, and even surface waves and free oscillations are not independent of the source and receiver locations. For instance, the lateral resolution is poorest mid-way between the station and event for a surface wave. The power with which different free oscillation modes are excited varies with position of source, while different station distributions are sensitive to different distributions of aspherical structure. It is very difficult to obtain good resolution of shallow structure. No global study has claimed to be able to resolve the upper 60km perfectly. Surface wave studies introduce crustal corrections, body-wave studies remove station statics. Deterministic models suffer also from aliasing in that they restrict the parameterization of the spectral dimension to a low degree, and even the highest resolution body-wave study of *Clayton and Comer* [1983] is equivalent to l=30 or so, well inside our long wavelength definition. We have showed that a lot of the travel time signal is due to much shorter wavelength structure. These coarse models most probably suffer from small-scale power being introduced into the lower harmonics. This effect is limited since it is only the smallest scale-lengths in the models that are most probably contaminated. Most other studies also have to contend with stations that are based on continents, and sources restricted to the

well-known tectonic seismogenic zones. This leads to problems in obtaining global measures of heterogeneity without being biased by these features.

Some of the scatter arises from uncertainty in the earthquake hypocenter parameters, but most of this scatter is scale invariant and is not inverted to produce spurious features in the model. In most other studies, errors in hypocentral parameters map into model errors. In surface-wave studies that model minor and major arc paths, one needs not only the location and origin time but also the source time function of the earthquake. All studies are sensitive to anisotropy, and attenuation. Attenuation can have a large effect on highly dispersive surface waves but does not have a large effect on body waves. Our summary ray formulation is largely oblivious to azimuthal anisotropy since most component rays travel the same direction. At the very largest scales we will consider many orientations, therefore part of the scatter at this scale could be due to azimuthal anisotropy.

1.6 Discussion

We have inverted scatter in P- and S-wave travel-times to obtain a description of the spectrum of lateral heterogeneity as a function of depth through the mantle. The technique is robust in getting estimates of the product of a characteristic scale-length of heterogeneity and the square of the mean amplitude of the slowness (reciprocal of velocity) variations. With less confidence we can evaluate the characteristic scale-length from the spectrum and hence

evaluate the mean amplitude of the slowness variations from the above product. By extrapolating the scatter to zero area we can also estimate the scale incoherent signal that is independent of structure and hence is a plausible lower-bound estimate of noise in the data-set.

1.6.a Intercepts

The estimates of the intercepts for the P- and S-wave studies are understandable from comparison with complexities on the travel-time curves. For the P-wave, studies these are high at regional epicentral distances, the distances at which we have large travel-time complexities from the presence of triplications due to the 400 and 660km discontinuities. Similarily, the intercepts are higher as we approach 100°, here PcP becomes asymptotic to P. Both P- and S-wave intercepts are lower over teleseismic distances away from these complexities.

The data from deeper source depths have lower intercepts; we suggest that this is partly due to a smaller degree of small-scale structure at depth. This is evidenced in the data as a sharper drop in variance as we approach the origin for the shallower curves. There is some evidence of this even in our smallest-scale data, which is actually relatively large-scale (see the original data in figures 1.4 and 1.5). A more subtle effect of small-scale structure arises since component rays of a summary ray originate at different source depths even inside the same source depth bin. If there is heterogeneity at the scale-length of the path difference, then this will in-

troduce scatter. Notice this scatter is independent of scale-length and hence will contribute to the intercept. It is also reasonable to expect more scatter from shallower events, since their arrivals are frequently complex and emergent, while deeper events generally have simpler impulsive arrivals. This could be a function of the source and also the complexity of near source structure.

If we assume that the depth variation of the intercept is primarily a function of small-scale structure and finite binning, then it is reasonable to take the deep source bin intercept estimate as an estimate of the non-structural signal in the data. For teleseismic P-waves this is $\approx 0.25s^2$ and $\approx 5s^2$ for the S-waves. Estimates of the structural signal are around $\approx 1s^2$ for P-waves and $\approx 12s^2$ for the S-waves. We estimate the signal-to-noise ratio for P-waves as $\sqrt{1}/\sqrt{0.25} = 2$; and $\sqrt{12}/\sqrt{5} = 1.5$. for S-waves.

1.6.b Discussion of seismic velocities

The most striking result of this study is the concentration of power in the upper mantle compared to the lower mantle and the similarity in the results for the shear and compressional wave structure. Given the fact that for teleseismic distances the lower mantle path length is appreciably larger than the upper mantle path length; the contribution of the lower-mantle on travel-time residuals will be larger than the low total power (the product of the half-width and the mean square amplitude of the slowness variations suggests [*Zhou et al.*, 1988]). If the scale-lengths of the heterogeneity were the same in both regions, then the amplitude of slowness fluctuations

would equally be greater in the upper than lower mantle. There is a suggestion in the results that the half-width in the lower mantle could be possibly 2 or 3 times larger. This is better resolved in the P-wave study than in the S-wave study. This decreases the relative amplitude of lower mantle slowness fluctuations compared to upper mantle slowness fluctuations, as based solely on total power. The results suggest that the large scale-length slowness fluctuations in the mid-upper mantle are more than an order of magnitude larger than the large scale-length fluctuations in the mid-lower mantle.

A comparison of the P-wave results in the lower mantle are made with the statistics of previous P-wave lower mantle deterministic studies in figure 1.29. It is seen that the lower mantle comparison is reasonably favorable. There are some variations in absolute magnitude with both deterministic studies showing higher heterogeneity than discovered in our statistical study. This additional signal could be the result of the aliasing of higher spectral power into the lower harmonics, spurious power due to inversion of noise into poorly resolved regions of the model, or leakage into lower mantle due to poor depth resolution in the deterministic inversions. Equally, it could be caused by insufficient power in the spectral model due to leakage of power out from low harmonic degrees into the higher harmonics. It is certain that all these effects are factors, but we have special reason to suspect the inversion of noise given the large noise level, (the results of *Dziewonski* [1984 #177] show large power in regions of the Earth where he has little or no coverage, and that the model of *Clayton and Comer* [1983 #122] has a large amount of high-frequency power (spectra are nearly white), hints that noise

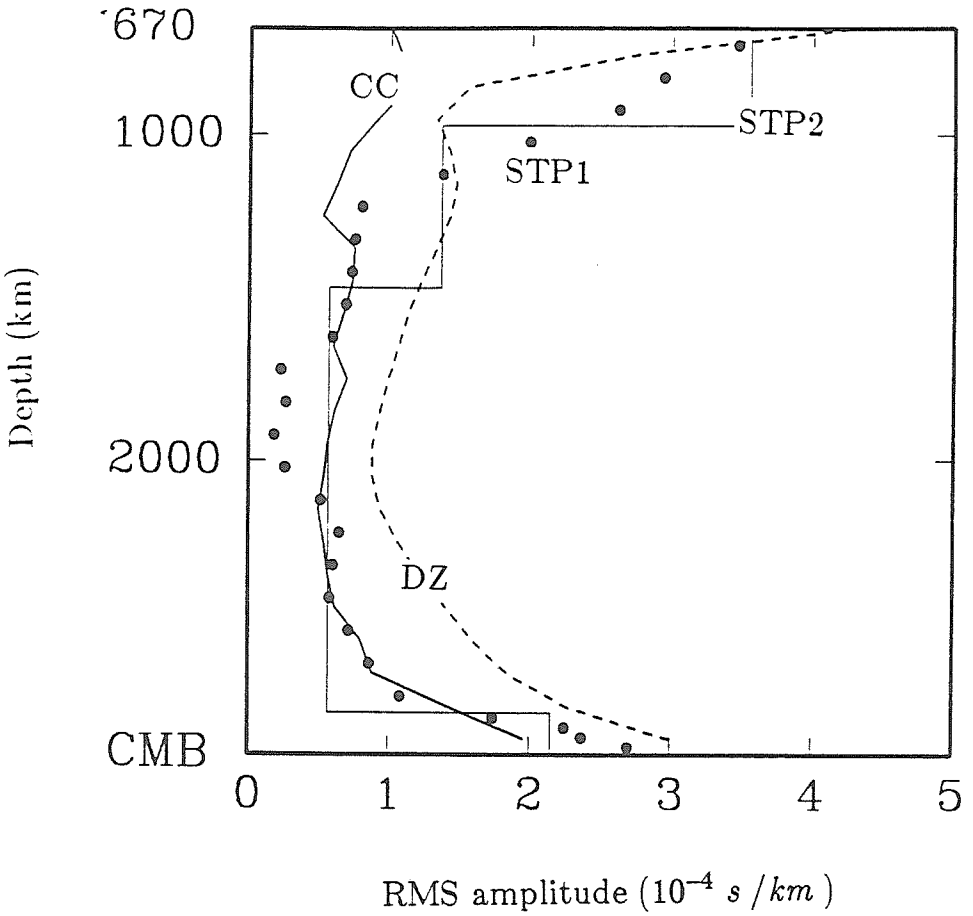


Figure 1.29 Comparison of stochastic P-wave models with deterministic P-wave models. CC refers to *Clayton and Comer* [1983] while DZ refers to *Dziewonski* [1984]. STP1 is the finely parameterized P-wave stochastic model while STP2 is the coarsely parameterized P-wave stochastic model. From *Gudmundsson* [1989].

may be entering these models) and the harmonic leakage due to the imperfect spectral resolution of stochastic method.

The only previous global shear body wave analysis of the lower mantle was *Sengupta and Toksoz* [1976]. They found lower mantle relative perturbations of around 1.2%. In a subsequent study [*Sengupta et al.*, 1981] they suggested that their shear wave data was not good enough to warrant a 3D model. *Tanimoto* [1989a] has derived a long-wavelength whole mantle shear wave model using surface waves, and the depth resolution is provided by including overtones as well as fundamental modes. He finds a minimum in power at around a depth of 2000km, with variations of the order of 0.7% peak to peak, i.e. mean amplitude variations of $\approx$ 0.2%. This study is out to angular order 6, scale-length of order 5,000km. The S-wave stochastic study suggests variations of order 0.1-1.0.% mean amplitude; with a half-width ranging from 500-5,000km. So our results (l =50) span Tanimoto's results (l =6), but if we assume that the shear-wave lower mantle variations are of a similar wavelength to the P-wave variations i.e. $\approx$ 1,200km, then the velocity variations are larger than Tanimoto's, (as expected, given the higher harmonic degree) at around 0.4% mean amplitude.

The global studies of *Dziewonski* [1984] and *Clayton and Comer* [1983], both removed most of the upper mantle structure by removing station statics from the travel-time residuals. This was to prevent leaking this poorly resolved signal into the lower mantle, which was the primary target of both studies. In contrast, the global P-wave study of *Inoue et al.* [1989], has not subtracted out station statics. They obtain r.m.s. slowness perturbations in the upper

100km of up to 0.7%, decreasing to 0.55% at 150km and 0.36% at 350km. This compares to our P-wave study, which suggests mean slowness perturbations of around 0.4% between 60 and 300km depth. Both of these studies have poor vertical resolution in the upper mantle. High resolution compressional wave studies of upper mantle regions by body-wave waveform modeling techniques and wave-field continuation techniques suggest that the difference between the velocities below the Canadian [*Le Fevre and Helmberger*, 1988] and Kola [*Given and Helmberger*, 1980] shields and those found beneath the Gulf of California [*Walck and Clayton*, 1984] suggest peak to peak variations of ≈ 5%; i.e. mean amplitude variations of ≈ 1.3%, and hence mean slowness variations of 1.3%. More notably though, the P-wave stochastic model finely parameterized in depth shows a dramatic drop of a factor of 2 in slowness variations over the depth range of 300-400km. All the high resolution studies have similar results below around 300km. If the high resolution results above show that oceans and cratons are different down to 300-400km, then our result tends to suggest that this might not just be limited to the areas studied above but might apply globally. This logic suggests that the seismic signature of oceans and cratons does not extend much below 300km. This is intermediate between the deep tectosphere beliefs of *Jordan* [1975; 1978; 1981] ≈400km and the thin plate beliefs of *Anderson* [1979a], and *Sclater et al.* [1981] ≈200km.

Presently there are no global body-wave upper-mantle shear-wave studies yet. There are a number of regional shear-wave models derived using high-resolution waveform modeling. For in-

stance, *Grand and Helmberger* [1984b] derived a model for the Canadian shield (SNA) and a model (TNA) appropriate for younger oceanic structures and the more tectonic parts of western North America. These show peak to peak variations of around ≈ 11% at around 100km depth; i.e. mean slowness amplitude variations of ≈ 2.5%. *Grand and Helmberger* [1984a] similarly derived a model for old ocean in the North Atlantic (ATL). This showed its largest differences from SNA at around 150km depth. They correspond to mean slowness variations of ≈ 1.5%. The difference between the oceanic/tectonic and shield models decreases linearly from ≈ 1% at 220km to zero at 400km. These were reconfirmed in cross-sections from oceanic/tectonic to shield regions [*Graves and Helmberger*, 1988; *Helmberger et al.*, 1985] and in a tomographic analysis of the mantle below the North American Plate [*Grand*, 1988]. The shear wave stochastic results suggest mean shear wave slowness variations of 1.6%, between 70 and 300km, and 2.9% from 300 to 600km. The seeming discrepancy is largely the result of imperfect spectral resolution in the stochastic study. This places nearly all the power between 300 and 600km in the long-wavelength spectrum while the majority of the power between 60 and 300km is placed in the short-wavelength spectrum, which isn't included in evaluating the long-wavelength slowness. Considering the uncertainty in whether the spectral resolution of the stochastic and waveform modeling technique is comparable, this comparison suggests that their results are not incompatible.

The above high-resolution shear wave studies were regional. Global comparisons can be made with the global upper mantle

shear wave models derived from surface wave studies. Our shear wave mean amplitude slowness estimates of ≈ 0.006s/km correspond to ≈ 3%. These would suggest ± 6% peak variations (i.e., ≈12% peak to peak) as compared to the ≈ ± 3% variations found in surface wave studies [*Dziewonski and Woodhouse*, 1987; *Nataf et al.*, 1984; *Tanimoto*, 1986; *Tanimoto*, 1987; *Tanimoto*, 1988; *Tanimoto*, 1989a; *Woodhouse and Dziewonski*, 1984]. The surface wave studies are limited in spectral degree to harmonic expansions out to, at most, degree 8, while the long-wavelength stochastic model extends to degree 50; hence one would expect more power in the stochastic study. These surface wave studies show a similar, if not so dramatic decrease in heterogeneity in going from 200 to 450km depth in the mantle. Note these studies have poor depth resolution (200km); similar to the stochastic analysis. Others have used fundamental and higher mode Rayleigh waves to illustrate the contrast between continent and oceans [*Lerner-Lam and Jordan*, 1987] and find that combined with travel-time variations of ScS [*Sipkin and Jordan*, 1976], that they can only be satisfied if the structure of continents extends to below 220km [*Jordan et al.*, 1989].

Free oscillations have also been used to estimate heterogeneity. *Masters et al.* [1982] study of spheroidal fundamental mode free-oscillations suggested a large quadrupolar pattern in the transition region. Splitting of free-oscillation multiplets has also been used to invert for aspherical structure, e.g., [*Ritzwoller et al.*, 1986]. They found perturbations in density in the lower mantle of up to 0.35%, and using their correlation of density variations to shear wave and compressional wave variations they suggest variations of up to 0.9%

in S-wave and 0.7% in P-wave. These are for only the longest wave-length variations and they have poor depth resolution in the lower mantle. They do have the advantage of averaging global volumes and are useful in comparing to higher resolution studies such as body-wave studies, where potential problems with biased sampling are much more prominent. *Giardini et al.* [1988] have also looked at the splitting of long-period normal modes. They show that the splitting functions are consistent with previous models of heterogeneity in the mantle [*Dziewonski*, 1984; *Woodhouse and Dziewonski*, 1984] provided that $d\ln(V_P)/d\ln(V_S) \approx 0.4\text{-}0.5$. Hence, to the degree that we correspond to the results of the above quoted studies, so we agree with the results for splitting of normal modes.

The small-scale power is very large in the upper 300km in both studies. If we suggest that a reasonable estimate of half-width for these heterogeneities is of order 10-50km (as suggested by array studies) then we find that the small-scale power is equivalent to velocity heterogeneity of order 1 - 4% for P-waves and of order 1 - 3% for S-waves. This compares favorably with the heterogeneity found in the array studies mentioned above. The rapid drop off of small-scale power at a depth of around 150km in P-wave is similar to the limit of the random medium from the array studies. Since we only have a coarse parmeterization in depth for the S-waves, we cannot resolve where this drop-off occurs and neither can we confirm that the average quoted above for 60-300km depth is actually representative of the variations above 150km. These figures from the stochastic study can be seen to be consistent with array studies but do not have the spectral or depth resolution or the data quality

(due to the intercept ambiguity) to be definitive. It is comforting that this method is compatible with the array studies and suggests that given better coverage this method can give the whole spectrum from the scale of the intrinsic resolution of the waves used to the whole Earth!

The results of both P- and S-wave studies suggest that the degree of small-scale structure in the lower mantle is indistinguishable from zero. This might not be a true reflection of the real Earth but rather of the wave propagation; i.e. the intrinsic averaging and wave-front healing discussed at length above. Using the coarse approximation of *Nolet* [1987] p.10 for the width of a seismic ray of $w \approx \sqrt{(\lambda L/8)}$, we find that the ray-width in the top of the upper mantle is around 70km, (equivalent to $l \approx 300$) increasing to 120km in the deep mantle ($l \approx 170$). Hence, most small-scale structure ($l > 50$) from the lower mantle, could be imaged by this technique. To convert harmonic degree to an equivalent length we have used the following approximation, length= $2\pi\ r/\ (l + 0.5)$, where r is the radius of the layer at which the conversion is made and l is the harmonic degree. Note also that for array studies a ray-width of 10km is achieved at a depth of 80-100km. Hence, drop-off of small-scale structure with depth below arrays could be a function of ray width being larger than these heterogeneities.

1.6.c Comparison of P- and S-wave studies

A comparison of the P- and S-wave studies is very interesting. Most striking is the consistency in the pattern of the radial

variation of lateral heterogeneity in both studies. They both show large variation in the upper 600km, decreasing as we approach the lowermost mantle. Both studies show a decrease in the power of lateral variation just above the 670km discontinuity as opposed to below the 670km discontinuity. The region just above the 670km discontinuity is poorly resolved. Only the consistency of this feature in both studies gives us encouragement as to its reality. Even the most poorly resolved region of both models, the uppermost 60km, has consistently less long-wavelength power in both studies. Short-wavelength power in the S-wave study is low while it is very high in the wave study. The shallow short wavelength power is what trades off with the intercept estimates. Hence, given the uncertainty in the intercept estimates, we expect the estimation of these parameters to be least robust.

The biggest deviation from correlation of S- and P wave power is the depth at which the peak occurs in the long-wavelength power. The P-wave study shows a peak at 60-300km depth; while the S-wave study shows its peak at 300-540km depth. Given the poor spectral resolution, it is possible that we have placed the power incorrectly into the short and long-wavelength spectra for these two depth bins. In the P study it is seen that the radial pattern of lateral variation is similar for both long and short-wavelength components. In comparison, the shear-wave pattern is different at the two different wavelengths for these two depth bins. The radial pattern of lateral variation of the sum of both components is very similar to the pattern exhibited by all the P-wave results. This gives us reason to believe that due to the lack of spectral resolution, we have placed too

much of the power between 60 and 300km into the short-wavelength component; while between 300 and 540km we have placed too much power into the long-wavelength component (in fact the short-wavelength component at this depth is negative). This separation of power at these depths is quite robust to the choice of damping parameters, and the spectra are very similar out to degree 25, hence the difference is due to relatively short-scale structure.

Since the radial pattern of variation is most closely correlated between the total power of both studies, we consider the ratio of this measure to be the most robust. Ratios are notably unstable, and we must be cautious in our interpretation. Since the shallowest layer is most poorly resolved, we will not consider this layer. Given the large relative error and the lack of power in the lowermost layer, this layer also has a poorly defined value for the ratio. The ratio of the total power in the S-wave compared to the P-wave study is largely constant from 540-1400km, and is around a factor of 80 ± 30. The half-widths are not too dissimilar between the two studies except for the lowermost mantle where the S-wave estimate of half-width is not very robust. The shear wave estimate of half-width are slightly lower than 400km, while the compressional wave estimates are slightly larger than 500km. Even assuming that the half-widths are actually similar (the difference could be due to the poorer spectral resolution of the shear wave study) for both studies, the ratio of the mean amplitude of slowness variations will be $\sqrt{80} \approx 9$. Hence, $\delta \ln(V_S)/\delta \ln(V_P) \approx 5$! From 60-300km we estimate $\delta \ln(V_S)/\delta \ln(V_P) \approx$ 4.5 and $\approx$ 7 from 300-540km.

These results are all much larger than found in previous studies. Values of $\delta\ln(V_S)/\delta\ln(V_P) \approx 2$ have been found previously. These include comparison of lower mantle body- and shear-wave studies [*Davies and Clayton*, 1986], comparison of P and S station statics [*Davies and Clayton*, 1986; *Doyle and Hales*, 1967], and the best conversion factor to convert a lower mantle P-wave model to an S-wave model to give the best match for free-oscillation splitting functions [*Giardini et al.*, 1988]. Can we be wrong? We have selected the more stable total power measure and assumed that the actual estimates of shear wave half-width are too low; both of these will tend to reduce estimates of the ratio. What about errors in the estimates of power? We could have estimated intercepts for the shear wave study that were too small and are inverting noise into the model. Our intercept estimate method is identical for both studies and the S-wave estimate is, if anything, the larger estimate since the smallest scale used is larger 1° versus 0.5° for the P-wave study. The actual value of the intercept is irrelevant for the longer scale-length power where we would get similar estimates of the ratio. Here it is only the increase in scatter as we go from one scale to the next that matters; this is demonstrated with the similarity of the long wavelength results following subtraction of the logarithmic or linear intercept estimates, i.e., figure 1.18 and 1.27. The error in the individual studies is large, at the outside they are 30%, hence the error in the ratio might be 45%. Even given these error bounds a ratio of 2 is improbable. Evoking high errors also leads to the problem of needing to explain why the shear wave results are too high for all the layers while the compressional results are too low for all the layers.

Unfortunately there are two weaknesses in the above studies, which could translate into a biased estimate of this ratio. Both studies do not have the same depth or spectral resolution. In particular, in the shear wave study the shallowest layer is very poorly resolved and its value is very heavily damped. This most probably leads to much of its signal being pushed slightly deeper into the mantle. The resolution of the P-wave study at this depth is also poor but is substantially better than the S-wave study. The S-wave study in comparison to the P-wave study has comparatively better resolution of the short-wavelength components compared to the long-wavelength components. This leads to the S-wave study suggesting a shorter half-width, and to explain the same variance it must have increased slowness. It's unclear whether the ratio is larger than 2 from a depth of 60-1400km; in the following discussion we shall assume so, but the reader is reminded that the last two issues raised bring it into doubt.

If the mantle was solid and homogeneous and the variations in seismic velocity were due to temperature, then one would expect the ratio to be around 1 from low-pressure laboratory measurements of prospective mantle minerals [*Sumino and Anderson*, 1984]. Hence, there was a problem in explaining values of the order of 2. The explanation proposed for explaining a ratio as high as 2 in the upper mantle has been that the mantle was homogeneous but that it was partially molten. Shear wave velocities decrease markedly in partially molten regions while compressional velocities are much less sensitive. Consider that all the velocity variation was

due to changes solely in the shear modulus (the rigidity); then it can be shown that

$$\frac{\partial \ln V_s}{\partial \ln V_p} = \frac{3(K + 4\mu/3)}{4\mu} \tag{1.66}$$

For a Poisson solid $K=5\mu/3$. This would suggest $\delta\ln(V_S)/\delta\ln(V_P) = 2.25$. This explanation requires the presence of variable amounts of partial melt everywhere. If there was no partial melt over a region large enough to be investigated for its lateral variation, then one would expect us to derive a value closer to 1 for the above ratio. For the uppermost mantle this explanation is seemingly not unreasonable given estimates of solidi of mantle assemblages and mantle geotherms. There is also clear evidence of partial melt of mantle origin at the mid-ocean ridges. Comparison of seismically-derived mantle seismic velocity and attenuation values with laboratory measurements suggests that the presence of partial melt is restricted to ocean plates younger than 40Ma [*Sato et al.*, 1989; *Sato et al.*, 1988]. High resolution studies discussed above show evidence of a low velocity zone even beneath shield regions. *Sato et al.* [1989] have also shown that the velocities decrease appreciably before there is any melt present. *Anderson and Minster* [1979b] had earlier argued for this behavior from relaxation of dislocations. The small low-velocity zones below shields needn't be evidence for a partially molten layer. Therefore, it seems improbable that partial melt is present globally. The experiments of*Sato et al.* [1989] show that as we approach the solidus, (they most probably can only iden-

tify ≈ 1% melt though) the relative decrease in shear velocity is approximately equal to the decrease in the compressional velocity (see their fig 3A). Hence, near solidus temperatures cannot explain a ratio of 2.

How does one explain a ratio of 5? With difficulty! The most probable explanation as discussed above involves the poor spectral and depth resolution, especially of the shallowest layers that are heavily damped. If more of the S-wave signal was placed in the shallowest rather than subsequent depths, the ratio might be lowered in the upper mantle but it cannot be made as low as 1 everywhere. The fundamental point is that signal of order $8s^2$ must be modeled for S-waves and signal of order $1s^2$ for the P-waves; if the scale-length of the variations are similar (a large difference would require the two fields being uncorrelated, which would be just as significant since it would demand mineralogical / compositional heterogeneity) this suggests that the signal averaged ratio will need to be 2. There is no way around this, hence there is no way of avoiding large ratios somewhere in the Earth.

Temperature variations of an homogeneous material are seemingly inadequate. Variations near the solidus also seem inadequate. Anyway, it seems improbable that there is partial melt present everywhere from a depth of 60-1400km. Note if $K >> 5/3\ \mu$ (i.e., not a Poisson solid), then in equation 1.66 the ratio becomes $=3K/4\mu$. Hence a ratio of 5 would require $K \approx 7\mu$. *Anderson* [1987a] has argued that the value of $\delta\ln(V_S)/\delta\ln(V_P) \approx 2$ suggested for the lower mantle implies that at high pressure the variations of seismic velocities are primarily due to changes in rigidity. At low pressure and high tem-

perature he claims that the effect of temperature on the velocity is primarily mediated through changes in volume (so-called extrinsic contribution). At higher pressures he suggests that this extrinsic effect is suppressed such that temperature effects are reduced and restricted to the rigidity. This has derived further support from *Agnon and Bukowinski* [1989] using lattice dynamic and molecular dynamic calculations of elastic properties as a function of depth and temperature. Our estimates of this ratio in the lowermost mantle is closer to 2 (though with large uncertainty) and the ratio seems to decrease from the high levels of ≈ 5 for the upper half of the mantle to a value closer to 2 in the lower half of the mantle. This effect could be present throughout the whole mantle, but it would not explain the higher ratio being at shallower depths. I suggest that the variations could be the result of mineralogical variations, due to either phase changes or compositional variations. This would require that mantle minerals have a wider range in shear velocities compared to compressional velocities. This is not true for low-temperature, low-pressure measurements. There are no velocity measurements of minerals at high temperatures and pressures. A variation in mineralogy with depth through the mantle would not be unreasonable with these variations being more readily stabilized in and near the cold rigid upper boundary layer. The lower boundary layer might be more unstable due to its lower viscosity and its higher temperature; dregs might be re-entrained into the mantle more easily [*Davies and Gurnis*, 1986] and the layer might contain higher velocities as material feeds into expected uprising ridges/plumes. This might explain

the possible decrease of this ratio away from 5, towards 2, with increasing depth.

What is the mineralogical variation that we might postulate? There are different candidates, and it might be a combination of different candidates, with different ones being dominant at different depths. At shallow depths, oceanic and continental mantle are probably different. Continental mantle is lighter [*Tanimoto*, 1989b], cooler [*Pollack and Chapman*, 1977], contains more incompatible elements, and most probably is more refractory and depleted in volatiles [*Jordan*, 1979]. Continents have thick crusts (up to 70km) compared to oceans (6km) and continental crust is very silica-rich compared to mantle. Sub-continental mantle might equally be enriched in silica relative to oceanic mantle. The mantle undergoes adiabatic decompression and melting near mid-ocean ridges, hence different portions of the oceanic mantle will contain different amounts of the basaltic component. Garnet and clinopyroxene melt first, hence the residues after melting will be poorer in these minerals. The volume of olivine and orthopyroxene will vary less since they melt at higher temperatures. The seismic velocity of garnet is appreciable faster and while diopside is appreciably slower compared to olivine or orthopyroxene [*Duffy and Anderson*, 1989], at low pressure, though this is true for both the P- and S-wave velocity. Hence we require that the shear velocities of different minerals at high pressure will show more variation, while the P-wave velocities will become more similar. *Anderson*, [1988] shows that the relative variance of the dimensionless logarithmic anharmonic {DLA} parameters among ionic solids is

larger for $\{G\}_T$ than for $\{K\}_T$. The rigidity has 2.5 times more variance than the bulk modulus.

Since the earliest days of plate tectonics the descending oceanic crust has been recognized as a major intrinsic heterogeneity. This transforms to eclogite at depth; therefore regions of high subduction might be expected to have a higher proportion of eclogite and also larger amounts of residue with the basalt component removed. Another contributor to an inhomogeneous mantle is the presence of temperature-sensitive phase changes. These include the olivine to β-spinel, β-spinel to γ-spinel, basalt to eclogite, partial melting, orthopyroxene to majorite, clinopyroxene to garnet, garnet-majorite to ilmenite solid solution, ilmenite to perovskite, and γ-spinel to magnesiowüstite and perovskite. Most of these changes have non-zero Clapeyron slopes, and given the lateral temperature variations in the mantle, the phase changes will be achieved at different depths in the mantle. These provide substantial lateral differences in seismic velocities, especially near the large temperature gradients of subducting slabs. *Anderson* [1987b] discusses their importance not only for interpretation of seismic velocities but also for their dynamical effect since density changes due to phase changes can be much greater than those due to thermal expansion. The only samples from the mantle are shallow, no greater than ≈ 200km. Two main classes of xenoliths are found; most commonly peridotitic xenoliths (mainly olivine and orthopyroxene) and less commonly eclogitic xenoliths (mainly clinopyroxene and garnet). Hence there are surface observations of major element variations in the mantle. Also, isotopic studies show large degrees of heterogeneity at all

length scales [*Hart and Zindler*, 1989; *Zindler and Hart*, 1986]. As discussed below, it is uncertain whether these are related to major element variations or are a reflection of processes that are independent of the major element distribution.

There is no shortage of candidates for mineralogical variation. The iron-to-magnesium ratio might vary with depth *Anderson*, [1989]. "Mineralogical" variations could lead to the shear fluctuations being much larger than the compressional fluctuations only if the compressional velocities of different component minerals become more alike at high pressure and temperature compared to their shear velocities. We suggest that this might be expected at high pressure since all minerals go to higher co-ordination and in fact most upper mantle minerals end up transforming to perovskite. At higher pressure they become stiffer in compression since the ions are already pushed together at the steep part of their potentials, which gives them repulsion strength. In contrast, under shear the structures undergo deformation without a reduction in volume, and hence without necessarily approaching the same steep part of the electrostatic repulsion as under compression. As noted by *Jeanloz* [1989], the anion-anion distance decreases with increasing co-ordination but anion-cation distances can increase and hence structures could be weaker. This has also been discovered in *ab initio* potential calculations. Shock wave experiments through equation of state modeling can give estimates of bulk moduli as a function of depth [*Jeanloz*, 1989]. They can also be measured indirectly using x-ray diffraction of samples inside diamond anvil apparatuses. Unfortunately, the shear modulus is not a thermodynamic quantity (it is not related to a

thermodynamic potential) and hence cannot be obtained simply at high temperature and pressure. It could be measured by using Brillouin spectroscopy.

We are postulating two different possibilities. First, that the Earth, especially from 60-1400km, contains lateral variations in mineralogy and that at high pressure the variation of shear velocities compared to compressional velocities of different minerals will be much larger. A second suggestion is that compressional velocities will show less variation with temperature at high pressure compared to the changes in shear velocity. Either one or the other effect could be sufficient, but extrapolating low pressure laboratory results for an homogeneous mantle will not work.

For the lowermost mantle the S-wave half-width result is not very robust with respect to the damping parameters, and the favoured half-width has large errors. If we make the assumption that the P and S half-widths are identical at this depth, this leads to the result that the ratio is around 1.5 $\pm^{\infty}_{1.5}$ here. If we take the shear half width to be 2400km rather than 1200km, then the ratio is reduced to $\approx$1.1. *Giardini et al.* [1988], obtained a value of 2.5-3 for this ratio. They obtained this ratio by evaluating the best ratio to convert a compressional velocity model to a shear velocity model to give the best fit for the splitting of free oscillations (which are more sensitive to shear rather than compressional velocities). We notice that they would obtain only a marginally poorer fit at a ratio of 4, while values as low as 1 produce much poorer fits. They mentioned the possibility that their obtaining a value of 2.5 for this ratio could be the result of comparing data at different periods, short-period 1Hz body waves

and long-period free oscillations. Our data are both at the same period and they also give a high ratio (but with large error bounds), hence a high ratio is unlikely to be the result of different seismic behavior at different period ranges. It is difficult to conceive of a mechanism to generate large-scale mineralogical heterogeneities (it is possible, though unlikely that they are left-over from accretion, due to the vigorous convection expected [*Stevenson*, 1989]). The claimed correlation of the S- and P-wave velocities [*Dziewonski and Woodhouse*, 1987] would also tend to argue against a mineralogical effect. This study though can make no statement on the correlation of the variations since we have thrown away all positional information.

How do we explain the higher ratio found in this study than in previous studies? The comparison of statics is an integration of lateral variation down to an ill-defined depth. The body-wave statics are related to the average structure below a station down to a depth defined by an average of all the various receiver angles and the local Fresnel zone. Hence, this will have its largest contribution from the small-scale shallow power. The ratio for this depth is not well defined by our study due to the poor depth resolution at this depth. It is possibly smaller and the near surface layers possibly conceal the larger ratio deeper in the mantle. Since there is no global P-wave upper mantle study, there has been no comparison at the depth postulated. The high resolution studies have found S-wave variations of up to $\approx 11\%$, and these suggest mineralogical/compositional variations beacause if these S-wave variations were only due to temperature, this would signify variations of 1000K. Variations of this magnitude would lead to larger amounts of

melting and larger heat flow anomalies than are observed. The lower mantle studies have developed this comparison by using body-wave studies to produce a shear model with which to match free oscillation splitting. These have also required estimation of the correlation between density and the compressional velocity; the surprising ratio of compressional to shear-wave variations found in this study suggests that low pressure measurements might not be suitable for estimating these correlations at depth in the Earth. The minima of 2.5 found by *Giardini et al.* [1988] was asymmetric with the fit degrading rapidly as the ratio was decreased, but largely insensitive as the ratio was increased. Given this fact and the other uncertainties, we believe that the lower mantle ratios are not incompatible. Also, the splitting functions involve global averaging over large depths of the mantle, and the higher ratios might be hidden by potentially lower values of the ratio in the lower half of the mantle. Due to the large errors in the lower mantle, the ratio from this study is not very robust; and a value of 2 is not unreasonable. The value at the top of the lower mantle is larger than 4 but the lower mantle P-wave studies suffer from leakage from the upper mantle. S-wave studies of the lower mantle have to date used surface wave overtones [*Tanimoto*, 1989a], they have assumed a value for this ratio. Given a different value for the ratio the results might change and hence the ratio and the assumed ratio could be made self-consistent.

A further possibility to explain this high ratio is that the assumptions of the method don't hold and that their failure conspires to increase the ratio. One distinct possibility is the non-ideal sampling. Even given the correction for the clustering of events and sta-

tions, this is a simple distribution that is derived from a sub-set of the data and is assumed to hold for the real, whole, complex data-set. In general we believe that our correction is very good but we have an indication that it might have problems at the largest separations at the largest scale-length. There is the further problem that this distribution might not be the same for all depth bins. Given the restricted seismicity at depth, one would expect even more clustering and that this effect would lead to an underestimate of scale-length at depth. A more critical problem is the interdependence of sampling and structure. This is a larger problem at depths of 100-700km where our model might be more representative of the variations near slabs in subduction zones than of the mantle in general. We know that there are compositional variations, partial melting, displaced phase transitions and large temperature contrasts near subduction zones. There is ample opportunity for complications in the interpretation of this ratio if our sampling heavily biases our model towards subduction zones. Given the poor vertical resolution in the upper mantle, we could also be suffering from leakage of shallow crustal signal. All these effects complicate whether the interpretation can be extended to the whole mantle, but none of the above suggest that a ratio greater than 1 isn't found somewhere in the Earth, be it solely in the crust or subduction zones.

Comparisons of the shear and compressional velocities have been used previously to get additional information in crustal studies. *Walck* [1988] in a back-projection tomographic study discovered areas of very low shear velocity, which she ascribed to geothermal systems and the presence of partial melt. *Holbrook et al.*

[1988] used refraction data to make a petrological model, by comparing their values to laboratory derived values. Our comparison is of fluctuations away from a mean, rather than the actual velocities. We need to be confident in our spherical Earth model before we proceed with mantle petrological models, but given the reasonably symmetric distribution of residuals around zero for the teleseismic range, we can be relatively confident of the magnitude of this ratio.

This ratio is a useful *in situ* datum for deep Earth mineralogy, and should be investigated and confirmed. Developing an understanding of this ratio should lead to a better understanding of the elastic properties of materials at high pressures and temperatures in general, a field presently lacking in direct laboratory measurement and also receiving little theoretical attention.

Many of the global seismological models have depended on assuming a correlation between the shear and compressional wave velocities. A very high ratio such as 5 would bring into question whether there is any correlation, but if there is one, then such a figure is higher than previously assumed, (e.g.,, *Masters et al.* [1982], 1.25; *Woodhouse and Dziewonski* ,[1984], 1.25; *Ritzwoller et al.* [1986], 1.25;*Tanimoto* [1989], 1.3-2.0; *Giardini et al.* [1988], 2.5-3.0). This ratio might also vary with depth. The kernels in these studies are primarily sensitive to the S-wave velocity. It is unclear how sensitive the results are to the choice of this ratio. The low variance reduction might be an indication of a poor model suggestive of a poor choice for this constant. These free-oscillation studies have also had difficulty explaining some core modes, could possibly admitting that there is no correlation allow them to be fit? These free-oscillation

and long-period surface wave studies generally also assume a correlation between the velocities and the densities.

As mentioned above, the possible high value of the above ratio suggests that assuming any correlation might be poor, due to the hint of the influence of mineralogical and compositional effects. This would also bring into doubt all studies that convert seismic heterogeneities into buoyancy forces in the Earth's mantle and drive flow, to predict surface observables such as the long-wavelength geoid, and the poloidal component of the surface plate velocities [*Hager et al.*, 1985; *Peltier et al.*, 1989; *Ricard et al.*, 1989]. A high ratio does not require a lack of correlation between density and seismic velocity and is only suggestive. As discussed above, for the very longest wavelength heterogeneity it is difficult to consider long-term sources of large-scale heterogeneities, but the magnitude of the smaller scale heterogeneities might be so large that there might be the possibility of "dynamical aliasing" (sufficient small-scale forcing such that the resulting flow has a large-scale). All the subduction circum-Pacific suggests that such a large-scale component isn't out of the question. This result has potentially wide ranging implications and it deserves closer scrutiny from all directions.

1.6.d Ratios of half-widths

The half-widths in both studies follow a similar pattern. They are constant through the upper half of the mantle; while they potentially increase in the lower half of the mantle. Given the poor spectral resolution, the exact values are questionable, especially in

the lower half of the mantle where the model power is so low and at the shallowest depth with the poorest depth resolution. The half-width values for the P-wave studies varies from ≈450-550km in the upper mantle through to 1200km in the lower mantle, decreasing again to < 400km at the CMB. The S-wave studies similarly show half-widths of ≈ 350-600km in the upper half of the mantle, increasing to order 3,000km in the lowermost mantle. As previously described, the S-wave lower mantle value is not robust to variations in the value of the damping parameter. Slight variations in damping parameter can lead to this half-width value approaching 400km.

Hence, within the errors and robustness of the values a ratio of unity for the P and S half-widths throughout the mantle is perfectly feasible. The details of the variations are different between the two studies and it is equally possible that the ratios are really different. If they are different, then the two fields cannot be correlated. This will have to await a study that can achieve better spectral resolution. We should note that these half-widths are derived only from the long wavelength component of the power, hence the real half-width is much smaller. This suggests that the characteristic scale of seismic velocity variations through the mantle is very small; i.e., the spectrum is nearly white. This is not unexpected considering the thin structures characteristic of high Rayleigh number convection.

1.6.e Implications for deterministic seismic inversions

These results obviously have implications for deterministic inversions. First, for inversions using the ISC P- and S-wave

data-sets, our intercepts provide estimates of the intrinsic noise; this technique also can be extended to measure the structural noise by estimating the variance that is present at the scale-length of the inversion. This information can be used to help in deciding the degree to which to damp the solution. Equally, the variance as a function of scale can be analyzed at various stages in an inversion, and the reduction of the variance at different scales can be observed. This gives information on which scale features are being rejected by the inversion and which features are accepted at each step. This methodology has been applied by *Zhou* [1988] to investigate the properties of inversions of subduction in the North-West Pacific. Noise and signal estimates as a function of scale can be used to select appropriate grid-sizes, and also to give the data error covariance matrix from which estimates can be made of the model error covariance.

The results of our inversions give estimates of the spectra as a function of depth. This provides information with which one can design the best model parameterization for the objective in mind. Equally, one can then evaluate the problem due to aliasing, arising from structures smaller than the minimum scale-length of the inversion. Also, one can estimate the problems due to contamination arising from poor resolution of features leading to power being mislocated.

Since both the P and S models show much larger power in the upper mantle than in the lower mantle, this suggests that unless the upper mantle structure is well resolved, slight leakage into the lower mantle would lead to appreciable contamination. This effect has been investigated using synthetic models and representative ray

geometries [*Gudmundsson*, 1989]. It was found to be a large problem for lower mantle structures and potentially an insurmountable problem for body-wave analyses of the core-mantle boundary and the inner core. Using "waves" with large wave-lengths (e.g., surface-waves, and long-period free oscillations) can alleviate these problems since they intrinsically average the structure at larger scales; but these generally have poor vertical resolution and hence can still lead to contamination. Note the potential need of crustal corrections for surface wave models [*Nataf et al.*, 1986; *Woodhouse and Dziewonski*, 1984].

1.6.f Temperature variations?

It would be interesting to assume that these large scale-length variations were due to temperature and to evaluate the variations implied by these slowness variations. If we assume that the upper mantle is primarily olivine then estimates of $\delta V_p/\delta T$ are $\approx -0.5\ 10^{-3}\ kms^{-1}K^{-1}$ from laboratory experiments [*Sumino and Anderson*, 1984]. These measurements were obtained below the Debye temperature. *Isaak et al.* [1989a] suggest that for forsterite it decreases from $-0.49\ 10^{-3}\ kms^{-1}K^{-1}$ at 300K to $-0.69\ 10^{-3}\ kms^{-1}K^{-1}$ at 1700K. All these measurements are at 1 atm. pressure. Theoretical calculations suggest that these temperature derivatives increase with pressure, (i.e., decrease in absolute magnitude), such that temperature variations have a smaller effect on velocities [*Isaak et al.*, 1989c]. The net result is that the real temperature derivative might

not be very different from the room temperature, atmospheric pressure measurement.

Creager and Jordan [1986] inferred a temperature derivative of -0.4 10^{-3} $kms^{-1}K^{-1}$ to match predictions of residual sphere observations using a thermal model of a subduction zone. Their derivative assumes that the signal they see is caused solely by temperature, and not due to temperature sensitive phase changes or compositional variations. This derivative suggests that long-wavelength temperature variations of mean amplitude 70K are required to explain the 0.4% P-wave slowness variations.

By contrast the laboratory estimate of $\delta V_S/\delta T$ is around -0.3 10^{-3} $kms^{-1}K^{-1}$. Hence the shear wave upper mantle variations of 2% suggest temperature variations of mean amplitude 350K. This is the same difference as encountered in the $\delta \ln(V_p)/\delta \ln(V_s)$ discussion above. The explanation put forward to explain the difference in the upper mantle is the presence of partial melting, which affects the shear wave velocity much more than the compressional wave velocity. Notice that estimates of small-scale structure (l >50) of 5% for the compressional waves would lead to mean temperature variations of ≈ 1000K! , at this scale.

Unfortunately there are no similar experiments on the velocity temperature derivatives of perovskite, the primary lower mantle mineral [*Jeanloz*, 1989]. Assuming that they are not too different from magnesio-wustite (the other significant postulated lower mantle mineral), for which *Isaak et al.* [1989b] found $\delta V_p/\delta T$ = -0.6 10^{-3} $kms^{-1}K^{-1}$ and $\delta V_S/\delta T$ = -0.45 10^{-3} $kms^{-1}K^{-1}$ at 1800K , we find that the lower mantle temperature variations are ≈ 14-70K (lower

value for P-wave, higher for S-wave). A similar figure (15K) is arrived at by *Hager and Clayton* [1989] from compressional body-wave tomography for the very long wave-length temperature variations ($l <= 3$), in their modeling of the Earth's long wave-length geoid. Since our figure is for $l < 50$, the Hager and Clayton estimate is larger. We see that this suggests that the lower mantle large scale-length temperature variations are appreciably less than the variations in the upper-mantle.

If the above assumptions are correct, (remember both the contribution of composition and the value of the velocity temperature derivatives are uncertain) it would suggest that there is a change in the thermal fluctuations in going from the lower to the upper mantle. These might signify a change in the mode of convection and possible partial isolation of the two layers, this needn't be the result of a chemical barrier but could be the result of an increase in viscosity. There is some support for this also in the structure in both stochastic images of mean slowness across the 670km discontinuity. The layer above the 670km discontinuity is poorly resolved, but since it trades off with layers that are shallower, (especially the shallowest layer, which we expect to be more heterogeneous) we suspect that the model value is an overestimate. If there was perfect unimpeded convection across an homogeneous region with no phase change at this depth, then one would expect a smooth gradation in velocity perturbations. Unfortunately there is also a major change in the co-ordination of silicon (6- versus 4-fold) and the mineral phases change from γ-spinel, pyroxene and garnet to perovskite and magnesiowustite, at this depth. We presently cannot rule out the fact that

the change in seismic behavior could result from the lower mantle phases having appreciably different temperature derivatives compared to the upper mantle phases.

The assumption that all the variations are due to temperature is questionable. *Tanimoto* [1989b], has shown that continents averaged over the upper 200km are less dense than the surrounding mantle. Hence their faster than average velocity cannot be due solely to cooler temperatures of mantle material. Therefore part of the velocity contrast must be explained by compositional differences. Unfortunately the surface wave density kernels are very small deeper in the mantle and hence the method cannot be extended deeper into the mantle. From the observations of both eclogitic and peridotitic xenoliths from the upper mantle it is clear that the upper mantle isn't compositionally homogeneous. It is clear that subduction also introduces compositional anomalies into the mantle. Their fate is unclear. Do they get passively stretched and destroyed [*Allegre and Turcotte*, 1986; *Gurnis and Davies*, 1986; *Hoffman and McKenzie*, 1985; *Olson et al.*, 1984a; *Olson et al.*, 1984b], or do they collect like megaliths at the 670km discontinuity [*Ringwood*, 1982; *Ringwood*, 1985] or at the CMB [*White and Hofmann*, 1982]? Whatever their fate, it is clear that they survive for an appreciable time. I shall argue that since the source of present day heterogeneities is subduction; and that they have a relatively small scale-length (perpendicular to slab ≈ 35km, along strike, up to a few thousand kilometers), the very longest anomalies are due to temperature. As one considers shorter length-scales, so the effects of composition will increase. I suspect that the compositional effects are sufficient such

that one should not expect perfect correlation of shear and compressional velocities even if they were perfectly resolved.

A further implication of the magnitude of the temperature variations stems from the requirement that heat generated inside the Earth needs to be transported to the surface. Net heat flow across a layer is a function of the cross-correlation of the temperature and velocity fluctuations. If the heat was transported solely by the large-scale (20K) temperature fluctuations and the velocity was perfectly correlated with temperature, one would require velocities of the order of 1 cm/yr to carry 40mWm^{-2} through the lower mantle. (This value [40mWm^{-2}] corresponds to 15% of the surface heat flow, [remember, the area is four times smaller at the CMB than at the surface] which is close to the minimum estimates of the heat flow from the core and neglects any internal heating in the lower mantle.) Given estimates of lower mantle viscosity of 10^{22} Pas, coefficient of thermal expansion of 4.0 10^{-5}K^{-1}, and 20K temperature variations of 1200km scale-length, Stokes Law suggests velocities of $\approx$ 3.6 cm/yr. If we consider shorter scale-length variations, then since the velocity is proportional to the square of the scale–length, one finds lower velocities.

If the heat flux through the lower mantle were greater, the coefficient of thermal expansion smaller [*Anderson*, 1987a; *Chopelas and Boehler*, 1989] or the viscosity higher [*Hager*, 1984; *Hager and Clayton*, 1989; *Sammis et al.*, 1977] then we would have lower velocities. This could lead to a possible contradiction. The velocities required to carry the heat flow could not be produced by the temperature variations postulated; similarly a contradiction is pro-

duced if the temperature variations and velocities are largely uncorrelated. A solution in this case might be that the broad scale temperature variations do not carry all the heat-flow but rather it is concentrated in thin unresolved narrow plumes. These plumes have much larger temperature increases and higher velocities due to lower viscosities. The cooling of the mantle could also be due to very narrow downwellings, with large temperature differences. If the plumes have to be solely responsible for the heat flow, then one can constrain the structures of plumes because they must transport appreciable heat but their thermal signature must be restricted since they are not observed in our seismic spectra. If the above arguments are correct there is presently little need to postulate plumes to transport the lower mantle heat flux.

Applying the same logic to an upper mantle of viscosity of 10^{21} Pas we estimate a velocity of 0.15cm/yr, for heterogeneity of radius 500km. Hence temperature fluctuations of order 350K from S-waves gives a heat flux of $100mWm^{-2}$, while temperature fluctuations of order 70K from P-waves gives a heat flux of only 20 mWm^{-2}. If the Earth's heat flow was radially in steady-state and the correlation of velocity and temperature was perfect then one could use the seismic variations to predict the temperature variations. Hence from the heat flux we could predict the velocity. The scale-length is constrained by the scale-length of the seismic variations hence the velocity can only be achieved at a certain viscosity. This could be used to evaluate the mantle's viscosity stratification. Unfortunately the correlation is unlikely to hold since the thermal dependence of viscosity couples together different modes and leads to a cascade

from one to the next. In convection of a fluid with temperature-dependent viscosity, the high velocity plumes are more concentrated than their thermal halos, i.e., the thermal signature seems of longer wavelength than the velocity signature. Also one would need to be able to evaluate the contribution of composition to the seismic heterogeneity and the distribution of internal heat sources.

1.6.g Comparisons to convection models

What seismic variations might one expect from a convecting mantle? *Jarvis* [1985], and *Jarvis and Peltier* [1986; 1989] have considered this for the case of a steady, incompressible, constant viscosity, 2D unit aspect ratio convection. They have found that one would expect the boundary layers to have very red spectra with the spectra becoming whiter in the middle of cells. Also the magnitude of the variations should be largest at the boundaries. Introducing internal heating introduces power into the even fourier components, where previously, due to the symmetry of the problem, there had only been power in the odd fourier components. Also the spectrum loses its peak at the lowest fourier component with the peak moving to a slightly higher components. At higher Rayleigh numbers they found that the convection was more vigorous and that the boundary layers are much thinner. Hence the spectra become whiter and stronger.

At the base of the mantle and also at the top of the mantle we have an increase in heterogeneity but a slight decrease in scale-length. The greater heterogeneity would be expected with

boundary layers but we would also expect an increase rather than a decrease in scale-length. Our inability to discern dramatic changes in characteristic length scales could be a function of our poor spectral resolution ($l < 50$), poor vertical resolution, or the fact that these boundary layers are also sites of compositional variations due to their positions at the extremes of the mantle. Scum, (continents) floats to the top and dregs collect at the bottom?!

The evidence for a bottom boundary layer in our data is weak. *Gudmundsson* [1989], showed in a synthetic study that due to poor coverage a similar feature can be reproduced in a stochastic inversion if the data is contaminated with realistic amounts of noise. Spherical geometry increases the strength of the bottom boundary layer since it is the same thickness but has a quarter of the area, hence it must carry a much higher temperature contrast to carry the same heat flux. It has been found that when cells are internally heated the bottom boundary layer is decreased, disappearing if heated completely from within.

In 3D compressible convection it has been discovered that the planform consists primarily of upwelling plumes and downwelling sheets; as the degree of internal heating is increased the number of plumes increases and the downwelling sheets become more vigorous [*Bercovici et al.*, 1989]. *Houseman* [1988] in 3D cartesian geometry has also found that at higher Rayleigh numbers that the pattern of convection changes as the mode of heating is changed from internal to basal heating. With all the heating being internal the model develops slab-like downwellings surrounded by a pervasive warm upflow. In contrast with all bottom heating we have plumes

rising through a pervasive downflow. With primarily internal heating but also some basal heating ≈ 20%; he found more plumes and also slabs. In fact the structures at the base of the box start as linear features but at their junction they coalesce into plumes as they rise. These ridge-like structures might explain the two characteristic scale-lengths of heterogeneity postulated at the base of the CMB from scattering studies. The slabs also break up into fingers as they proceed down, an observation claimed for subducting plates [*Zhou*, 1989]. All this suggests that the degree of bottom heating is low, otherwise one would expect a more dramatic seismic signature. Hence we suggest that most of the Earth's heat flow is due to internal heating and secular cooling.

A critical question in discussing predictions of mantle convection is deciding the extent and form of the convection. This depends critically on the behavior of slabs in and around the 670km discontinuity. Some people have argued for layering from the observation that seismicity ceases below the 670km discontinuity, the fact that deep focal mechanisms are compressive and since Rb-Sr, Nd-Sm and inert gas isotopic studies can all be satisfied with layered models. Others have argued for whole mantle convection since kinematic models incorporating whole mantle convection could explain slab dips [*Hager and O'Connell*, 1978], convection simulation aspect ratios could be more easily reconciled with the large length scale of the tectonic plates, and residual sphere studies [*Creager and Jordan*, 1984; 1986] seem to require penetration. While even others have argued for leaky convection, primarily as a means to satisfy both families of evidence [*Silver et al.*, 1988].

Many of these arguments are linked to explaining geochemical and isotopic data. It is becoming clear that the isotopic evidence might be largely independent of major-element chemistry. It isn't clear what the host phase is for many of the REE, alkalies, LIL, including the parents and daughters of many isotopic schemes. Possibly they shall be found primarily at grain boundaries rather than any crystal phase; in this case their relationship to the major elements in the rocks will be weak. The constraint that there is an isotopic reservoir that has been independent of the convective process can be readily accounted for by involving a pristine lower mantle and also by recycling components that have been stabilized in the sub-continental lithosphere.

Equally many of the other observations mentioned above are uncertain, controversial, or are incompatible with later work. There is much uncertainty regarding the observations of *Creager and Jordan* , [1984; 1986] who claim to see the signature of the aseismic extension of slabs below the 670km discontinuity in the pattern of travel-time residuals on the focal sphere. *Zhou et al.* [1988] and *Grand and Ding* [1989] claim that given different corrections for the data one does not get the pattern. *Hager and O'Connell* [1978], in the study where they matched the dips of slabs by driving the flow inside the Earth using the surface plate velocities as boundary conditions, discovered that the dips fit best if there was no viscosity contrast across the 670km discontinuity. *Hager et al.*, [1985] argued from the modeling of the geoid using the tomographic results as input that there is a viscosity jump between the upper and lower mantle of approximately 30. They also argued against a chemical in-

terface. A suggestion to resolve this has been that the slabs in the dip study were too thick since there was no variable viscosity to make them thin and rigid. *Peltier et al.* [1989] have argued from using the tomographically obtained velocity contrasts converted into buoyancy forces that they can match the surface velocity fields if the viscosity contrast at the 670km is ≈ 8.

What would be the seismic signature of these three different scenarios in our results? One of the uncertainties is the lack of information on the temperature derivatives of the seismic velocities of many of the phases present in the mantle. We shall assume for discussion's sake that they do not vary greatly from the derivatives already measured. We shall also assume that there are no significant lateral compositional differences.

Given the above assumption a two layer mantle would lead to a thermal boundary layer at this depth and slabs would also cause appreciable deflection of the chemical boundary. Therefore one would expect increased seismic heterogeneity above and below the discontinuity. The thermal boundary layer would lead to longer scale-lengths but the deformation component will contribute a scale-length characteristic of the convection in the upper and lower layers. One would also expect a different scale-length for the upper and lower mantle given their different depths and chemistry. In whole mantle convection there would be no thermal boundary layer but there would be deflection of the phase boundary at the 670km discontinuity. If the downwelling slabs are more focussed than the upwelling plumes at this depth, then it is possible that the deflection of the phase boundary is largely restricted to the subduction zones. A

leaky system would exhibit intermediate characteristics. As envisioned by *Silver et al.* [1988], it would have a weak thermal boundary layer and weaker deflection of the chemical interface. This is because the primary attacks on the interface by slabs and plumes would penetrate it, and their penetration would lead to increased heat transport.

Our results, under the above assumptions, would tend to favour whole mantle convection, but given the lack of mineralogical data, the poor spectral and depth resolution of the stochastic method and the lack of understanding of the effect and influence of compositional variations, this must remain a tentative conclusion. It will be interesting to extend our analysis to regional distances to improve our depth resolution in the neighborhood of this discontinuity.

Glatzmaier [1988] presents one of the more complex simulations of mantle convection which also evaluates spectra of the thermal variations. He considers compressible 3D convection, with large Rayleigh numbers. He found large variations in behavior as he increased the Rayleigh number. At $Ra=10^5$ he found a network of narrow cold downwellings and broad regions of upwellings but at $Ra=10^7$ this was replaced by hot thin plumes in a broad region of cold downwelling. His model produced maximum temperature deviations of around 200K at the largest Rayleigh number. His model went up to harmonic degree 65. At $Ra=10^5$ he found that the thermal variance decreased by nearly an order of magnitude from its peak at degrees 3-10 out to $l = 50$. While at $Ra=10^7$ he found that by harmonic degree $l=50$ the thermal variance had only decreased by a

factor of 2. This is similar to the findings in much simpler 2D models of convection by *Jarvis and Peltier* [1986].

We are sensitive to not only the degree of lateral heterogeneity but also to its scale-length as a function of depth through the Earth. Our half-widths suggest length scales shorter than 500km, since this is the scale-length with the short-scale power excluded! *Weinstein et al.* [1989] have found that large aspect ratio cells can exist, but like others they find that at larger Rayleigh numbers the boundary layer instabilities become more vigorous and give the flow a "multi-scale character," i.e., a whiter spectra. Introducing plates above the flow allows the flow to attain much larger aspect ratio [*Davies*, 1988] and hence would tend to make the spectra "redder."

Davies [1984] has suggested that much of the geochemical evidence can be explained by distributed heterogeneity throughout the mantle generated by the plate tectonic process. It is conceivable that different portions of the spectra of seismic heterogeneity are controlled by different factors. Hence the longest scale-length features might be controlled by temperature, the very shortest scale-length features controlled by composition while the intermediate scale-length features are controlled by both and also phase changes. The very small half-width tends to favour high Rayleigh number convection and internal heating, but it could also be suggestive of distributed heterogeneity. The question of whether these heterogeneities can be considered as passive or whether their spectra actually leads to variations that are sufficiently large spatially to alter the flow has yet to be firmly resolved. A better description of the spectra of seismic heterogeneity combined with modeling of thermal

convection might allow one to assess the contribution of non-thermal heterogeneity to the seismic spectra and hence allow one to estimate its influence on mantle dynamics.

The lack of short-scale power in the lower mantle can be explained by the primary convective features being much smaller than averaging ray width ≈ 100km, or that they are sampled by insufficient rays or that they don't exist and only the large-scale features exist. The first two points are very reasonable, since we would expect very narrow features in a high Rayleigh number convective system with temperature dependent viscosity. The third point is also feasible from a heat flux scenario as discussed above. This would imply very weak convection, i.e., a low Rayleigh number. This could be a function of increasing viscosity [*Sammis et al.*, 1977] and a decreasing coefficient of expansion, [*Anderson*, 1987a; *Chopelas and Boehler*, 1989]. If so, then this weakly convecting region could allow heterogeneities to survive a very long time before being sampled at the surface to satisfy isotopic data without requiring separate convection. Of course if the viscosity stratification is too large then separate convection can develop. *Davies* [1977] estimated that separation requires a viscosity increase of three orders of magnitude.

The most prominent feature in our models is the decrease in power below 300-400km. This can be attributed to the difference between continents and oceans as a function of composition, phase and temperature. Being such prominent features suggests that they are important features of the system, possibly even dynamically. The continents might introduce a blanketing effect as demonstrated by *Gurnis*, [1988], leading to dynamical feedback.

The lack of periodicity in the data suggests that we don't have steady state convection; hence this puts a lower bound on the Rayleigh number. We note that compressible convection leads to thermal turbulence, "chaos," setting in at much lower Rayleigh numbers than is the case for Boussinesq convection.

All the above simulations are obviously of a reconnaissance nature. Many other facets of mantle convection have been investigated. For example it has been found that for constant viscosity rheology, compressibility damps the flow, while with temperature dependent viscosity the effects are more complicated and it can lead to enhanced downwellings being decoupled from the majority of the flow due to flanking low viscosity regions [*Schmeling*, 1989]. Stress dependent rheology has been found to smooth temperature dependent flows [*Christensen*, 1984]. At high Rayleigh numbers the flow field for stress dependent but temperature independent viscosity becomes much more vigorous. There is less of an effect on the thermal field [*Christensen and Yuen*, 1989]. The limitations of the above simulations serve to remind us that the Earth is more complex. Moreover all the above factors affect the spectrum of temperature variations differently. Hence we should continue to view our interpretations with a healthy dose of skepticism.

1.6.h Spectral comparisons of the shear wave study results

The spectra are shown in figure 1.23. The spectra from 60-1400km are reasonably similar, while the deepest depth bin has

a spectra that decreases more rapidly as a function of harmonic degree. The P-wave spectra are similar to the S-wave spectra, but the numbers in the following discussion are derived solely from the S-wave study.

The seismic spectra could be used to predict the shape of the geoid spectrum using the kernels of *Richards and Hager* [1984]. They claim that the lower mantle signature is responsible primarily for the signal at degrees 2 and 3. While the slabs are largely responsible for the signature at degrees 4 to 9, a strong correlation has been found between degree 6 of the spectrum and the hot-spot distribution. Here we have variations through the whole mantle and with spectra that extend to a very high degree. The poorest resolution, near surface, and the largest errors, near CMB, introduce the least to the geoid since they are near rigid boundaries.

Unfortunately, our spectra suggest that the spectra of temperature variations would involve variations at all scale-lengths. This would suggest large lateral variations in viscosity. The methodology of *Richards and Hager* [1984] assumes a spherically symmetric viscosity. *Ashby and Verrall* [1977] have developed a theory to permit the extrapolation of laboratory measurements to mantle conditions. Their work suggests that a temperature decrease of 100K can increase viscosity by an order of magnitude. This is a critical link in the following argument and direct experimental confirmation of the above result would be important. *Richards and Hager* [1989] have investigated their spherically symmetric viscosity assumption and suggest their method is robust provided that the variations in viscosity are no greater than an order of magnitude. Our results sug-

gest that the long-wavelength temperature variations in the lower mantle are less than 80K (20K from P-wave study). Hence if our spectra (long-wavelength variations of < 80K) and their interpretation is confirmed it would suggest that the spherically symmetric viscosity assumption of their method is valid. This assumption has also been used by Forte and Peltier in their prediction of the poloidal field from tomographically imaged heterogeneities (buoyancy forces), and by *Ricard et al.* 1989, who have used it to predict both the geoid and surface velocities.

Assuming laterally constant viscosity uncouples the different harmonic degrees. Coupling is significant and is demonstrated by the fact that the power in the toroidal component of the plate velocity field is nearly identical to the power in the poloidal velocity field. Density variations in a mantle with laterally constant viscosity cannot generate a toroidal field. From surface tectonics it is known that the interiors of plates generally undergo very little deformation while plate boundaries undergo extreme deformation (are actually discontinuities). Hence it is at the surface of the Earth that we most probably should expect the largest variations in viscosity, and where the coupling of poloidal to toroidal field is strongest. Contributions to the geoid from masses near the surface are very small, and most probably remain small even given the large lateral variations in viscosity.

The spectra of other surface fields have also been evaluated. They vary reasonably smoothly as power laws of harmonic degree with constant exponents. Heat flow has been characterized to decrease $l^{-0.5}$; this is much flatter than the geoid that is $l^{-2.0}$ or the

velocity fields that are $l^{-2.3}$ or the topography that varies like $l^{-1.5}$ [*Kaula*, 1980]. These are defined as the root mean square magnitude of the normalized coefficients; i.e., the equivalent spectra for the seismic heterogeneities would be $\sqrt{(Q_l)}$ as a function of l . From the above it is clear that much of the heat flow is due to small structures, while the topography is sensitive to larger structures while the geoid and plate velocities are sensitive to only the very largest structures.

In figure 1.30 we present the logarithm of the spectra versus the logarithm of the harmonic degree. Since the lines are curved rather than straight this suggests that the spectra of seismic heterogeneities do not follow a power law but would be better described by a power law where the exponent was a function of scale, with the exponent becoming more negative as the harmonic degree increased. Out to around harmonic degree thirty the spectra can be approximately described by a power law with a constant exponent as follows $l^{-0.65}$, shallow and $l^{-1.1}$ in the lower mantle. Both numbers are not very robust and have large (≈50%) error bounds.

Glatzmaier [1988] in his simulations finds that the thermal variance out to degree 50 is an approximately linear function of the logarithm of the harmonic degree. In figure 1.31 we present the normalized spectra versus the logarithm of harmonic degree, and we find a similar behavior for our results. If this correlation can be confirmed as being real then it would give us a powerful tool with which to constrain mantle convection and hence such poorly constrained quantities as viscosity. Hence if *Glatzmaier* [1988] plotted the logarithm of the thermal variance versus the logarithm of harmonic degree he would equally find a power law that is a function of scale. At

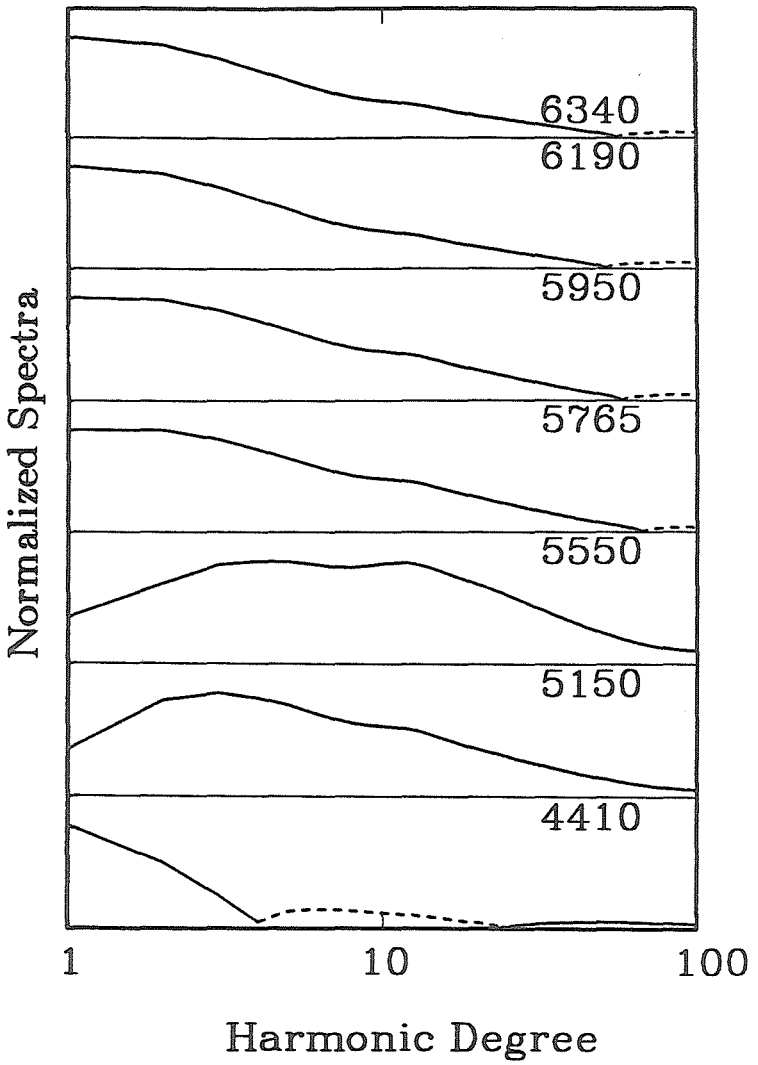


Figure 1.30 Plot of Normalized Spectra of S-wave model versus harmonic degree on a logarithmic axis. Dashed line represents negative power.

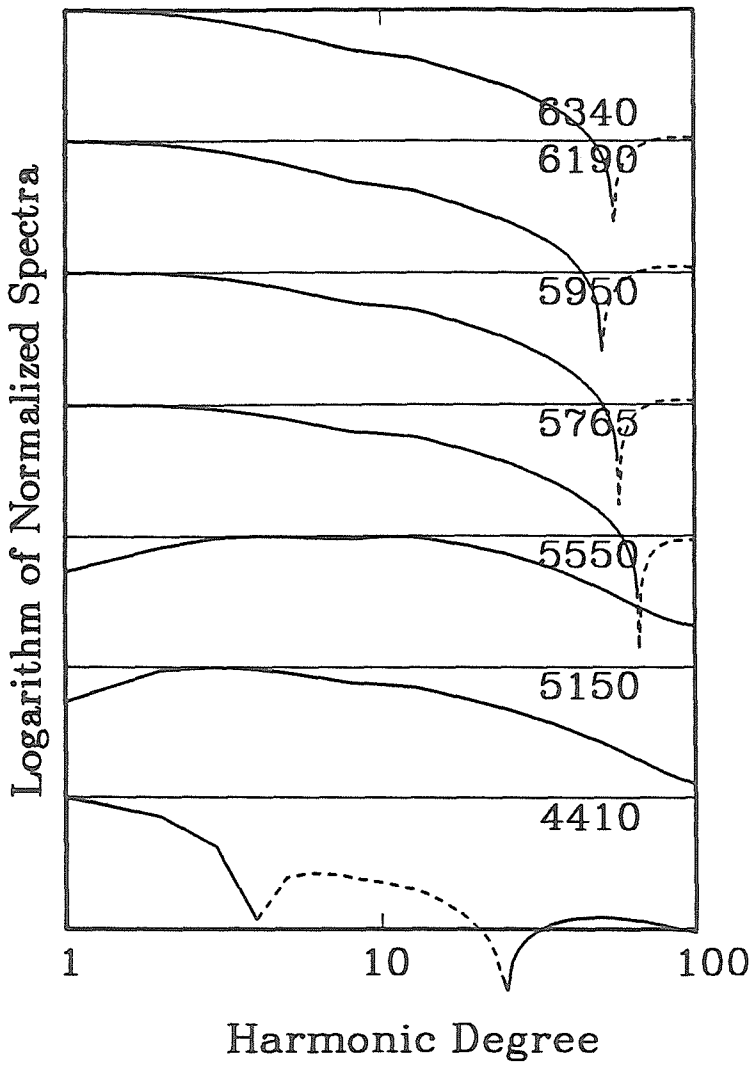


Figure 1.31 Plot of Logarithm of normalized spectra of S-wave model versus logarithm of harmonic degree. Dashed line represents negative power.

a Rayleigh number of 10^6 he finds that the thermal variance decreases by an order of magnitude out to degree 50; hence using the end points this would define an approximate power law with an exponent $\approx$ -0.6, while with a Rayleigh number of 10^7 he only finds a decrease of a half over the 50 degrees, (i.e., an exponent of $\approx$ -1.4). If these approximate power law exponents are similar to the estimates derived from figure 1.30, and both our results and Glatzmaier's simulation were perfect, then this would suggest that mantle convection could be characterized by a Rayleigh number of $\approx 10^6$. He finds that the velocity variance decreases faster as a function of harmonic degree and is a weaker function of the Rayleigh number.

Comparing the seismic spectra to the thermal variance of convection calculations might be an indirect method to estimate the Rayleigh number and hence the convective regime of the Earth. This would be feasible provided that the thermal variance of convection was discovered to be a strong function of Rayleigh number but a weak function of other potential complications (e.g., temperature dependent viscosity, Glatzmaier's simulations are for constant viscosity). Of course we would assume that the low harmonic degree seismic spectra solely reflected temperature variations.

In figures 32(a), 33(a), and 34(a) we present for completeness the equivalent of figure 31 (the logarithm of the normalized spectra versus logarithm of harmonic degree) for the other three cases considered (the alternative damping, the logarithmic intercept estimates and the raised asymptotes). Equally, figures 32(b), 33(b) and 34(b) are the equivalent of figure 30. The consistency of the spectral shape for the four cases from 60-1400km gives us confi

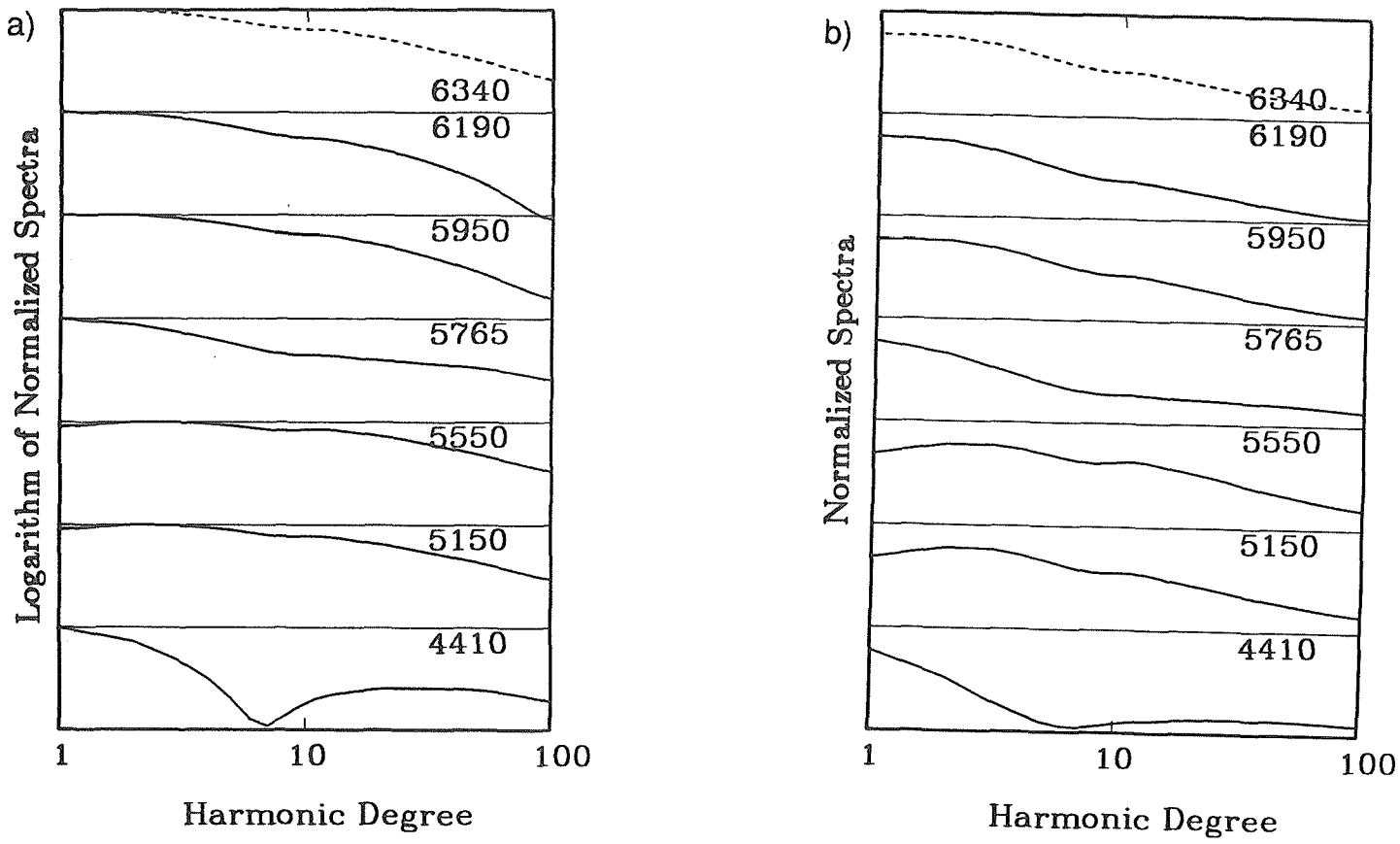


Figure 1.32 (a) Plot of Normalized Spectra of S-wave model with alternative damping parameters versus harmonic degree on a logarithmic axis. (b) Plot of Logarithm of normalized spectra of S-wave model with alternative damping parameters versus logarithm of harmonic degree.

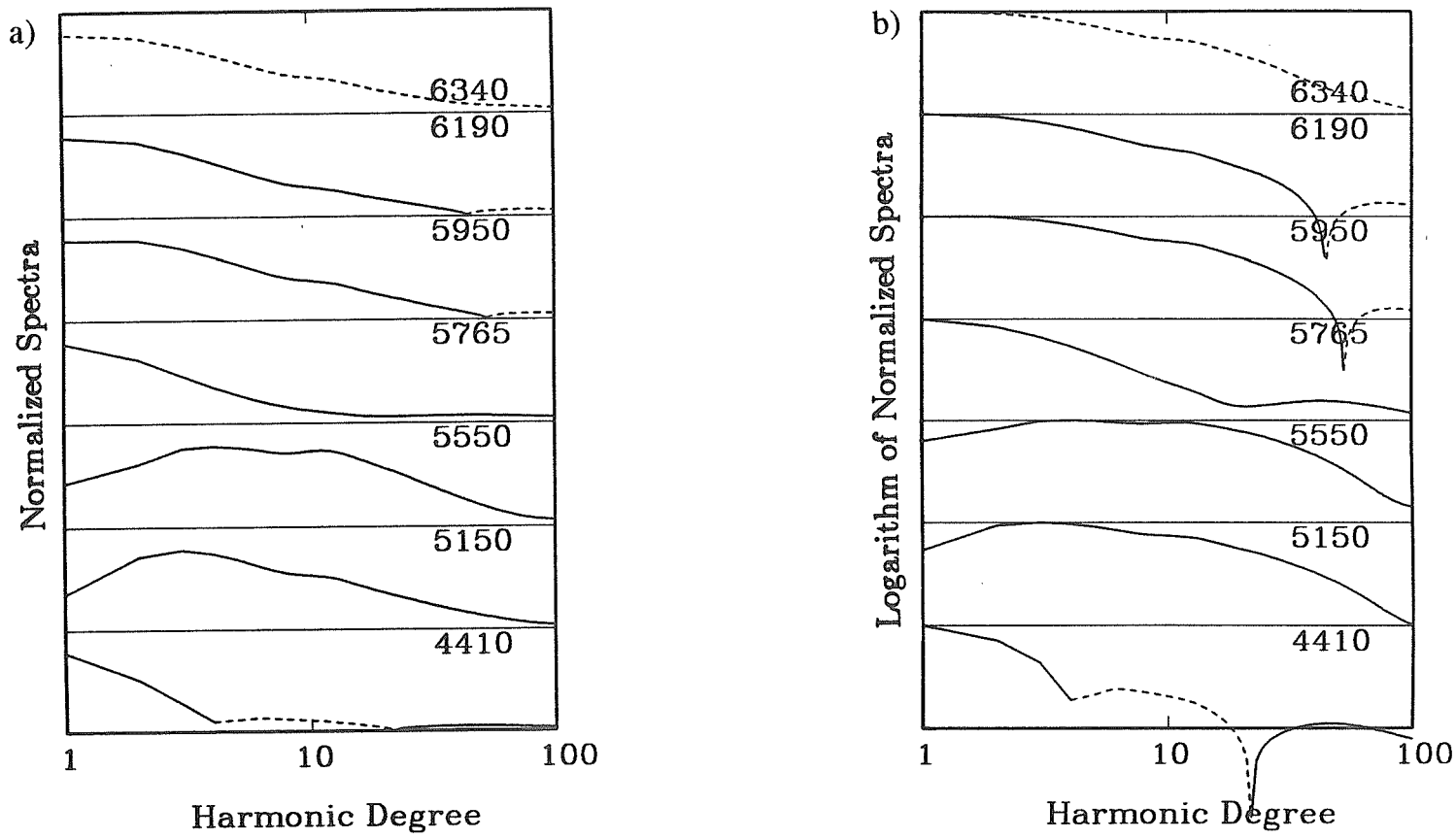


Figure 1.33 (a) Plot of Normalized Spectra of S-wave model with logarithmic intercept estimates versus harmonic degree on a logarithmic axis. (b) Plot of Logarithm of normalized spectra of S-wave model with logarithmic intercept estimates versus logarithm of harmonic degree.

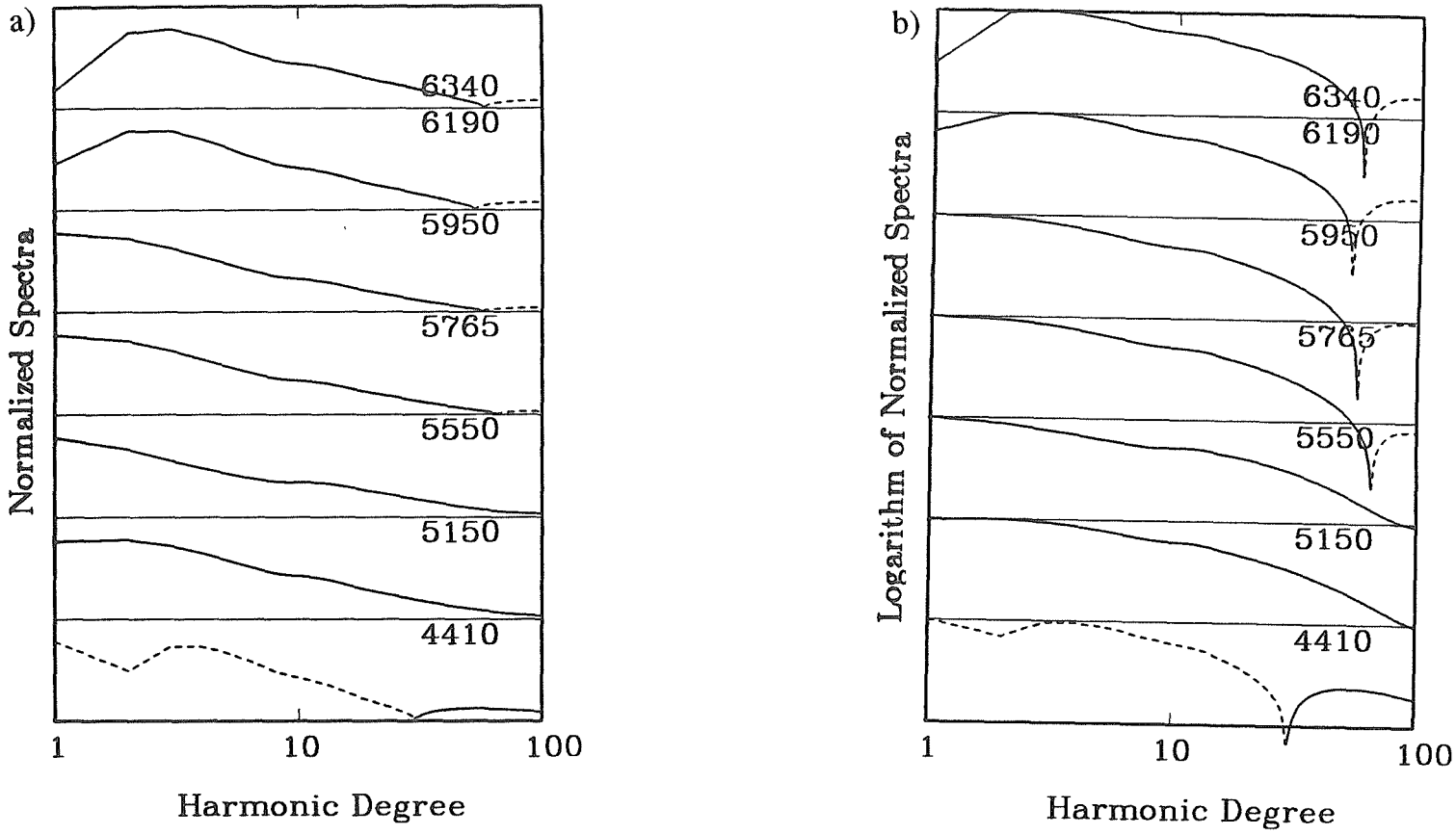


Figure 1.34 (a) Plot of Normalized Spectra of S-wave model with lifted asymptotes versus harmonic degree on a logarithmic axis. (b) Plot of Logarithm of normalized spectra of S-wave model with lifted asymptotes versus logarithm of harmonic degree.

dence in its shape, but in contrast the variability in the spectral shape of the upper and lowermost layers suggests a lack of robustness in the spectral description of these layers.

It would also be interesting to extend the analysis of *Kaula* [1980], where he used the spectra of surface fields to constrain material properties such as the mean lithospheric viscosity. We could use the seismic spectra as a function of depth to derive and constrain material properties at depth. The heat flow discussion above is one example, as is matching the geoid and plate velocity spectra to constrain the radial variation of viscosity [*Peltier et al.*, 1989; *Richards and Hager*, 1984].

If it's true that there are compositional variations, then it is easy to believe that the distribution of heat sources in the mantle could be non-uniform radially and laterally. Due to the link between high temperature and magmatism and the incompatible nature of the majority of heat sources (U, Th and K), there is obviously a feed-back between the two. The influence of an irregular distribution of heat sources on mantle convection simulations has yet to be investigated. We would expect it to lead to a whiter thermal variance spectra.

It is clear that this method has already presented us with much interesting information, but like tomography, these early results from the global stochastic models are more appetizers than meals. This method can potentially provide information about the whole spectrum of heterogeneity. This method shall bear great fruit if the spectral resolution can be improved. This can be achieved only by analyzing the data at many more separations. For instance, for uniform sampling, to differentiate between signal at harmonic degree

100 and 101, we would require a stable estimate of the variance estimate over an area of scale 400km as compared to an area of state 396km! Sampling is insufficient to allow this degree of spectral resolution presently, but given more computational power, we can certainly improve on the present study. As shown above, the spectral interpretations are also limited by the depth resolution in the shallow mantle. This could potentially be addressed by using regional data ($\Delta<30^o$).

1.7 Conclusion

We have developed a method to image the spectrum of the Earth's heterogeneity as a function of depth. It allows us an indication of the incoherent noise in the data-set, and can be adapted to estimate the amount of signal that is due to structures smaller than the grid-size (or harmonic parameterization) of a deterministic inversion. This allows estimation of aliasing, error variance and optimal parameterizations of deterministic inversions.

We discover that the Earth's heterogeneity is concentrated in the upper 400km of the mantle. There is evidence of a slight increase in heterogeneity at the base of the mantle. This suggests a dramatic deviation away from uniform convection as presently simulated (provided that the seismic velocity temperature derivatives vary little across the mantle). Hence there must be significant radial variations in material properties. The pattern of lateral heterogeneity is very similar for both the compressional and shear velocities. The ratio of S-wave to P-wave variations for much

of the mantle is $\delta \ln(V_S)/\delta \ln(V_P) \approx 5$. This high value could be a reflection of poor spectral and depth resolution in the upper mantle. A signal averaged value of 2 is suggested by the raw data following the subtraction of the intercept estimate, assuming perfect correlation of the two fields and equivalent sampling. This ratio is also appreciably larger than expected from low pressure laboratory measurements (≈ 1). In the uppermost mantle this has been explained as being due to partial melting. In the lowermost mantle there is a suggestion that it is the natural consequence of the effect of compression on materials [*Anderson*, 1987a].

There is a marginal feature at the 670km discontinuity, which might suggest the deflections of a phase boundary or a weak thermal boundary layer if there were no change in the temperature derivatives of material across this boundary; it also be a result of the high damping and poor resolution of the bin just above the 670km discontinuity. There is no strong evidence of periodicity in the travel-time auto-covariance, hence the underlying flow is unlikely to be periodic and the convection is unlikely to be steady. The results show a major decrease in power in going from the upper to the lower mantle. They also show an abrupt drop in heterogeneity at around 300-400km. This has previously been interpreted as being the contrast between continental shields and oceans. Hence we suggest that this signature is present globally, and is the most dominant seismic heterogeneity of the Earth's mantle.

REFERENCES

Agnon, A., and M. S. T. Bukowinski, δ_S at High Pressure: Is the lower mantle sub-solidus?, *EOS, Trans. Am. Geophys. Un.*, **70**, [43], 1212, 1989.

Aki, K., Analysis of the seismic coda of local earthquakes as scattered waves, *J. Geophys. Res.*, **74**, 615-631, 1969.

Aki, K., Scattering of P waves under the Montana LASA, *J. Geophys. Res.*, **78**, 1334-1346, 1973.

Aki, K., and B. Chouet, Origin of coda waves: source, attenuation, and scattering effect, *J. Geophys. Res.*, **80**, 3322-3343, 1975.

Aki, K., and P. G. Richards, *Quantitative Seismology*, W.H. Freeman, San Francisco, 1980.

Allegre, C. J., and D. L. Turcotte, Implications of a two-component marble-cake mantle, *Nature*, **323**, 123-127, 1986.

Anderson, D. L., The deep structure of continents, *J. Geophys. Res.*, **84**, 7555-7560, 1979a.

Anderson, D. L., The frequency dependence of Q in the Earth and implications for mantle rheology and Chandler wobble, *Geophys. J. Roy. Astr. Soc.*, **58**, 431-440, 1979b.

Anderson, D. L., A seismic equation of state 2. Shear properties and thermodynamics of the lower mantle, *Phys. Earth Planet. Int.*, **45**, 307-323, 1987a.

Anderson, D. L., Thermally induced phase changes, lateral heterogeneity of the mantle, continental roots and deep slab anomalies, *J. Geophys. Res.*, 13968-13980, 1987b.

Anderson, D. L., Temperature and Pressure derivatives of Elastic Constants with application to the Mantle, *J. Geophys. Res.*, **93**, 4688-4700, 1988.

Anderson, D. L., Where on Earth is the Crust?, *Physics Today*, **42**, 38-46, 1989.

Ashby, M. F., and R. A. Verrall, Micromechanisms of flow and fracture and their relevance to the rheology of the mantle, *R. Soc. Lond. A.*, **288**, 59-95, 1977.

Bercovici, D., G. Schubert, and G. A. Glatzmaier, Three-dimensional spherical models of convection in the Earth's Mantle, *Science*, **244**, 950-955, 1989.

Berteussen, K. A., A. Christoffersson, E. S. Husebye, and A. Dahle, Wave scattering theory in analysis of P-wave anomalies at NORSAR and LASA, *Geophys. J. Roy. Astr. Soc.*, **42**, 403-417, 1975.

Berteussen, K. A., E. S. Husebye, R. F. Mereu, and A. Rau, Quantitative assessment of the crust-upper mantle heterogeneities beneath the Guaribidanur seismic array in Southern India, *Earth Planet. Sci. Lett.*, **37**, 326-332, 1977.

Capon, J., Characterization of crust and upper mantle structure under LASA as a random medium, *Bull. Seism. Soc. Am.*, **64**, 235-266, 1974.

Capon, J., and K. A. Berteussen, A random-medium analysis of crust- and upper-mantle structure under NORSAR, *Geophys. Res. Lett.*, **1**, 327-328, 1974.

Chernov, L. A., *Wave propagation in a random medium*, McGraw-Hill Book Company, Inc., New York, 1960.

Chopelas, A., and R. Boehler, Thermal measurements at very high pressure, systematics, and a case for a chemically homogeneous mantle, *Geophys. Res. Lett.*, **16**, 1347-1350, 1989.

Christensen, U., Convection with pressure- and temperature-dependent non-Newtonian rheology, *Geophys. J. Roy. Astr. Soc.*, **77**, 343-384, 1984.

Christensen, U. R., and D. A. Yuen, Time- dependent convection with non-Newtonian viscosity, *J. Geophys. Res.*, **94**, 814-820, 1989.

Claerbout, J. F., *Imaging of the Earth's Interior*, Blackwell, Oxford, 1985.

Clark, I., *Practical Geostatistics*, Applied Science Publishers, London, 1979.

Clayton, R. W., and R. P. Comer, A tomographic analysis of mantle heterogeneities from body wave travel times, *EOS, Trans. AGU*, **62**, 776, 1983.

Creager, K. C., and T. H. Jordan, Slab penetration into the lower mantle, *J. Geophys. Res.*, **89**, 3031-3049, 1984.

Creager, K. C., and T. H. Jordan, Slab penetration into the lower mantle beneath the Mariana and other island arcs of the northwest Pacific, *J. Geophys. Res.*, **91**, 3573-3580, 1986.

Davies, G., Role of the lithosphere in mantle convection, *J. Geophys. Res.*, **93**, 10451-10466, 1988.

Davies, G. F., Geophysical and isotopic constraints on mantle convection: an interim synthesis, *J. Geophys. Res.*, **89**, 6017-6040, 1984.

Davies, G. F., and M. Gurnis, Interaction of mantle dregs with convection: Lateral heterogeneity at the core-mantle boundary, *Geophys. Res. Lett.*, **13**, 1517-1520, 1986.

Davies, J. H., and R. W. Clayton, Lower-mantle S-wave tomography, *Trans. AGU, EOS*, **367**, 1099, 1986.

Davies, J. H., O. Gudmundsson, and R. W. Clayton, Errors and small scale structure inferred from areal statistics of ISC residuals, *Seism. Res. Lett.*, **359**, 40, 1988.

Doyle, H. A., and A. L. Hales, An analysis of the travel times of S waves to North American stations, *Bull. Seismol. Soc. Am.*, **57**, 761-771, 1967.

Duffy, T. S., and D. L. Anderson, Seismic velocities in mantle minerals and the mineralogy of the upper mantle, *J. Geophys. Res.*, **94**, 1895-1912, 1989.

Dziewonski, A., and D. L. Anderson, Preliminary reference Earth model, *Phys. Earth Planet. Int.*, **25**, 297-356, 1981.

Dziewonski, A. M., Mapping the lower mantle: Determination of lateral heterogeneity in P velocity up to degree and order 6, *J. Geophys. Res.*, **89**, 5929-5952, 1984.

Dziewonski, A. M., and J. H. Woodhouse, Global Images of the Earth's Interior, *Science*, **236**, 37-48, 1987.

Flatte, S. M., and R. Wu, Small-scale structure in the lithosphere and asthenosphere deduced from arrival time and amplitude fluctuations, *J. Geophys. Res.*, **93**, 6601-6614, 1988.

Frankel, A., and R. W. Clayton, A finite-difference simulation of wave propagation in two-dimensional random media, *Bull. Seism. Soc. Am.*, **74**, 2167-2186, 1984.

Frankel, A., and R. W. Clayton, Finite difference simulations of seismic scattering: Implications for the propagation of short-period seismic waves in the crust and models of crustal heterogeneity, *J. Geophys. Res.*, **91**, 6465-6469, 1986.

Gao, L. S., N. N. Biswas, L. C. Lee, and K. Aki, Effects of multiple scattering on coda waves in three dimensional medium, *Pageoph*, **121**, 3-15, 1983a.

Gao, L. S., L. C. Lee, N. N. Biswas, and K. Aki, Comparison of the effects between single and multiple scattering on coda waves for local earthquakes, *Bull. Seismol. Soc. Am.*, **73**, 377-389, 1983b.

Giardini, D., X. Li, and J. Woodhouse, Splitting functions of long-period normal modes of the Earth, *J. Geophys. Res.*, **93**, 13716-13742, 1988.

Given, J. W., and D. V. Helmberger, Upper mantle structure of Northwestern Eurasia, **85**, 7183-7194, 1980.

Glatzmaier, G. A., Numerical Simulations of Mantle Convection : Time-Dependent Three-Dimensional Compressible Spherical Shell, *Geophys. Astrophys. Fluid Mech.*, **43**, 223-264, 1988.

Gradshteyn, I. S., and I. M. Ryzhig, *Table of integrals series and products corrected and enlarged edition*, Academic Press, New York, 1980.

Grand, S. P., Tomographic Inversion for Shear Velocity Beneath the North American Plate, *J. Geophys. Res.*, **92**, 14065-14090, 1988.

Grand, S. P., and X. Ding, Residual Spheres and Slab Penetration into the Lower Mantle, *EOS, Trans. Am. Geophys. Un.*, **70**, [43], 1322, 1989.

Grand, S. P., and D. V. Helmberger, Upper mantle shear structure beneath the northwest Atlantic Ocean, *J. Geophys. Res.*, **89**, 11,465-11,475, 1984a.

Grand, S. P., and D. V. Helmberger, Upper mantle shear structure of North America, *Geophys J. Roy. Astr. Soc.*, **76**, 399-438, 1984b.

Graves, R., and D. Helmberger, Upper mantle cross section from Tonga to Newfoundland, *J. Geophys. Res.*, **93**, 4701-4711, 1988.

Gudmundsson, O., Some problems in global tomography: modeling the core mantle boundary and statistical analysis of travel time data, Ph. D. thesis, pp. 225, California Institute of Technology, 1989.

Gudmundsson, O., J., H. Davies, and R. W. Clayton, Stochastic Analysis of global travel time data : Mantle Heterogeneity and random errors in the ISC data, *Geophysical Journal (in press)*, 1989.

Gudmundsson, O., J. H. Davies, and R. W. Clayton, Stochastic Analysis of Global Travel Time Data, *Trans. Am. Geophys. Un. EOS*, **69**, 1308, 1988.

Gurnis, M., and G. F. Davies, Mixing in numerical models of mantle convection incorporating plate kinematics, *J. Geophys. Res.*, **91**, 6375-6395, 1986.

Haddon, R. A. W., and J. R. Cleary, Evidence for scattering of seismic PkP waves near the mantle-core boundary, *Phys. Earth Planet. Int.*, 8, 211-234, 1974.

Hager, B. H., Subducted slabs and the geoid: Constraints on Mantle Rheology and Flow, *J. Geophys. Res.*, 89, 6003-6015, 1984.

Hager, B. H., and R. W. Clayton, Constraints on the structure of mantle convection using seismic observations, flow models, and the geoid, in *Mantle convection: plate tectonics and global dynamics, the fluid mechanics of astrophysics and geophysics*, vol. 4, edited by W. R. Peltier, pp. 657-764, Gordon and Breach Science Publishers, Montreaux, 1989.

Hager, B. H., R. W. Clayton, M. A. Richards, R. P. Comer, and A. M. Dziewonski, Lower mantle heterogeneity, dynamic topography and the geoid, *Nature*, 313, 541-545, 1985.

Hager, B. H., and R. J. O'Connell, Subduction zone dip angles and flow driven by plate motion, *Tectonophysics*, 50, 111-133, 1978.

Hart, S., and A. Zindler, Constraints on the nature and development of chemical heterogeneities in the mantle, in *Mantle convection: plate tectonics and global dynamics, The fluid mechanics of astrophysics and geophysics*, vol. 4, edited by W. R. Peltier, pp. 261-388, Gordon and Breach Science Publishers, Montreaux, 1989.

Helmberger, D. V., G. Engen, and S. Grand, Upper-Mantle cross section from California to Greenland, *J. Geophys.*, 58, 92-100, 1985.

Henley, S., *Nonparametric geostatistics*, Applied Science Publishers Ltd, London, 1981.

Herrin, E., Introduction to the 1968 Seismological Tables for P phases., *Bull. Seism. Soc. Am.*, **58**, 1193, 1968.

Hoffman, N. R. A., and D. P. McKenzie, The destruction of geochemical heterogeneities by differential fluid motions during mantle convection, *Geophys. J. Roy. Astr. Soc.*, **82**, 163-206, 1985.

Holbrook, W. S., An interpretation of wide-angle compressional and shear wave data in Southwest Germany: Poisson's ratio and petrological implications, *J. Geophys. Res.*, **93**, 12081-12106, 1988.

Houseman, G., The dependence of convection planform on mode of heating, *Nature*, **332**, 346-349, 1988.

Inoue, H., Y. Fukao, and Y. Ogata, Whole mantle P-wave travel time tomography, *submitted to Phys. Earth Planet. Int.*, 1989.

Isaak, D. G., O. L. Anderson, and T. Goto, Elasticity of single crystal forsterite measured to 1700 K, *J. Geophys. Res.*, **94**, 5895-5907, 1989a.

Isaak, D. G., O. L. Anderson, and T. Goto, Measured Elastic Constants of Single-Crystal MgO up to 1800K, *Phys. Chem. Mineral.*, **16**, 704-713, 1989b.

Isaak, D. G., R. E. Cohen, and M. J. Mehl, Calculated Elastic and Thermal Properties of MgO at High Pressures and Temperatures, *Preprint*, 1989c.

Jarvis, G. T., The long-wavelength component of mantle convection, *Phys. Earth Planet. Int.*, **40**, 24-42, 1985.

Jarvis, G. T., and W. R. Peltier, Lateral heterogeneity in the convecting mantle, *J. Geophys. Res.*, **91**, 435-451, 1986.

Jarvis, G. T., and W. R. Peltier, Convection models and geophysical observations, in *Mantle convection: plate tectonics and global dynamics, The fluid mechanics of astrophysics and geophysics*, vol. **4**, edited by W. R. Peltier, pp. 479-594, Gordon and Breach Science Publishers, Montreaux, 1989.

Jeanloz, R., High pressure chemistry of the Earth's mantle and core, in *Mantle convection: plate tectonics and global dynamics, The fluid mechanics of astrophysics and geophysics*, vol. **4**, edited by W. R. Peltier, pp. 203-260, Gordon and Breach Science Publishers, Montreaux, 1989.

Jeffereys, H., The times of P, S and SKS, and the velocities of P and S, *Mon. Not. Roy. Astr. Soc. (geophys. suppl.)*, **4**, 498-533, 1939.

Jeffereys, H., and K. E. Bullen, *Seismological Tables*, British Association, Gray Milne Trust, 1940.

Jordan, T., The continental tectosphere, *Rev. Geophys. Space Phys.*, **13**, 1-12, 1975.

Jordan, T. H., Composition and development of the continental tectosphere, *Nature*, **274**, 544-548, 1978.

Jordan, T. H., The deep structure of the continents, *Sci. Am.*, **240**, 92-107, 1979.

Jordan, T. H., Global tectonics regionalization for seismological data analysis, *Bull. Seismol. Soc. Am.*, **71**, 1131-1141, 1981.

Jordan, T. H., A. L. Lerner-Lam, and K. C. Creager, Seismic imaging of boundary layers and deep mantle convection, in *Mantle*

convection: plate tectonics and global dynamics, *The fluid mechanics of astrophysics and geophysics*, vol. 4, edited by W. R. Peltier, pp. 97-203, Gordon and Breach Science Publishers, Montreaux, 1989.

Kaula, W. M., Material properties for Mantle Convection Consistent with Observed Surface Fields, *J. Geophys. Res.*, **85**, 7031-7044, 1980.

Le Fevre, L. V., and D. V. Helmberger, Upper mantle P-velocity of the Canadian shield, *Preprint*, 1988.

Lerner-Lam, A. L., and T. H. Jordan, How thick are the continents?, *J. Geophys. Res.*, **92**, 14007-14026, 1987.

Lindsey, J. P., The Fresnel zone and its interpretive significance, *Geophysics: The Leading Edge*, **8**, 33-39, 1989.

Masters, G., Observational Constraints on the chemical and thermal structure of the Earth's deep interior, *Geophys. J. Roy. Astr. Soc.*, 507-534, 1979.

Masters, G., T. H. Jordan, P. G. Silver, and F. Gilbert, Aspherical earth structure from fundamental spheroidal-mode data, *Nature*, **298**, 609-613, 1982.

Nataf, H.-C., I. Nakanishi, and D. L. Anderson, Anisotropy and shear-velocity heterogeneities in the upper mantle, *Geophys. Res. Lett.*, **11**, 109-112, 1984.

Nataf, H.-C., I. Nakanishi, and D. L. Anderson, Measurements of mantle wave velocities and inversion for lateral heterogeneities and anisotropy, III, Inversion, *J. Geophys. Res.*, **91**, 7261-7307, 1986.

Nolet, G., Seismic wave propogation and seismic tomography, in *Seismic tomography with applications in global seismology and exploration geophysics*, edited by G. Nolet, pp. 1-23, D. Reidel Publishing Co., Dodrecht, 1987.

Olson, P., A. Yuen, and D. Balsinger, Convective mixing and the fine structure of mantle heterogeneity, *Phys. Earth Planet Int.*, **36**, 291-304, 1984a.

Olson, P. D., A. Yuen, and D. Balsinger, Mixing of passive heterogeneities by mantle convection, *J. Geophys. Res.*, **89**, 425-436, 1984b.

Peltier, W. R., G. T. Jarvis, A. M. Forte, and L. P. Solheim, The radial structure of the mantle general circulation, in *Mantle convection: plate tectonics and global dynamics, the mechanics of astrophysics and geophysics*, vol. 4, edited by W. R. Peltier, pp. 765-816, Gordon and Breach Science Publishers, Montreaux, 1989.

Pollack, H. N., and D. S. Chapman, On the regional variation of heat flow, geotherms, and lithospheric thickness, *Tectonophysics*, **38**, 279-296, 1977.

Powell, C. A., and A. S. Meltzer, Scattering of P-waves beneath SCARLET in Southern California, *Geophys. Res. Lett.*, **11**, 4811-484, 1984.

Ricard, Y., C. Vigny, and C. Froidevaux, Mantle heterogeneities, geoid and plate motion: A Monte Carlo inversion, *J. Geophys. Res.*, **94**, 13739-13754, 1989.

Richards, M. A., and B. H. Hager, Geoid anomalies in a dynamic earth, *J. Geophys. Res.*, **89**, 5987-6002, 1984.

Ringwood, A. E., Phase transformations and differentiation in subducted lithosphere: implications for mantle dynamics, basalt petrogenesis, and crustal evolution, *J. Geol.*, **90**, 611-643, 1982.

Ringwood, A. E., Mantle Dynamics and Basalt Petrogenesis, *Tectonophysics*, **112**, 17-34, 1985.

Ritzwoller, M., G. Masters, and F. Gilbert, Observations of anomalous splitting and their interpretation in terms of aspherical structure, *J. Geophys. Res.*, 10203-10228, 1986.

Sammis, C. G., J. C. Smith, G. Schubert, and D. A. Yuen, Viscosity-depth profile of the Earth's mantle: effects of polymorphic phase transitions, *J. Geophys. Res.*, **82**, 3747-3761, 1977.

Sato, H., I. S. Sacks, and T. Murase, The use of laboratory velocity data for estimating temperature and partial melt fraction in the low-velocity zone: comparison with heat flow and electrical conductivity studies, *J. Geophys. Res.*, **94**, 5689-5704, 1989.

Sato, H., I. S. Sacks, T. Murase, G. Muncill, and H. Fukuyama, Attenuation of Compressional Waves in Peridotite Measured as a Function of Temperature at 200MPa, *Pageoph*, **128**, 433-447, 1988.

Schmeling, H., Compressible convection with constant and variable viscosity: The effect on slab formation geoid and topography, *J. Geophys. Res.*, **94**, 12463-12481, 1989.

Sclater, J. G., B. Parson, and C. Jaupart, Oceans and Continents : similarities and differences in the mechanisms of heat loss, *J. Geophys. Res.*, **86**, 11535-11552, 1981.

Sengupta, M. K., R. E. Hassell, and R. W. Ward, Three-dimensional seismic velocity structure of the Earth's mantle using body wave travel times from intra-plate and deep-focus earthquakes, *J. Geophys. Res.*, **86**, 3913-3934, 1981.

Sengupta, M. K., and M. N. Toksoz, Three dimensional model of seismic velocity variations in the Earth's Mantle, *Geophys. Res. Lett.*, 84-86, 1976.

Silver, P., R. W. Carlson, and P. Olson, Deep slabs geochemical heterogeneity and the large-scale structure of mantle convection: Investigation of an enduring paradox, *Ann. Rev. Earth Planet. Sci.*, **16**, 477-541, 1988.

Sipkin, S. A., and T. H. Jordan, Lateral heterogeneity of the upper mantle determined from the travel times of multiple ScS, *J. Geophys. Res.*, **81**, 6307-6320, 1976.

Stevenson, D. J., Formation and early evolution of the earth, in *Mantle convection: plate tectonics and global dynamics, The mechanics of astrophysics and geophysics*, vol. 4, edited by W. R. Peltier, pp. 817-874, Gordon and Breach Science Publishers, Montreaux, 1989.

Stewart, C. A., and D. L. Turcotte, The route to chaos in Thermal Convection at Infinite Prandtl Number, 1, Some Trajectories and Bifurcations, *J. Geophys. Res.*, **94**, 13707-13718, 1989.

Sumino, Y., and O. L. Anderson, Elastic constants of minerals, **3**, 39-137, 1984.

Tanimoto, T., The Backus-Gilbert approach to the three-dimensional structure in the upper mantle, I, Lateral Variation of

surface wave phase velocity with its error and resolution, *Geophys. J. Roy. Astr. Soc.*, **82**, 105-124, 1985.

Tanimoto, T., The Backus-Gilbert approach to the three dimensional structure in the upper mantle, II, SH and SV velocity, *Geophys. J. Roy. Astr. Soc.*, **84**, 49-70, 1986.

Tanimoto, T., The 3-D shear-wave structure in the mantle by overtone waveform inversion. I . Radial seismogram inversion, *Geophys. J. Roy. Astr. Soc.*, **89**, 713-740, 1987.

Tanimoto, T., The 3-D shear-wave structure in the mantle by overtone waveform inversion. II . Inversion of X waves, R waves and G waves, *Geophys. J. Roy. Astr. Soc.*, **93**, 321-334, 1988.

Tanimoto, T., Long wavelength S-wave velocity structure throughout the mantle, *Geophys. J. Roy. Astr. Soc. (in press)*, 1989a.

Tanimoto, T., Waveform inversion for three-dimensional density and S-wave structure, *Preprint*, 1989b.

Tatarski, V. I., *Wave propagation in a turbulent medium*, McGraw-Hill Book Company, Inc., New York, 1961.

Walck, M. C., Three-dimensional Vp/Vs variations for the Coso region California, *J. Geophys. Res.*, **93**, 2047-2052, 1988.

Walck, M. C., and R. W. Clayton, Analysis of Upper Mantle Structure Using Wave Field Continuation of P waves, *Bull. Seism. Soc. Am.*, **74**, 1703-1719, 1984.

Weinstein, S. A., P. L. Olson, and D. A. Yuen, Time-Dependent large aspect-ratio thermal convection in the Earth's Mantle, *Preprint*, 1989.

White, W. M., and A. W. Hofmann, Sr and Nd isotope geochemistry of oceanic basalts and mantle evolution, *Nature*, **296**, 821-825, 1982.

Wielandt, E., On the validity of the ray approximation for interpreting delay times, in *Seismic tomography with applications in global seismology and exploration geophysics*, pp. 85-98, D. Reidel Publishing Co., Dodrecht, 1987.

Woodhouse, J. H., and A. M. Dziewonski, Mapping the upper mantle, three-dimensional modeling of Earth structure by inversion of seismic waveforms, *J. Geophys. Res.*, **89**, 5953-5986, 1984.

Wu, R. S., and K. Aki, Elastic waves scattering by a random medium and the small-scale inhomogeneities in the lithosphere, *J. Geophys. Res.*, **90**, 10261-10273, 1985.

Wu, R. S., and K. Aki, Introduction: Seismic Wave Scattering in a three dimensionally heterogeneous Earth, *Pageoph*, **128**, 1-6, 1988.

Zhang, Y. S., and T. Tanimoto, Three-dimensional Modelling of upper mantle structure under the Pacific Ocean and surrounding area, *Geophys. J. Int.*, **98**, 255-269, 1989.

Zhou, H., How well can we resolve the deep seismic slab with seismic tomography?, *Geophys. Res. Lett.*, **15**, 1425-1428, 1988.

Zhou, H., D. L. Anderson, and R. W. Clayton, Modeling of residual spheres for subduction zone earthquakes in the northwest Pacific, *J. Geophys. Res.*, submitted 1988, 1988.

Zhou, H.-W., Travel time tomographic studies of seismic structures around subducted lithospheric slabs, Ph. D. thesis, pp. 397, California Institute of Technology, 1989.

Zindler, A., and S. Hart, Chemical Geodynamics, *Ann. Rev. Earth Planet. Sci.*, **14**, 493-571, 1986.

Appendix A

Method for simultaneous derivation of model and estimation of intercept

We have estimated intercepts first and then inverted the resulting data to obtain the model. Given the slight offsets of model predicted data from data for all scales of individual reference rays, it suggests that better models could potentially be obtained if the intercept was estimated simultaneously with the model. A simultaneous solution would also give us a quantitative measure of the trade-off between model and intercept. The advantage of the separate solution includes its efficiency, simplicity and flexibility.

The problem of simultaneous solution for a continuous model and a discrete set of parameters is also encountered in the estimation of a local velocity field from local earthquakes. One needs to solve for the earthquake locations and also the velocity perturbations. A simultaneous solution is advantageous but has only been sparingly undertaken. More commonly the earthquakes are located first, then the velocity field is derived, then the iteration is repeated, until convergence is reached.

Following *Engdahl and Gubbins* [1988] we can undertake the simultaneous inversion as follows.

$$\mathbf{D} = \mathbf{G}\mathbf{X}\mathbf{F}^{\mathrm{T}} \qquad \text{(A1)}$$

is expression of the forward problem (equation 1.60 of

Theory section). This can be converted to the simultaneous problem by including the full data on the left-hand side, i.e. before the intercepts are removed, and include a vector of the intercepts on the right hand side.

$$\mathbf{D} = \mathbf{G}\mathbf{X}\mathbf{F}^{T} + \mathbf{K}\mathbf{D}_{0} \tag{A2}$$

where $\mathbf{D}_0$ is the matrix of intercepts, a diagonal matrix with the intercepts on the diagonal, and $\mathbf{K}$ is a matrix full of elements of value unity. Hence we can post multiply all terms by $\mathbf{F}[\mathbf{F}^{T}\mathbf{F}]^{-1}$, and since $\mathbf{D}_0$ is diagonal it commutes with this term, hence we have

$$\mathbf{D}\mathbf{F}[\mathbf{F}^{T}\mathbf{F}]^{-1} = \mathbf{G}\mathbf{X} + \mathbf{K}\mathbf{F}[\mathbf{F}^{T}\mathbf{F}]^{-1}\mathbf{D}_{0} \tag{A3}$$

Now if we let $\mathbf{A}=\mathbf{G}$, $\mathbf{T}=\mathbf{D}\mathbf{F}[\mathbf{F}^{T}\mathbf{F}]^{-1}$ and $\mathbf{B}=\mathbf{K}[\mathbf{F}(\mathbf{F}^{T}\mathbf{F})^{-1}]$; then we get

$$\mathbf{T} = \mathbf{A}\mathbf{X} + \mathbf{B}\mathbf{D}_{0} \tag{A4}$$

and $\mathbf{D}_0$ and $\mathbf{X}$ can be solved for as follows.

$$\begin{aligned} \mathbf{X} &= (\mathbf{O}\mathbf{A})^{-1}\mathbf{O}\mathbf{T} \\ \mathbf{D}_{0} &= (\mathbf{B}^{T}\mathbf{B})^{-1}(\mathbf{B}^{T}\mathbf{T} - \mathbf{B}^{T}\mathbf{A}\mathbf{X}) \\ &\textit{where} \\ \mathbf{O} &= (\mathbf{A}^{T} - \mathbf{A}^{T}\mathbf{B}(\mathbf{B}^{T}\mathbf{B})^{-1}\mathbf{B}^{T}) \end{aligned} \tag{A5}$$

This involves the inversion of two matrices. $\mathbf{B}^{T}\mathbf{B}$ and $\mathbf{O}\mathbf{A}$. Not all the parameters can be determined, hence one must damp the solution. These matrices can again be inverted using

singular value decomposition and the damping can be achieved by limiting oneself to the larger singular values. There is no *a priori* information on the relative damping of the intercepts and the model, hence a trade-off surface would have to be defined.

Another solution of the simultaneous relocation / velocity problem has been proposed by *Pavlis and Booker* [1980]. They show that it is possible to derive the continuous part of the model using a set of equations that is independent of the discrete part of the model, provided that there are more data than the number of discrete parameters inverted for. These equations, which are independent of the locations in this application (the intercept in ours) are called the annulled set. The discrete part of the model can be derived from a second set of equations that relates both the continuous and discrete parts of the model. Following *Menke* [1989] ,the Pavlis and Booker algorithm can be understood as follows. Let $\mathbf{d}=\mathbf{G}\mathbf{m}$ where $\mathbf{m}_1$ is the vector of earthquake locations and $\mathbf{m}_2$ is the vector of velocity parameters. Then the forward problem can be written $\mathbf{d}=[\mathbf{G}_1,\mathbf{G}_2]\ [\mathbf{m}_1,\mathbf{m}_2]^T= \mathbf{G}_1\mathbf{m}_1+\mathbf{G}_2\mathbf{m}_2$. Assume that $\mathbf{G}_1$ has a SVD $\mathbf{G}_1=\mathbf{U}_1\mathbf{W}_1\mathbf{V}_1^T$ with n nonzero singular values, then $\mathbf{U}_1$ can be partitioned into two submatrices, one corresponding to the nonzero singular values $\mathbf{U}_{1n}$ and the other to the zero singular values $\mathbf{U}_{10}$. Hence the data can be split as follows $\mathbf{d}=\mathbf{d}_1+\mathbf{d}_2=\mathbf{K}\mathbf{d}+(\mathbf{I}-\mathbf{K})\mathbf{d}$, where $\mathbf{K}=\mathbf{U}_{1n}\mathbf{U}^T_{1n}$, the data resolution matrix of $\mathbf{G}_1$. If we premultiply the forward problem by $\mathbf{U}_{10}$, the submatrix of $\mathbf{U}_1$ with the zero singular values then we get

$$\mathbf{U}^T_{10}\mathbf{d}_2=\mathbf{U}^T_{10}\mathbf{G}_2\mathbf{m}_2 \qquad \text{(A6)}$$

because $\mathbf{U}^T_{10}\mathbf{U}_{1n}=\mathbf{0}$. While if we premultiply by $\mathbf{U}^T_{1n}$ we get

$$\mathbf{U}^T_{1n}\mathbf{d}_2=\mathbf{U}^T_{1n}\mathbf{G}_1\mathbf{m}_1+\mathbf{U}^T_{1n}\mathbf{G}_2\mathbf{m}_2. \qquad \text{(A7)}$$

We have split the problem into two, notice equation 1.A6 is independent of the $\mathbf{m}_1$ part of the model; i.e. these are the annulled equations discussed above. While we solve for $\mathbf{m}_2$ by inverting equation 1.A7. Hence, the annulled equations are the equations relating to the zero singular values of the kernel matrix $\mathbf{G}_1$, (which in this application is the matrix relating the earthquake hypocentral parameters to the travel-times; notice it is block diagonal and the singular value decomposition can be done on the resulting submatrices).

Converting this elegant method to apply to our problem is non-trivial since our forward problem is $\mathbf{D}=\mathbf{G}\mathbf{X}\mathbf{F}^T+\mathbf{d}_0$. $\mathbf{D}$ and $\mathbf{X}$ are matrices but these can be strung out to give vectors, this is not insurmountable (of course $\mathbf{G}$ and $\mathbf{F}$ will also need to be changed, and in fact they must be combined to give one matrix $\mathbf{P}$, a tedious but straight forward task). If the initial equation in subscript form is

$$D_{i,j} = G_{i,k}X_{k,l}F_{l,j} + D_{i,j}^{\,0}$$

then

$$d_p = P_{p,q}x_q + d_p^{\,0}$$

where

$$p = i + I \times j;\ \ q = k + K \times l;\ \ and\ P_{p,q} = G_{i,k}F_{l,j} \qquad \text{(A8)}$$

Notice p and q must uniqely define i,j,k, and l; therefore, I is equal to the number of i elements and similarily K is the number of elements indexed by k.

The subsequent solution is no problem since now the forward problem has the same format as above. We can just proceed as above and we get the following two equations

$$\mathbf{U}^T_{10}\mathbf{d}_2=\mathbf{U}^T_{10}\mathbf{G}_2\mathbf{m}_2 \tag{A9}$$

$$\mathbf{U}^T_{1n}\mathbf{d}_2=\mathbf{U}^T_{1n}\mathbf{P}\mathbf{m}_1+\mathbf{U}^T_{1n}\mathbf{G}_2\mathbf{m}_2 \tag{A10}$$

An advantage with the above formulation is that the interpretation of P is changed, it now relates the data to a particular spectral bin in a particular depth bin in the Earth. Hence, resolution estimates of this matrix would tell us how well specific spectral components were resolved at specific depths. This formulation also has advantages for the actual forward formulation, since standard inversion techniques could be applied. The disadvantage is that matrix P is very large, of order (IxJ) x (KxL), whereas G was of order IxK and F was of order LxJ. Hence, the problem would be more expensive. For example, for the the coarsely parameterised model for the shear-wave data, P would be 990 by 350! For the similar P-wave model, the matrix would have over a million elements! Notice for an N by N matrix, memory requirement is $O(N^2)$, while expense is $O(N^3)$. Hence, limitation of solving the problem as outlined above is processor speed.

REFERENCES

Engdahl, E. R., and D. Gubbins, Simultaneous travel-time inversion for earthquake location and subduction zone structure in the central Aleutian Islands, *J. Geophys. Res.*, **92**, 13855-13862, 1988.

Menke, W., *Geophysical Data Analysis: Discrete Inverse Theory*, Revised edition, Academic Press, San Diego, 1989.

Pavlis, G. L., and J. R. Booker, The mixed discrete-continuous inverse problem: application to the simultanieous determination of earthquake hypocenters and velocity structure, *J. Geophys. Res.*, **85**, 4801-4810, 1980.

PART 2

Physical Model of Source Region of Subduction Zone Volcanism

Abstract

The thermal structure of a generic subduction zone (SZ) is investigated to elucidate the source region of subduction zone volcanics (SZV). The steady-state thermal field is evaluated for a model SZ where the plates are prescribed by kinematic boundary conditions. The resulting model suggests that the slab is not melted extensively, if at all, and hence is not the source of SZV. The temperature field is high enough to produce dehydration of the slab and melting at both the water-saturated and amphibole buffered peridotite solidi in the mantle wedge. It is proposed that due to the reaction of water with peridotite, the net transport of water can be horizontal even though its motion when free might be vertical. Provided the relative convergence velocity is $>6(\pm 2)$ cm/yr, the mantle wedge will be hot enough at the limit of the lateral water transport mechanism to generate melting at the amphibole buffered solidus. It is conceived that the melt migrates to the base of the overriding lithosphere by means of cracks. The direction of propagation of such cracks is controlled by the direction of the least compressive stress; for a corner flow regime this direction transports the melt toward the wedge corner. The nearly constant distance of the volcanic front above the seismogenic zone can be explained by the primarily pressure controlled dehydration reaction, the short horizontal distance that the water travels before inducing melting and the focusing of cracks toward the wedge corner. The model was then extended to include the effect of localized sources of buoyancy (melt, residue, etc.) as a stationary body force. Best estimates

of local buoyancy sources in a constant viscosity wedge suggest that it is difficult to generate appreciable flow reversal; but due to shear heating, presence of water and dehydration/hydration reactions the slab and wedge might be locally decoupled, allowing a local flow reversal. Such a reversal of the mantle wedge flow, if it persists, leads to cooling of the mantle wedge. Hence, flow reversal cannot be a steady state mechanism. Instead we might expect the flow reversal of the mantle wedge to lead to a modulation of the wedge flow and of the water induced melting at a time scale of O(1 Ma). The time dependence of this model is reinforced by the need to clear the wedge of infertile material. Trying to reconcile the difference between estimates of the amount of water subducted and the amount returned to the crust in melts, leads us to propose that large volumes of hydrated melts are trapped in the lithospheric mantle. Later, these might become the precursors for slow diapiric uprise of tonalitic and granitic plutons.

2.1 Introduction

There have been many attempts to derive the thermal structure of a subduction zone (SZ) since the advent of plate tectonics. These models have had many motives; explanation of depth of seismicity [*McKenzie*, 1969], stress generation for seismicity [*Goto et al.*, 1985], back arc basins [*Andrews and Sleep*, 1974], andesite volcanics and volcanic front heat flow [*Hsui et al.*, 1983;*Honda*, 1985], focal sphere travel time residuals [*Creager and Jordan*, 1986], SZ mechanics [*Kawakatsu*, 1986], dip of SZ [*Furlong et al.*, 1982], and inverted metamorphic gradients [*Cloos*, 1985]. This paper concentrates on elucidating the thermal regime relevant to the identification of the source of Subduction Zone Magmas (SZM).

This paper frequently uses a few generic terms to describe the geometry of the subduction zone. To introduce the reader to them and prevent confusion the idealized geometry is presented in figure 2.1, labeled with these terms.

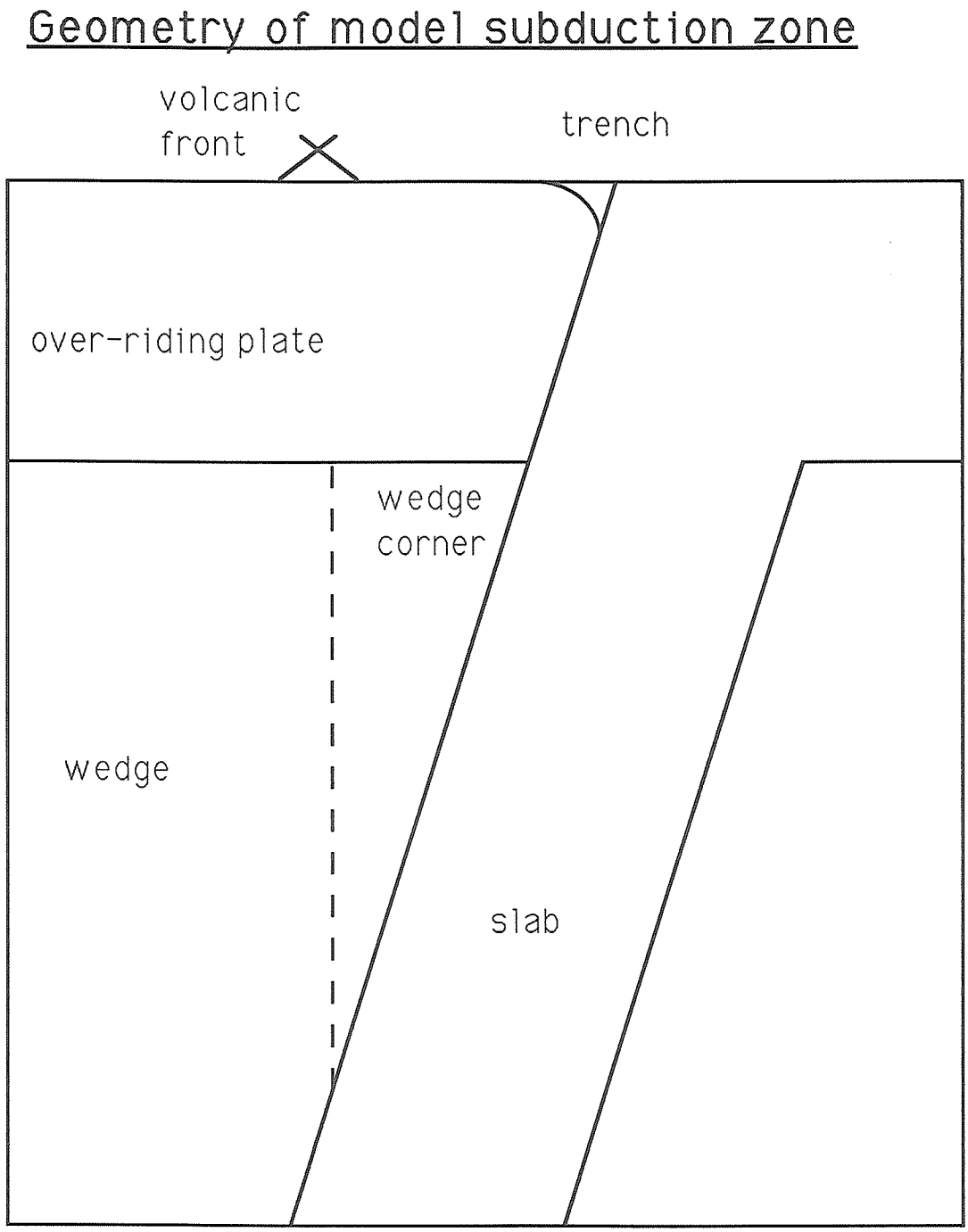


Figure 2.1 Cartoon of terms describing geometry of model subduction zone. Note that by over-riding plate we mean the mechanical lithosphere. When referring to a point, the wedge corner means the apex at which the over-riding plate and slab meet; while when referring to volume, we mean roughly the volume of the wedge below the line joining the trench and volcanic front.

Petrologists and geochemists have advanced numerous hypotheses regarding the source region of SZM. *Marsh* [1979] argues for the slab to be the source since melting experiments on eclogite produce melts similar to the high-alumina arc basalts. He claims that the REE distribution is satisfied provided that the garnet melts nearly completely (degree of melting $\approx 55\%$). These and further arguments are discussed in the papers of *Brophy and Marsh* [1986] and *Myers et al.* [1985]. This viewpoint has also been advanced by *Maaloe* and *Petersen* [1981].

Potentially, the most convincing case against the mantle wedge being the source is that the possibly primary high-alumina basalts do not have olivine or orthopyroxene at their liquidus above 15 kbar - 20 kbar, but rather garnet and clinopyroxene, suggesting an eclogitic as opposed to a peridotitic source [*Johnston*, 1986]. Others have argued that the olivine and orthopyroxene could be in a reaction relationship with the basalt at these pressures [*Hess,* 1989]; as suggested by the work of *Yoder* [1976]. Of course one needs to identify the primary magma and this is also a bone of contention. Is it the high-Al basalt or is it the high magnesium basalt [*Myers et al.*, 1985]? Depending on the source perceived, the appropriate primary magma is chosen (wedge, high-Mg basalt; slab high-Al basalt) and a mechanism developed to explain the appearance of the other basalt. *Myers et al.* [1985] conceive that the high-Al basalts from the slab are the primary magmas while the high-magnesium basalts are the result of contamination by melting of a path through the mantle wedge. Others consider the high-magnesium basalts to be the primary magmas arising from a source in the mantle wedge; in this scenario the high-Al basalts arise from extensive mafic fractionation of the primary basalts [*Perfit et al.*, 1980; *Kay*, 1980; *DeBari and Coleman*, 1989] or the accumulation of plagioclase phenocrysts [*Crawford et al.*, 1987; *Brophy*, 1989].

Workers who argue for the wedge to be the primary source; also involve the incorporation of components from the slab and sediments. These models frequently allow additional assimilation of sub-volcanic front crust [*Hildreth and Moorbath*, 1988]. The relative contribution from the different sources are hard to quantify and from *DeBari et al.* [1987] there is a suggestion that for different elements the contribution of different sources varies, suggesting preferential leaching of different elements by a mobile phase.

Even among the supporters of the wedge source there is a controversy as to what type of source exists. They compare it to either the MORB or OIB source; the other two major oceanic basalt types.

Morris and Hart [1983] argued from the similarity of Sr, Nd and Pb isotopic histograms and LILE abundance ratios that Aleutian lavas and OIB lavas had the same mantle source. Following a similar suggestion by *Saunders et al.* [1980] they proposed to explain the LREE (La) depletion by invoking the survival of a Titanium rich mineral in the source residue, e.g., perovskite, rutile, sphene or ilmenite.

In contrast *Perfit et al.* [1980] argue that major elements of primitive SZM and MORB are alike and differ from OIB, and hence that the source is MORB-like. They claim that evolved MORB and SZM are different since the fractionation is dominated by different phases due to the different depth of crystallization and different water activity. They accept that the SZM are enriched in LIL, volatiles, and Cl. They suggest that this is due to a hydrous phase in the source. *Green and Pearson* [1986] have suggested that it is difficult for Ti to be controlled by residual phases in the SZ mantle wedge, as proposed by *Saunders et al.* [1980].

These two different sources could reflect different degrees of melting. Hence different thermal models might be suggested for different sources. The MORB source is a more depleted source while the OIB source is more enriched and

possibly is more fertile, having a lower solidus. Little petrological work has been done to quantify experimentally the differences that these minor element variations make in the melting behavior. It is widely believed that source rocks rich in alkalis have lower solidi and this might extend to other incompatible elements. Experimental work, however suggests that there is little difference in the solidus temperature for the range of fertility commonly considered, [e.g., *Falloon et al.*, 1987].

2.1.a Thermal Models

The earliest thermal models of subduction zones [e.g., *Griggs,* 1972; *McKenzie,* 1969] had the explanation of deep seismicity as their focus and solved the zeroth order problem of the sinking of a cold slab into a hot constant temperature mantle. These thermal models did not address the thermal interaction of the slab on the wedge, or the mechanical interaction. Most have attempted to model the heat flow and one can obtain different thermal regimes depending on the assumptions one makes regarding how the system behaves; witness the wide range of thermal fields derived over the years given different assumptions. This seems to suggest that there are no constraints on the thermal models, but this is too pessimistic. In fact many of the assumptions of the old thermal models are accepted to be unreasonable or unsuitable and in this light the difference between models with reasonable assumptions are actually small; sufficiently small to allow reasonable predictions from phase diagrams. For example, *Hasebe et al.* [1970] had viscous shear heating as a major heat source while they credited the enhanced heat flow at the volcanic front to advected heat by magma flow; both viscous heating and advection by magma are argued below to be of secondary importance, and not relevant in obtaining the first order model. Another example is the purely viscous model of *Bodri and Bodri* [1978], which allowed the induced wedge

flow to lead to extreme ablation of the wedge corner; their extreme ablation is unreasonable but otherwise the general principle of ablation is well supported and their thermal model is not too different from other thermal models including induced flow and no major heat sources. We shall now begin by reviewing previous models to illustrate what was learned, and how, when one assesses them critically, one arrives at the conclusion that the thermal field of a subduction zone now more constrained than assumed in previous syntheses (e.g., *Wyllie* [1988]) .

The earliest suggestion was that the subducting plate melted due to the high frictional shear stresses on the fault between the two plates [*Turcotte and Schubert,*1973] or the high viscous dissipation due to the shearing of a fluid plate-wedge interface [*Turcotte and Schubert,*1968]. Their thermal models assumed that this process was the explanation for the volcanics and hence set the stresses at the level required to melt the plate. This required large stresses on the order of several kilobars. The models of *Toksöz et al.*, [1971] have been widely cited primarily due to their inclusion of many heat sources. Frictional heating was also a significant factor in these models, although they ignored viscous coupling of the slab and the wedge. Following the work of *Yuen et al.* [1978] and *Yuen and Schubert* [1979] high shear stresses were shown to be unlikely. *Beukel and Wortel* [1987] have shown that given reasonable rheology and heat flow constraints, shear stresses are likely to be about 250 bars, extend down to about 40 km, and locally lead to temperature increases of no more than 250°C. *Anderson et al.* [1976] argued that endothermic dehydration reactions in the plate would absorb most if not all the heat produced in the frictional heating in the shear zone. Therefore frictional heating has been shown not to be a major heat source and thermal models dominated by this premise are dubious. This leads to difficulties for models suggesting that the source is an extensively melted slab.

Following the demise of frictional heating, the next candidate for a major heat source has been advected heat, carried by an induced flow generated in the mantle wedge by viscous coupling to the subducting plate [e.g. *McKenzie*, 1969; *Andrews and Sleep*, 1974; *Bodri and Bodri*, 1978; *Anderson et al.*, 1980; *Hsui et al.*, 1983; *Honda*, 1985]. The resulting flow under certain circumstances can bring quite high temperatures close to the subducting plate. *Hsui et al.* [1983] claimed that it is possible for induced flow to raise the temperature in the wedge corner sufficiently to melt the subducting plate. *Honda* [1985] developed a detailed model for the Tohoku subduction zone, in Japan, incorporating an induced flow in the mantle wedge. The model included other heat sources and sinks and was constrained primarily by heat flow observations. He needed to propose relatively high shear stress on the interface to satisfy the surface heat flow. Recent studies have shown that there is substantial continental generation of heat in Japan and with its inclusion one can satisfy the surface heat-flow with low shear stresses on the thrust zone [*Nagao and Uyeda* 1989; *Furukawa and Uyeda* 1989]. Petrological and seismological constraints were also satisfied.

The calculation of *Hsui et al.* [1983] could not be conclusive for three reasons. First, a thermal boundary condition needed to be prescribed at the slab mantle interface. In one case the temperature at this interface was prescribed and hence couldn't be evaluated while in the second case a zero heat flux condition was applied at the interface, which is obviously incorrect, and only an approximation as *Hsui et al.* [1983] discuss. Second, (as they also discuss) their grid spacing is insufficient to resolve the actual temperature at the interface and in the neighboring mantle wedge. Third, the geometry is arguably artificial and directly leads to the potential melting being located $\approx$ 100 km below the volcanic front; large shear stresses would develop in the corner leading to ablation and a more reasonable geometry would most probably not have such an angular form. The resultant

"corner" would be expected at a much shallower depth and hence their explanation of the depth of the slab below the volcanic front wouldn't hold. *Honda* [1985] addressed many of the above problems but posed the answer in the form of a question, i.e., the mantle temperature below the Japan Sea is 1400°C, and how is the Japan Sea this hot? Honda couldn't answer this question since the area he modeled is quite small, less than 200 km in depth.

This study was undertaken so as to avoid the limitations of the previous studies outlined above. Induced flow needs to be considered with a more detailed grid such that the resolution problems of *Hsui et al.* [1983] and the extent problems of *Honda* [1985] are avoided. Also the interface boundary conditions are avoided by placing the slab in the middle of a grid so that the side boundary conditions do not involve the slab.

The fundamental paradox of SZV is the presence of melt and high heat flow adjacent to an enormous heat sink, the cold subducting slab; see *Uyeda* [1982]. In reviewing the previous thermal models we have touched on some of the different proposals for the source region of subduction zone volcanics. The proposal with most support to date produces melting of the subduction zone wedge by fluxing it with volatiles derived from the slab which lower the solidus, see *Gill* [1981]. *Marsh* [1979] has been a strong advocate that the slab melts and he achieves this by having extreme induced flow melt the wedge in a "magic corner". This mechanism was investigated numerically in *Hsui et al.* [1983], and was reiterated in *Toksoz and Hsui* [1987], but as discussed above was not definitive. Others have proposed that the melting arises due to adiabatic decompression of the wedge on upwelling. Proponents of this mechanism and variants thereof include *Tatsumi et al.* [1983], *Nye and Reid* [1986] and *Plank and Langmuir* [1988]. The arguments have arisen from petrologic experiments suggesting high equilibrium temperatures in the wedge and correlation of major element chemistry with crustal thickness.

Ida [1983,1987] suggested that the proposal of *Tatsumi et al.* [1983], of diapirs small enough to be heated by conduction but large enough to rise faster than the induced slab flow, was unrealistic; hence he argued that the reverse flow wasn't individual diapirs but rather was a continuous updraft; his source is the subducting slab. *DeBari et al.* [1987] also proposed a continuous updraft rather than individual diapirs, though their source is in the mantle wedge rather than the slab. There are also proposals of an intermediate nature such as *Nicholls and Ringwood* [1973] which involve diapirs rising from the slab, hybridizing the wedge and forming a garnet pyroxenite, which subsequently gets extensively melted.

2.1.b Analytic Solution

Before describing the numerical thermal models we would like to suggest analytic approximations of various aspects of the problem to give the reader a qualitative feeling for the behavior of the model. Since the plate is forced to subduct by imposing kinematic boundary conditions, we are left with three regimes: one in which we have only conduction (the rigid slabs), the second dominated by advection (the mantle wedge) and third, boundary layers where advection balances conduction. The boundary layers extend from the slab out into the wedge until the velocity of the mantle wedge perpendicular to the isotherms is faster than the characteristic "velocity" of conduction $\approx \kappa/d$. Here κ is the heat diffusivity and d is the conduction length scale. This comparison can be evaluated in the fluid dynamic jargon that the Peclet number $Pe = v\ d/\kappa > 1$. If $v = 1$cm/yr and $Pe \approx 1$ then $d \approx 3$km, while if $v = 6$cm/yr then for $Pe \approx 1$, $d \approx 0.5$km. The geometry of the flow field is defined by the boundary conditions.

The subducting slab is assumed to drag down mantle at its sides. By continuity the mantle below the overriding lithosphere must approach the wedge corner to replace the material descending with the slab. Hence the slab induces a

regional flow. As this mantle descends with the slab, it heats the slab, and is correspondingly cooled. The extent to which this cooling extends into the mantle increases with depth into the mantle. This is a boundary layer and its thickness is about $\sqrt{\kappa t}$, or equivalently $\sqrt{\kappa x/v_p}$; where t is time, x is distance down along the slab and v_p is the plate velocity. The boundary layer is thinnest within 10 km of the wedge corner. We shall refer to this as the boundary layer's initial thickness. An estimate of the initial thickness ($\approx$5km) and a more complete analytic solution suggested to us by A. Howard and B. Hager (personal communication 1985), are presented in appendix 2.A. This more complete analytic solution shows us that the thickness of the boundary layer is inversely proportional to the plate width and demonstrates the $\sqrt{\kappa x/v_p}$ behavior.

It is found that the temperature at the interface of two different temperature blocks remains virtually constant at the mean temperature of the two blocks as they equilibrate [*Carslaw and Jaeger*,p.55,1959]. It can be expected that the temperature at the slab-mantle interface will be approximately constant, an average of the mantle temperature away from the slab and the coldest temperature in the slab.

2.2 Method

Before introducing the method we will briefly discuss the framework envisaged for the model. Since the initiation of subduction and subsequent dynamic modeling of its evolution is in its infancy [*Gurnis* 1989], we have used the more robust kinematic modeling of the flow of the slab and the surrounding wedge. Since we are restricting our attention to potential source regions for SZV, we only attempt to model accurately the thermal field of the subduction zone close to the wedge corner down to a depth of around 150 km. The model outside this region does not require great care. Provided the region of interest is largely

insensitive to the exact choices at the model periphery, we do not care which choice is made. In particular we make no account for variations in the large global flow away from the slab, the details of the fate of the slab, or local sources of thermal buoyancy. For example we make no account of possible thickening [*Vidale*1987], slowing [*Vassiliou et al.* 1984], spreading on [*Giardini and Woodhouse* 1984], penetration of the 670 km discontinuity by the slab [*Creager and Jordan* 1986] or non-penetration [*Zhou et al.* 1988]. Their effect on the flow within 200 km of the corner will be insignificant. The insensitivity of the thermal field to this type of variation was demonstrated for a spreading and penetrating slab by running two 700 km depth box models.

At a larger distance from the corner and over larger time scales the global convection is significant. Suggestions for this can be seen in the 3D, internally heated, convection calculations of *Bercovici et al.* [1989] (spherical geometry) and *Houseman* [1988] (cartesian geometry). They find the upflow to be pervasive and the downflow to be in the form of linear, slab-like features. From their figures, it can be seen that the hottest regions in the upper part of their models 'halo' the cold downwellings. The 'halo' is virtually stationary, because its intrinsic buoyancy is counteracted by coupling to the cold downwelling nearby. This material heats up, since its internal heat generation is greater than the heat it can conduct away (since it is largely stationary, there is no advection). This could be a partial explanation for the formation of back-arc basins. Similarly, on long time scales, the global flow might be sufficiently chaotic to clear the wedge of infertile material. Also, the global flow is critical in adjusting the local deformation and the global plate velocities to give approximately constant local convergence velocities ≈ 7 cm/yr [*Otsuki* 1989].

The temperature field was modeled using a finite element convection code. The grid used for the 60° dipping slab, together with the location of the plates

and the thermal boundary conditions is shown in figure 2.2. A similar but even finer resolution grid was used for the 30° dipping slab. Using such a grid had two advantages over a finite-difference grid. First, it was easy to implement the variable cell size allowing cells to range from $\approx$ 5 km across to 80 km across, (grid of 60° model) giving us the ability to combine resolution in the region of interest together with distant boundary conditions such that the region of interest was insensitive to the particular boundary conditions. Second, it allowed us to have the slab interface be flat rather than stepped as it would be with a rectangular grid.

The following equations were solved :

The equation for the conservation of mass (i.e., equation of continuity),

$$\nabla \cdot \vec{v} = 0 \tag{2.1}$$

where $\vec{v}$ is velocity.

The equation for the conservation of momentum (i.e., Stokes Equation)

$$\nabla P = \nabla \cdot (\eta \nabla \vec{v}) + \Delta \rho g \hat{z} \tag{2.2}$$

where P=p-ρgz, where p is the total pressure and z is the depth below the surface, η is viscosity, $\Delta\rho$ is deviation in density, g is the acceleration of gravity and $\hat{z}$ is the unit vector in the downward radial direction. The above two equations were solved by the penalty parameter formulation such that equation (2.1) was automatically satisfied to 10^{-6}, and no boundary conditions are required on the pressure. For details see *Hughes et al.* [1979]. The equation for conservation of energy (i.e., the heat conduction-advection equation) is

$$\partial T/\partial t + \vec{v} \cdot \nabla T = \kappa \nabla^2 T \tag{2.3}$$

where T is temperature, t is time, κ is thermal diffusivity.

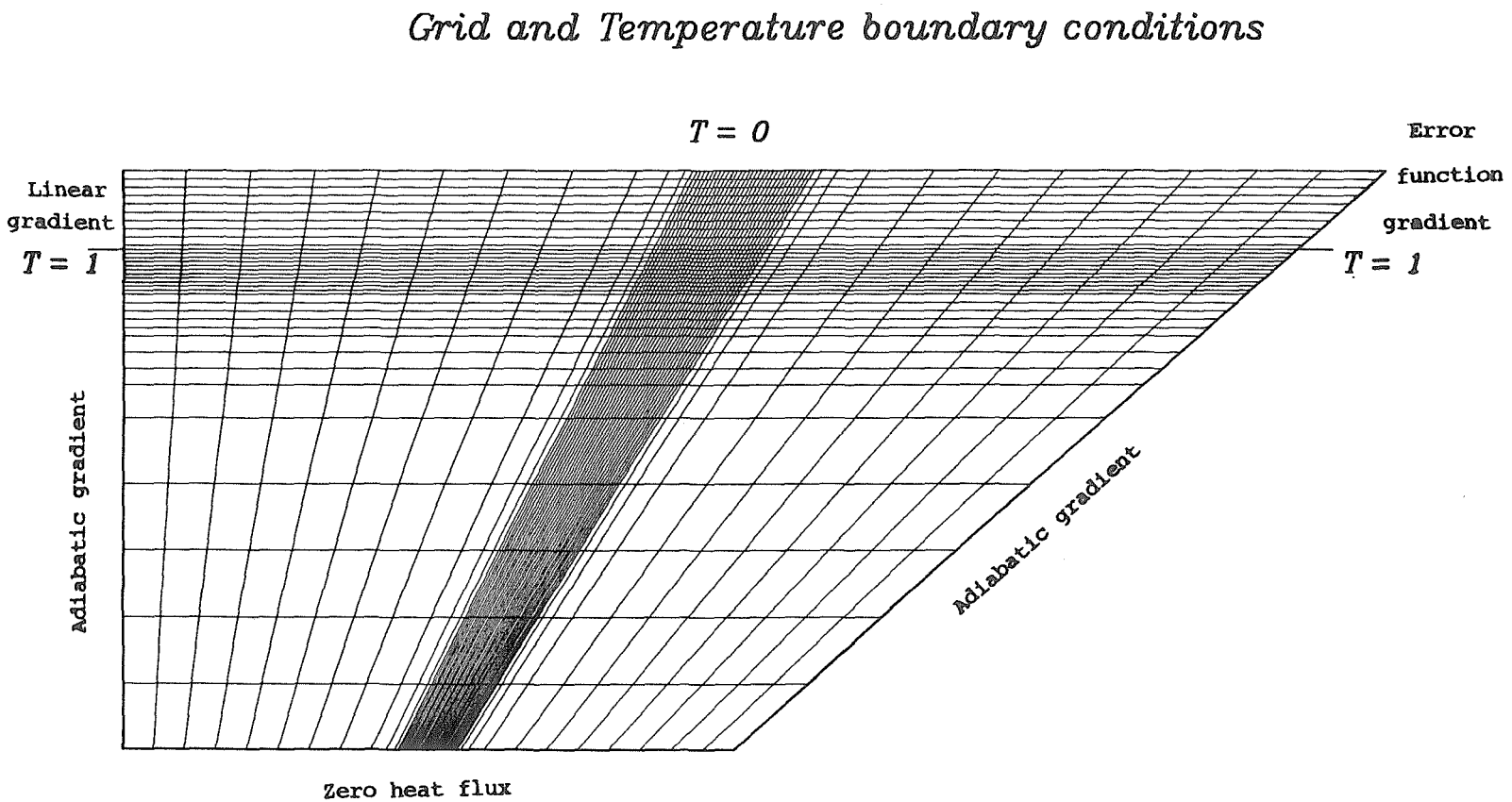


Figure 2.2 Temperature boundary conditions for 60^{o} slab. Note the fine grid near the region of interest to achieve better resolution. The marks on the two sides correspond to the thickness of the thermal lithosphere.

We have used the Boussinesq approximation [*Tritton* 1977], which ignores all variations in density other than as a body force in the Stokes Equation. One of the conditions for the applicability of the Boussinesq approximation is that the depth of the model is less than the pressure scale height; for the whole mantle this is a problem but is not for a model 400 km deep. This leads to the continuity equation reducing to a requirement for incompressibility, while the absolute value of the density cancels out of the Stokes equation. This is an approximation where we ignore a term $(-\alpha TDP/Dt)/\rho C_p$ on the left-hand side of equation (2.3). This is small compared to the other term provided $g\alpha dT_o/C_p\Delta T<1$; it is actually $\approx$ 0.15; where d is the depth of the system, T_o is a reference temperature in the system (arbitrary, 1075°C) , and ΔT (1000°C) is the temperature difference across the system. This term can be thought of as an unaccounted heat source/sink, due to adiabatic decompression. This term could be accounted for by using the potential temperature, $\Theta = [T - (T_a - T_o)]$, where T_a is the adiabatic temperature defined such that $dT_a/dz = g\alpha T_o/C_p$. We have also assumed that C_p, κ and α are constant. This allows us to take κ out from $\nabla\cdot(\kappa\nabla T)$ to give $\kappa\nabla^2 T$.

We used real rather than potential temperatures; this allowed us to compare the results directly with phase diagrams. The thermal models need be no more accurate than the phase diagrams to which they are compared. In light of this, using real temperatures leads to errors small in comparison to the uncertainty and disagreement in locating the phase boundaries; this is also true of the other approximations mentioned.

Using the robust kinematic boundary conditions, it was discovered that variable or constant viscosity made little difference to the flow produced near the wedge corner, our region of interest. This can perhaps be understood because using kinematic boundary conditions requires the slab to behave as if it has infinite viscosity. In comparison, any other viscosity contrast is insignificant. As a

result constant viscosity was used. Temperature dependent viscosity is potentially important in the Earth, but this will only be effectively implemented when kinematic boundary conditions are replaced by a dynamic system. Similarly, Newtonian viscosity was used as opposed to power law rheology, even though the high stresses expected suggest that the rheology is most probably non-linear, at least near the corner. For a discussion of power law rheology corner flow see *Tovish et al.* [1978]. Note if the plate is mechanically decoupled from the mantle wedge the kinematic boundary conditions no longer control the flow far away from the slab, hence in this case the rheology of the wedge will be important if there are local sources of buoyancy.

Since in our formulation the density and viscosity are independent of temperature, the solution of the flow and temperature field are decoupled. The flow need only be solved once and only the energy equation needs to be advanced through time, leading to computational savings. Hence we have forced convection rather than free convection. Using kinematic boundary conditions makes it difficult to evaluate the proportion of the global thermal buoyancy that has been accounted for with the kinematic velocity boundary conditions. It is expected that the flow resulting from the smooth large scale temperature variations are well accounted for by the boundary conditions, while the temperature variations whose flow generation are unaccounted for, will lead to only minor, local changes to the flow. To accurately include the effects of thermal buoyancy one will need a dynamic model with temperature dependent viscosity. By contrast, it is felt that large local compositional differences (e.g., due to the presence of melt and residue) can lead to substantial local flow variations, especially away from the forced flow of the slab boundary conditions. These are investigated in a later section using locally prescribed body forces.

These equations are non-dimensionalized as follows $x' = x/d$, $t' = \kappa t/d^2$, $P = \eta\kappa p'/d^2$. The primes correspond to non-dimensionalized quantities. Dropping the primes, we get the following non-dimensionalized equations

$$\nabla P = \nabla^2 \vec{v} + B_m \hat{z} \tag{2.4}$$

$$\partial T/\partial t + \vec{v} \cdot \nabla T = \nabla^2 T \tag{2.5}$$

where $B_m = d^3 \Delta\rho g/\kappa\eta$ is the buoyancy number, $\Delta\rho$ is the body force acting per unit volume, and a buoyancy number can be defined for each element in the grid.

The temperature boundary conditions are also illustrated in figure 2.2. The thermal lithosphere of the overriding plate was made 100 km thick and was characterized by a linear temperature distribution. The thermal lithosphere of the subducting plate was made of variable thickness varying from 20 to 120 km thick and was characterized by an error temperature distribution. All the figures are from models with oceanic lithospheres that are $\approx$ 90 km thick. The sides of the box below the thermal lithosphere were given a linear temperature profile corresponding to an adiabatic gradient increasing from T=1 at the base of the thermal lithosphere to T=1.14 at the bottom of the box (modeled to be 400 km). The thermal boundary condition at the bottom was one of zero heat flux.

The velocity boundary conditions are illustrated in figure 2.3. Note that the boundary conditions on the edges of the wedge are zero normal stress, implying that the derivative of the velocity in the direction normal to the boundary is zero. This gives a velocity field very similar to the corner flow solution found by *Batchelor* [1967] for an infinite wedge. Note the descending slab is prescribed by the kinematic boundary conditions as a non-migrating slab dipping uniformly at 60° or 30°. We modeled different local convergence velocities. For the moment we shall assume that only relative velocities are important. We ignore global flow, trench migration and internal plate deformation, (e.g., arc compression or back

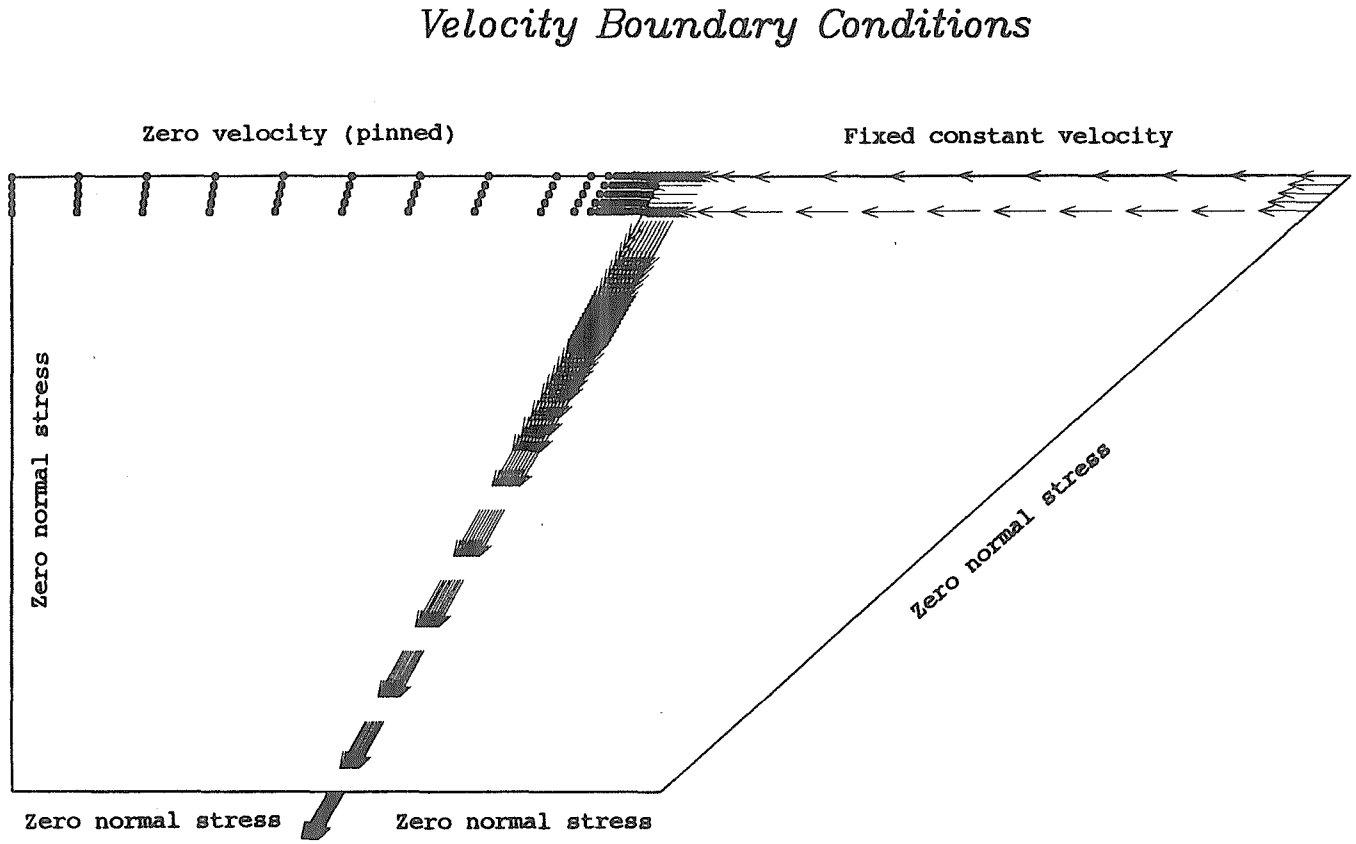


Figure 2.3 Velocity boundary conditions. The circles correspond to fixed zero velocity boundary conditions. Note the unconstrained side boundaries have zero normal stress boundary conditions.

arc spreading), and only consider the resulting local convergence velocity. *Wdowinski et al.* [1989] have modeled the deformation in the overriding plate as the response of a viscous mechanical lithosphere to an applied horizontal force and shear tractions from the underlying corner flow. The resulting modification of the corner flow is weak.

It was demonstrated that the boundaries of the model were far enough away by running experiments with altered boundary conditions. In one experiment the velocity boundary conditions along the side were made vertical and horizontal rather than the stress free boundary conditions. In another the zero heat flux boundary condition at the base of the box was replaced by a constant temperature boundary condition. It was found that neither made a difference to the temperature of the slab or wedge corner down to 200 km depth. As a corollary one should not interpret the thermal field close to the base or sides of the box.

The issue of slab shape is not fully addressed here, but because the flow is laminar and most of the heat flow is by advection, the exact shape is not critical to the global temperature field but would be important for temperature locally in regions where the slab shape is curved. *Spence* [1989] shows that the slab dip below $\approx$ 30 km is approximately uniform through 100 km depth to $\approx$ 200 km depth at least. *Barazangi and Isaacks* [1976] find that dips vary down dip but are largely constant at 100 km. Down to $\approx$ 30 km depths the dip is shallow $\approx 10^{\circ}$; this leads to more heating of the downgoing lithosphere compared to our model and equally a stronger cooling of the overriding lithosphere. Since the overriding lithosphere is cold at these depths and is made ever colder, this effect is minimal. This is illustrated in the work of *Sydora et al.* [1978]. The effect of geometry of subduction path was also considered by *Furlong et al.* [1982], but only the thermal field of the slab was evaluated; the wedge was effectively a stationary hot bath like the zeroth order model of *McKenzie* [1969]. The continental plate in our

model was stationary while the downgoing plate had a constant velocity, varying from 1.8 - 9 cm/yr. In estimating what real velocity corresponds to the non-dimensional model velocity, one needs to know the applicable diffusivity of the model. It has been found that thermal conductivity decreases appreciably with temperature but increases with pressure; as pressure increases radiative transfer makes an increasing contribution. *Fujisawa et al.* [1968] found that the diffusivity of forsterite at 30 kbar and 825°C was 9.4 $10^{-7}m^2s^{-1}$, and 10.3 $10^{-7}m^2s^{-1}$ for 725°C. Density increases slowly with depth, while above the Debye temperature, the specific heat capacity is largely constant. The relevant diffusivity is that of the slab and mantle wedge in the thermal boundary layer where conduction is important. Since it is cold here we estimate that the net result is a thermal diffusivity largely unchanged compared to the value at the surface. We use $\kappa = 10^{-6}m^2s^{-1}$. The uncertainty in the estimate of κ translates into a corresponding uncertainty in the estimate of the real plate velocity. We estimate that the applicable κ could range from 0.8-1.3 $10^{-6}m^2s^{-1}$. This translates to percentage uncertainties in modeled velocities of up to 30%.

The mechanical lithosphere is taken to be 40 km thick. This is arbitrary but our choice was motivated by the following 5 lines of reasoning. Given that the timescale of this process is on the order of 1Ma (>3 Ma and <10 Ma from initiation of subduction to generation of magma [*Gill*,1981, p29-32]) then for a Maxwell body this suggests that the relevant transition viscosity from elastic to viscous behavior is 10^{24} Pas. Given our high stress regime this suggests a critical isotherm of 700°C. For old overriding lithospheres it is not unreasonable for this geotherm to be at 40 km. Note that for midplate oceanic lithosphere at low strain rates, (i.e. long timescales and low stress regime), *Bergman and Solomon* [1984] suggest from seismicity that the critical isotherm is at $\approx$ 800°C while *Wiens and Stein* [1983] suggest 700°C for old oceanic lithosphere. Second, given the expected high

temperatures from the high heat flow and the relatively high stresses due to the corner flow, one expects this region to undergo some degree of ablation [*Andrews and Sleep* ,1974; *Bodri and Bodri*,1978]. In the case of oceanic lithosphere crossing a hotspot, it is found that the lithosphere thins rapidly to the thickness of 24Ma old oceanic lithosphere, $\approx$ 45km [*Crough*,1978; *Detrick and Crough*,1978;*Crough and Thompson*,1976]. This thickness seems to be independent of the original plate age, the velocity of the moving plate or the strength of the hot spot involved. The rate of thinning requires delamination or ablation, while the constant resulting thickness requires that it is controlled by material rheological properties that are independent of temperature. Third, the flexural thickness of oceanic lithosphere below island arcs ranges from 10-35 km [*Bodine et al.*,1981]. It is unlikely that the mechanical thickness is less than the flexural thickness. Fourth, island arc crust ranges from 5-35 km and hence suggests that crust is stable down to at least this depth. Our model has a problem when attempting to model the Andes. Here the crust is estimated to be up to 70 km thick [*James*,1971] and either (i) the wedge corner is full of crustal material, (ii) delaminated crust is replaced by magma production, (iii) these low seismic velocities could correspond to regions of high melt porosity in mantle rocks rather than real crust, or (iv) the rheological thickness is actually at least 70 km thick in this region. Fifth, the transition from elastic to ductile behavior ends at a depth of 60 km in the thrust zone (as defined by the extent of the rupture zone of thrust earthquakes) [*Astiz et al.*,1988]; this can be considered an upper bound. More direct estimates of the extent of rupture in the Andes by *Tichelaar and Ruff* [1989] give depth ranges of 40-50 km. From the combination of the above arguments we believe that a thickness of 40 km is most probably a reasonable choice, if such a thing as a reasonable choice exists for modeling a heterogeneous set. All workers interested in proposing a reversal in flow direction have required the mechanical lithosphere to be, at most, 50 km

thick [*Plank and Langmuir*, 1988; *Ida*, 1983,1987; *Nye and Reid*,1986]. We note that *Honda* [1985] argued from heat flow constraints and the lack of seismicity observed in the mantle that the mechanical lithosphere in Japan was the same thickness as the crust, i.e., 30 km thick. Similarly *Thatcher et al.* [1980] modeled the relaxation following the Riku-u earthquake of 1896 using a 30 km thick elastic plate on a viscoelastic asthenosphere of viscosity of 10^{19}Pas. The thickness of the rheological boundary in our model was kept constant away from the corner since the material properties controlling ablation and delamination are poorly understood. Experiments were made with thickening mechanical lithospheres away from the wedge and also appreciable but local ablation in the wedge corner. It was discovered that the resulting changes in the thermal field were minimal in potential source regions. Since no thermal buoyancy is evaluated other than that inherent in the kinematic boundary condition, the possible effect of sidewall cooling in a narrow rift below the volcanic arc is not allowed to develop, hence the effects of *Mutter et al.* [1988] were not observed. These might be important locally.

In the thermal model there are no heat sources or heat sinks. Potential heat sources are frictional heating on a thrust zone, viscous dissipation in the mantle wedge, hydration of the mantle, radioactivity and phase changes. Potential heat sinks include dehydration of the slab, and melting. Estimates of these are made in table 2.1. The most dramatic observation is that conduction of heat into the slab is the biggest single entry by a large margin; hence the first order thermal field will be dominated by its contribution. Given our limited understanding of other components of the model, such as melting, the SZ rheology, and phase diagrams; it is appropriate to ignore these other sources and sinks. We are assuming that the first order thermal field is responsible for the melting rather than some localized phenomenon, e.g., heat generated by the mechanical work done due to the volume change in the basalt-eclogite transformation if it occurred

Table 2.1. HEAT SOURCES

heat source/sink	power involved per m along slab (Wm^{-1})	Comments
conduction into slab	-10^{10}	lithosphere 100km thick
viscous dissipation	$+< 10^{6}$	>10km from corner
frictional heating	$+6\ 10^{3}$	σ 25MPa
melting	-10^{5}	SZV addition $1km^3yr^{-1}$
dehydration	$-5\ 10^{3}$	2wt%H_2O in slab
hydration	$+10^{4}$	
radioactivity in continental crust	$+2\ 10^{4}$	
radioactivity in oceanic crust	$+10^{3}$	
radioactivity in mantle wedge	$+\approx 0$	

very rapidly. Localized phenomena are unlikely to be the explanation since they cannot explain the very high heat flow unless the wedge is sufficiently hot to generate melt independently of the phenomenon. Note that the estimate of dehydration is not too dissimilar to the frictional and viscous dissipation, while the latent heat of melting and heat of hydration are not too dissimilar. Hence some of the sources and sinks that we are ignoring will partially counteract each other.

An effect that is not directly addressed in this model is the advection of heat by fluid components, e.g., H_2O, and melt. Water or melt from the slab is cooler than the wedge while melt from the source region is hotter than the overlying lithosphere, hence fluid advection could heat or cool. Estimates of heat advection by H_2O have been made by *Peacock* [1987] and he shows that one would need very large volumes of water for this to be a major effect. Similarly, *Honda* [1985] made an estimate of the melt flux required to explain the heat flux deficiency that his model had at the arc. Assuming thermal equilibrium between the melt and the surrounding country rock throughout, he required a melt flux of 0.05cm/yr as opposed to erupted fluxes of 0.001cm/yr. This ratio of 20:1 of intrusive to extrusive volume is on the large side of all current estimates. If instead we assume a thermal pipe model such as *Turcotte* [1989], then we find that a melt flux of $0.05 \mathrm{cmyr}^{-1}$ gives an excess heat flux estimate of $60 \mathrm{mWm}^{-2}$, but the heating of the intervening material would be negligible. We suggest that advection of heat plays a minor role in changing the thermal field globally, though it is accepted that channeled flows could have a large effect locally.

Since our calculations are done using actual temperatures rather than potential temperatures, we should account for adiabatic decompression/compression as a potential heat source/sink, as discussed above. The rate of change of heat $dQ/dt \approx g\rho\alpha T v_z$, is $\approx 2\ 10^{-6} \mathrm{Wm}^{-3}$. Hence for a region 100 km thick descending for 400 km the heat source per metre along arc is $\approx 1\ 10^{5} \mathrm{Wm}^{-1}$. Since we are

using adiabats as boundary conditions we are including the global average effect of adiabatic decompression.

Since we have few constraints on the initial conditions of subduction, it was decided to concentrate on the steady state thermal regime of the subduction zone. Hence an implicit convection code was used, allowing faster convergence to steady state. The model approaches steady state in a few Ma. If the wedge was purely rigid, and all the heat flow was by conduction, it would takes hundreds of Ma. Later it was discovered that many of the mechanisms do not work at steady state but could work in a periodic cycle. An explicit time dependent code was then used to model these situations, taking the results of steady state calculations as the initial temperature conditions.

Before comparing the model with phase diagrams, we need to dimensionalize the temperature. An estimate of the temperature below the lithosphere can be made assuming that the potential temperature below the lithosphere is largely constant and that it explains the relatively constant thickness of oceanic crust. *McKenzie and Bickle* [1988], applying this assumption, used experimental melting data and inferred a temperature of 1325°C below a 100 km thick lithosphere. Using an estimate of $0.6°Ckm^{-1}$ for the adiabatic gradient, one gets an estimate of 1500°C at 400 km. The dry solidus at 100 km depth is 1400°C and there is little evidence away from MOR for extensive melting, see *Sato et al.* [1988b,c]. Another reasonable constraint is the conditions for the phase change of the 400 km seismic discontinuity assuming it has been correctly equated with the olivine-spinel phase change. The depth of this discontinuity can only be resolved seismically to $\pm$ 10 km and its Clausius-Clapeyron slope is $4barsKm^{-1}$, hence *Jeanloz and Thompson* [1983] could only estimate a temperature of 1425 $\pm$ 300°C for this phase change. Of course the variability in the depth of this discontinuity could equally be of this order and hence the upper bounds on lateral temperature

variations at this depth are of the order of ± 300°C. Other workers have preferred to place the discontinuity at 1500°C or 1600°C when modeling its seismic jump from mineral physics [*Anderson and Bass*,1986; *Weidner and Ito*,1987]. Hence we model the temperature at 100 km as 1325°C, and 1500°C at 400 km.

2.3 Models and Discussion

We shall consider the effects of two sources of buoyancy on the dynamics. They are the long wavelength component of the temperature variation, which gives the descent of the slab seen in the seismicity, and the buoyancy due to the differentiation, (melt and residue) to form SZM. We shall first include only the slab induced regional flow. This is discussed in section 2.3.a below. Then in section 2.3.b we shall also include the potential effects of buoyancy as a result of melting, leading to local changes in the flow. This buoyancy can potentially give a local reversal in the mantle wedge flow; hence this section 2.3.b is entitled "Flow reversal in wedge."

2.3.a Induced Corner Flow

Using the above methodology a series of experiments was run using models in which the mantle flow was prescribed by the kinematic plate flow and there were no intrinsic sources of buoyancy i.e., $B_m = 0$. Two grids were considered, a 30° and 60° grid. These models were set up to address the question of whether an extreme induced corner flow could melt the slab and if any other regions could be made to melt. For these models the geometry of the streamlines is independent of the magnitude of the slab velocity provided we have the same dip, no thermal or intrinsic buoyancy forces, and a linear constant viscosity. Hence, the velocity plot in figure 2.4 illustrates the flow field for all the 60° models in this section, i.e., the velocity field only varies by a scaling from one model to another.

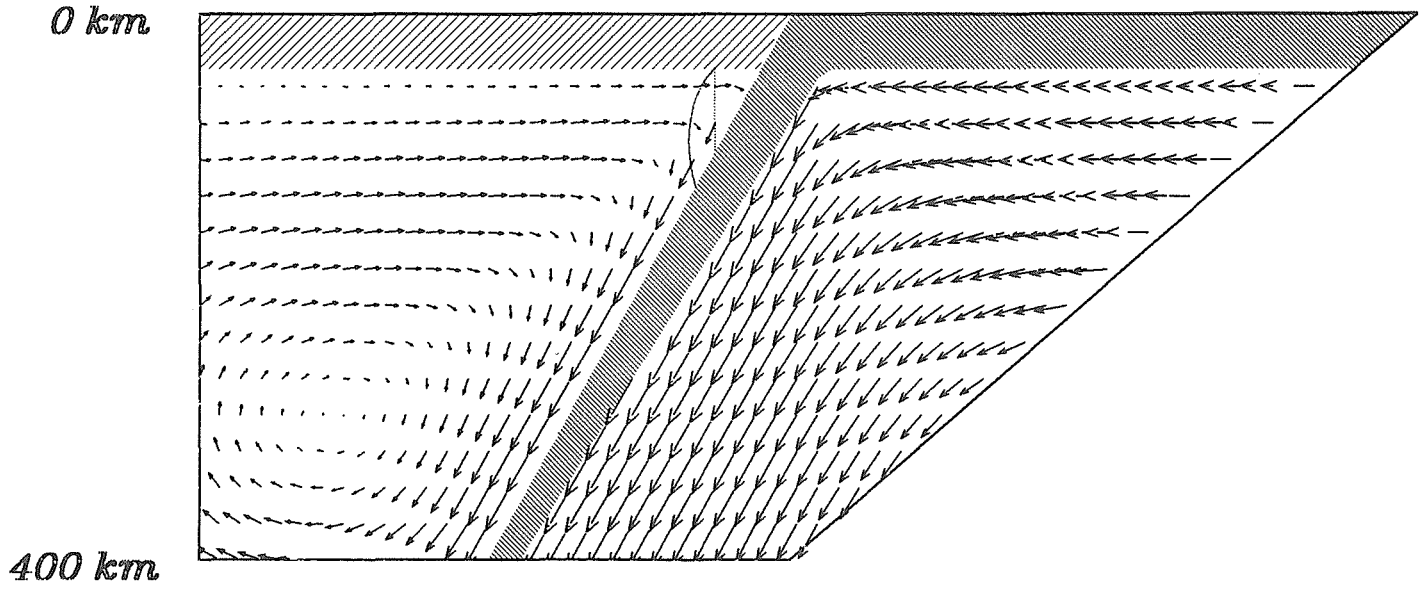


Figure 2.4 Plot of velocity field. The descending part of the slab is shaded. The velocity has been interpolated from the uneven spaced grid to an even grid to make the figure clearer.

The results of the 60° dipping slab at 3 convergence velocities, 1.8, 4.5, and 9 cm yr^{-1} are presented in figure 2.5. The thermal models are sensitive to convergence velocities. It is found that lower convergence velocities imply much cooler mantle wedges with much thicker thermal boundary layers. Similarly the slab interiors are hotter. It is found that for a plate velocity of 1.8cm/yr the wedge corner is not hotter than 800°C!

Varying the other input parameters had the following effects: shallower dips lead to similar thickness thermal boundary layers in the mantle wedge parallel to the slab, bringing these boundary layers closer to the cold upper lithosphere. This tends to squeeze out the hot tongue of mantle wedge that heads towards the apex of the wedge corner, and therefore we get marginally cooler mantle wedges and hotter slab interiors. A thicker mechanical lithosphere leads to the isotherms being pushed deeper, but relative to the mechanical lithosphere they are shallower and the wedge is hotter. Similarly, the interior and surface of the slab are also slightly hotter. Varying the thickness (or equivalently the age) of the subducting oceanic plate had no effect on the temperature in the mantle wedge, while thinner, younger, hotter slabs heated up more and as a consequence had higher temperatures at their interface with the mantle wedge.

In figure 2.6 we illustrate the anhydrous (A), wet (W) and damp (D) solidi in both the basalt (B) and peridotite (P) systems. We also present the dehydration curves of amphiboles in both the peridotite (AmP) and the basalt system (AmB). The damp solidus (amphibole buffered) is above the wet solidus (free water) in the regions where amphibole is stable above the wet solidus. Note that the amphibole buffered solidus (damp solidus, D) is for a rock that is perfectly dry, i.e., damp does not imply wetness but rather the presence of hydrous minerals. The curves are a synthesis of the work of Wyllie and his co-workers, compiled in *Wyllie* [1978], Green and his co-workers compiled in *Green* [1973], and other work

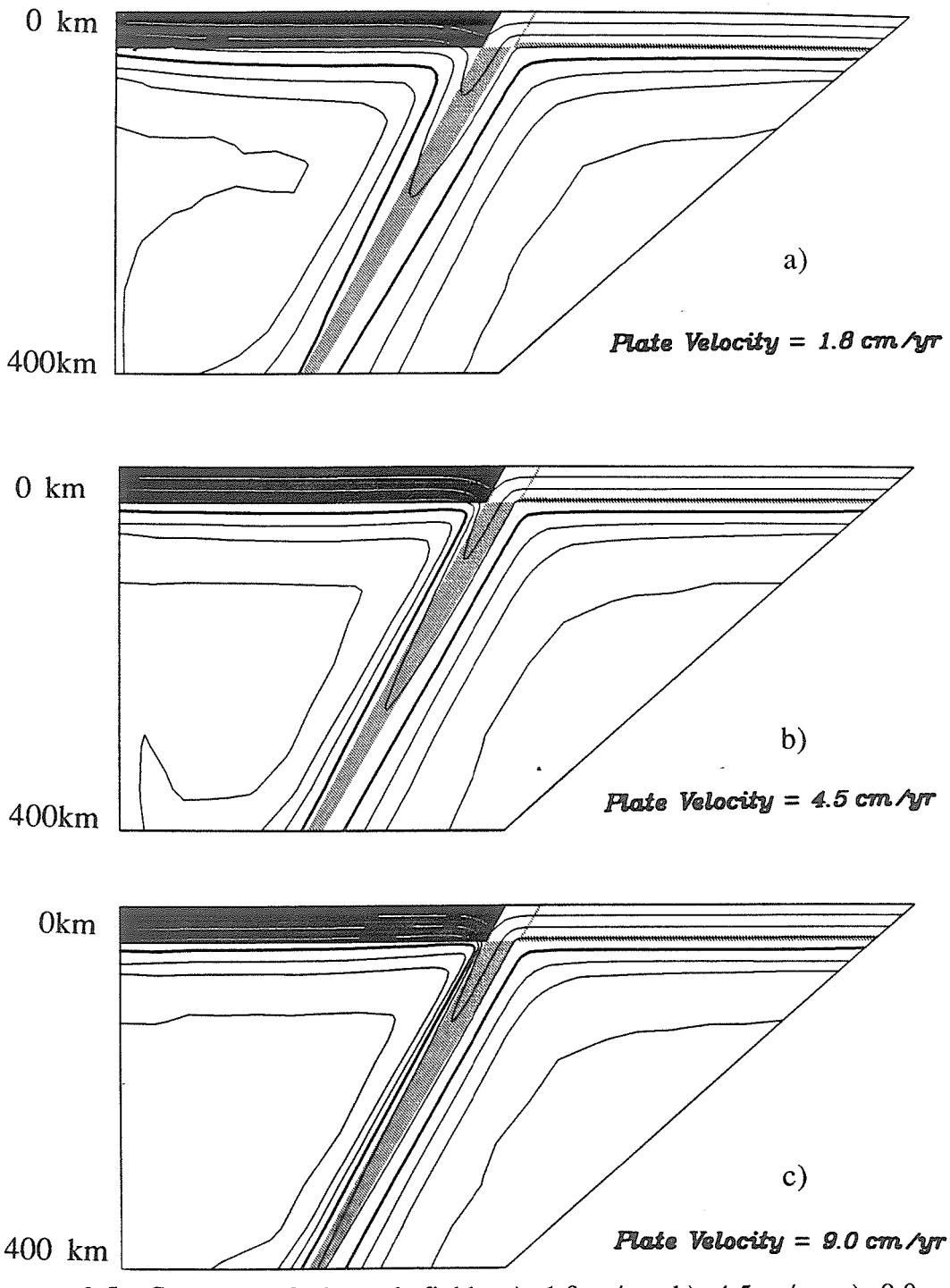


Figure 2.5 Contours of thermal field. a) 1.8cm/yr, b) 4.5cm/yr c) 9.0 cm/yr. The solid area is the mechanical over-riding lithosphere. The shaded area is the descending oceanic lithosphere. The heavy contour corresponds to 800°C, while the contours are spaced 200°C apart.

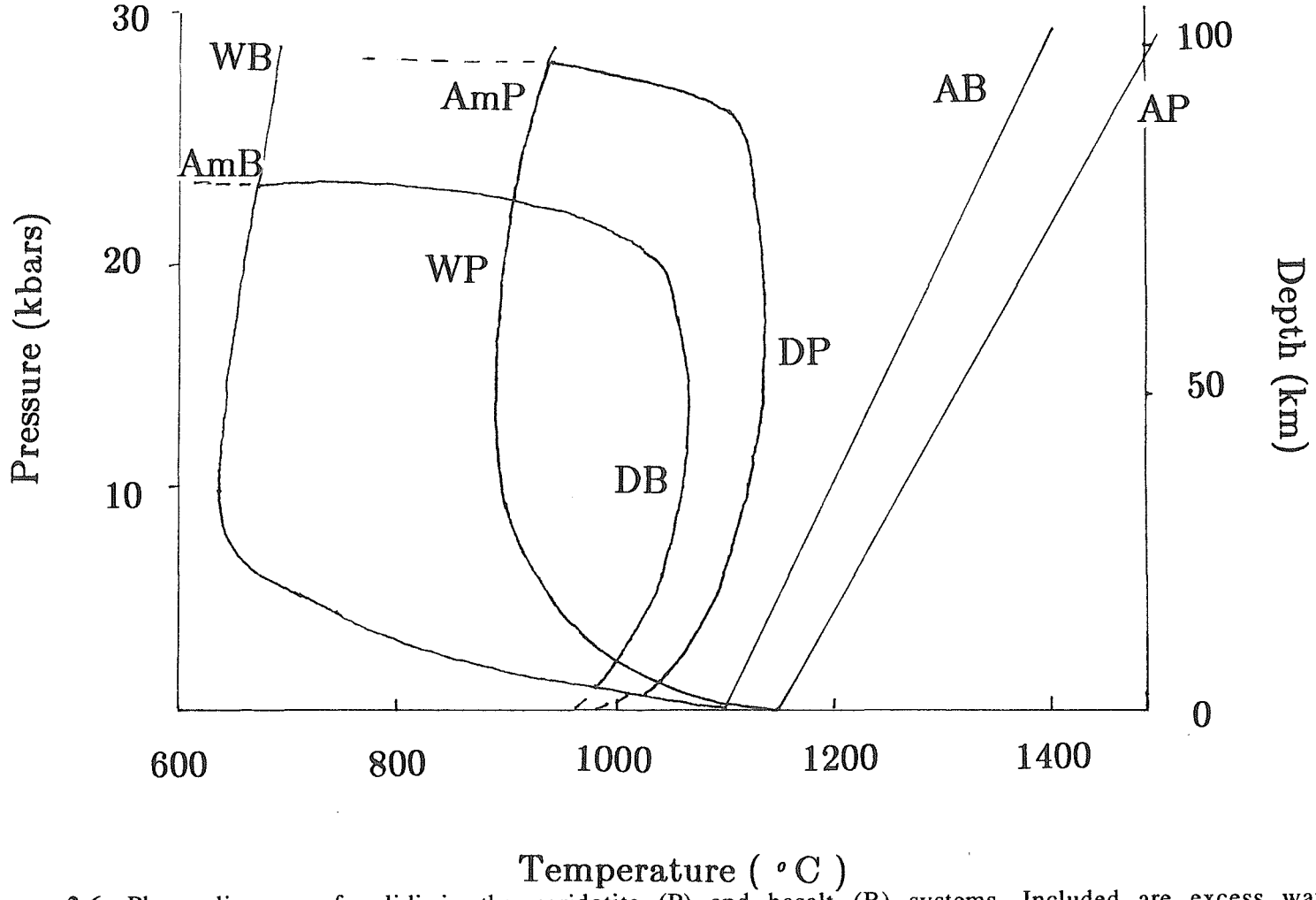


Figure 2.6 Phase diagram of solidi in the peridotite (P) and basalt (B) systems. Included are excess water (W-wet), amphibole buffered (D-damp) and dry (A-anhydrous) solidi from 0 to 30kbar. Also we have AmP and AmB, which are the dehydration curves of amphiboles in the peridotite and basalt system respectively.

collected in *BVTP* [1981]. An indication of the uncertainty can be obtained by noting that the DP solidus at 25kbars is placed at 1050°C by Wyllie and 1150°C by Green. The largest discrepancy away from the curves presented was the work of *Mysen and Boettcher* [1975] that suggested that the WP solidus was closer to 800°C than 1000°C at 30 kbar. *Green* [1976] suggests that *Mysen and Boettcher* [1975] had misinterpreted glassy deposits from the vapor phase as glass quenched from liquid. *Delaney and Helgeson* [1976] theoretically calculated the stability of hydrous phases in a subduction zone. Their calculations are of limited use though, since they did not consider the amphibole, pargasite hornblende, which is the hydrous phase expected from experimental work, e.g., *Green* [1973]. *Olafsson and Eggler* [1983] found that amphibole became unstable at 23kbar in peridotite, as opposed to 30 kbar of *Green* [1973] which we assume for our discussion. The system of *Olafsson and Eggler* [1983] had CO_2 as well as H_2O. *Wallace and Green* [1988] also looked at a peridotite system with CO_2 and found that amphibole was stable to $\approx$ 32 kbar, but that a carbonate rich melt was formed between 20 and 30 kbar while the amphibole remains stable. If the result of *Olafsson and Eggler* [1983] is applicable, then the initiation of melting would be shallower and quantitative predictions of our model would be changed. Qualitatively, the proposed behavior of the system would be unchanged.

The phase diagrams should be considered with a little caution, since the early experiments were frequently done at uncertain oxygen fugacities and all suffered problems of iron interaction with containers (loss to noble metals, gain from iron, reduction by graphite). Oxygen fugacity is now regularly controlled, while the container problem is circumvented by the 'sandwich technique' [*Stolper*,1980; *Falloon et al.*,1987]. We shall have to wait and see if these were significant problems. Since our dimensionalization of temperature (discussed above) is also based on similar experiments via *McKenzie and Bickle* [1988] this effect might be minimized.

The presence of CO_2 generally moves the solidi to higher temperatures at pressures below 22Kbars [*Falloon and Green*,1989; *Wyllie*,1979]. Above 22Kbars its influence is different since it is buffered by stable carbonates, but generally the solidus is not too different from the WP solidus. We suggest though that CO_2 is not a significant volatile in this tectonic environment for two reasons. First, *Gill* [1981] suggests that the amount of CO_2 in fumaroles is low, this direct observation is questioned by some petrologists, notably Barnes. Second, it is uncertain how much calcite is subducted, but it seems clear that calcite will remain stable and survive until deep into the mantle.

Very little melting is expected at the wet solidus since we have little free water, but at the damp solidus we can expect of order 10% melting [*Green*, 1972]. The degree of melting is a function of both the amount of free water and the temperature. At temperatures just below the damp solidus even small amounts of water will produce appreciable melting, but when the amount of free water is as small as $< 10^{-4}$ this is practically irrelevant. Therefore we expect the relevant solidus (as regards the point where appreciable melt is produced) to be the damp solidus. Notice that at high pressures and very low pressures the damp solidus intersects with the wet solidus (i.e., in regions where amphibole is unstable). Since we ignore the latent heat of fusion in our thermal models, an allowance can be made by considering the effective solidus to be higher (e.g., around 40°C higher for 10% degree of melting).

The lines for stability of amphibole and damp melting from figure 2.6 are interpreted on figures 2.7a and 2.7c. Figures 2.7b and 2.7d illustrate the whole thermal models from which the above figures are taken. The lines are shown as double lines. Notice that at low temperatures the stability curves of amphibole in both peridotite ($\approx$ 29 kbar) and basalt ($\approx$24kbars) are horizontal, and these reactions occur at nearly constant depth. Hence the horizontal double line at a depth

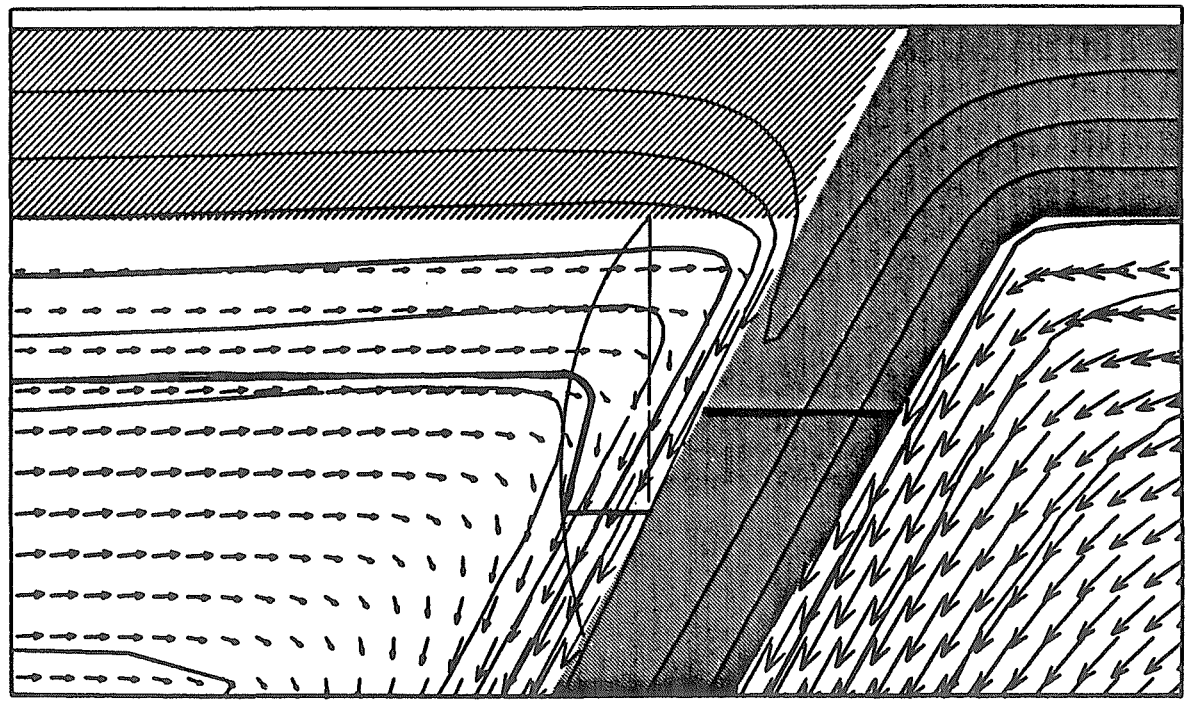


Figure 2.7a Thermal field of 60° slab with convergence velocity of 7.2cm/yr. The contours are spaced 200°C, while the heavy contour corresponds to 800°C. The double line in the plate displays the breakdown limit of amphibole. The deep horizontal double line in the wedge also displays the breakdown of amphibole while the vertical double line that turns into a shallow horizontal double line corresponds to the wet solidus in the wedge. Amphibole dehydration in mantle wedge is at a depth of 100km.

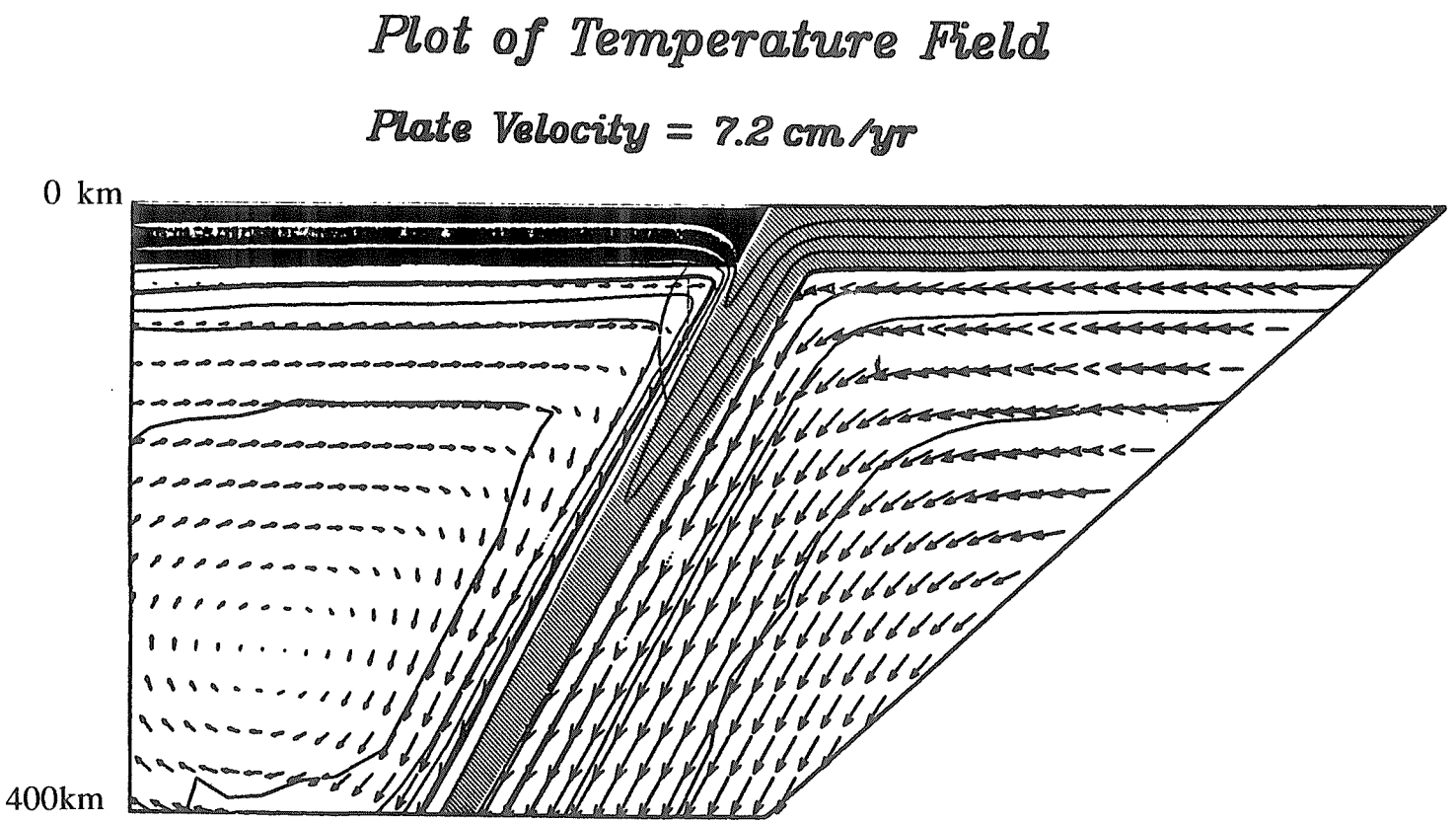


Figure 2.7b Identical to plot 7a but presents the whole model.

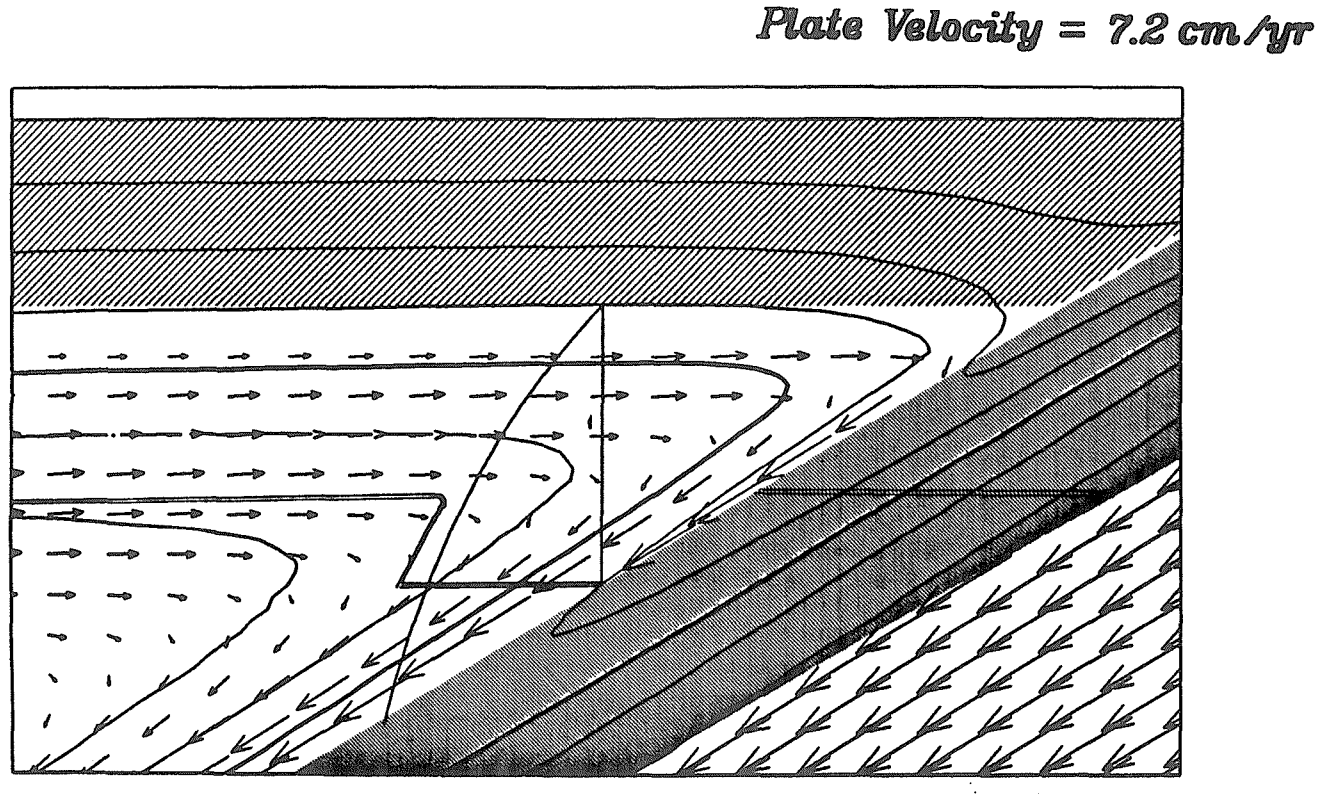


Figure 2.7c Thermal field of 30° slab with convergence velocity of 7.2cm/yr. The contours are spaced 200°C, while the heavy contour corresponds to 800°C. The double line in the plate displays the breakdown limit of amphibole.The deep horizontal double line in the wedge displays the breakdown of amphibole, while the vertical double line that turns into a shallow horizontal double line corresponds to the wet solidus in the wedge. Amphibole dehydration in mantle wedge is at a depth of 100km.

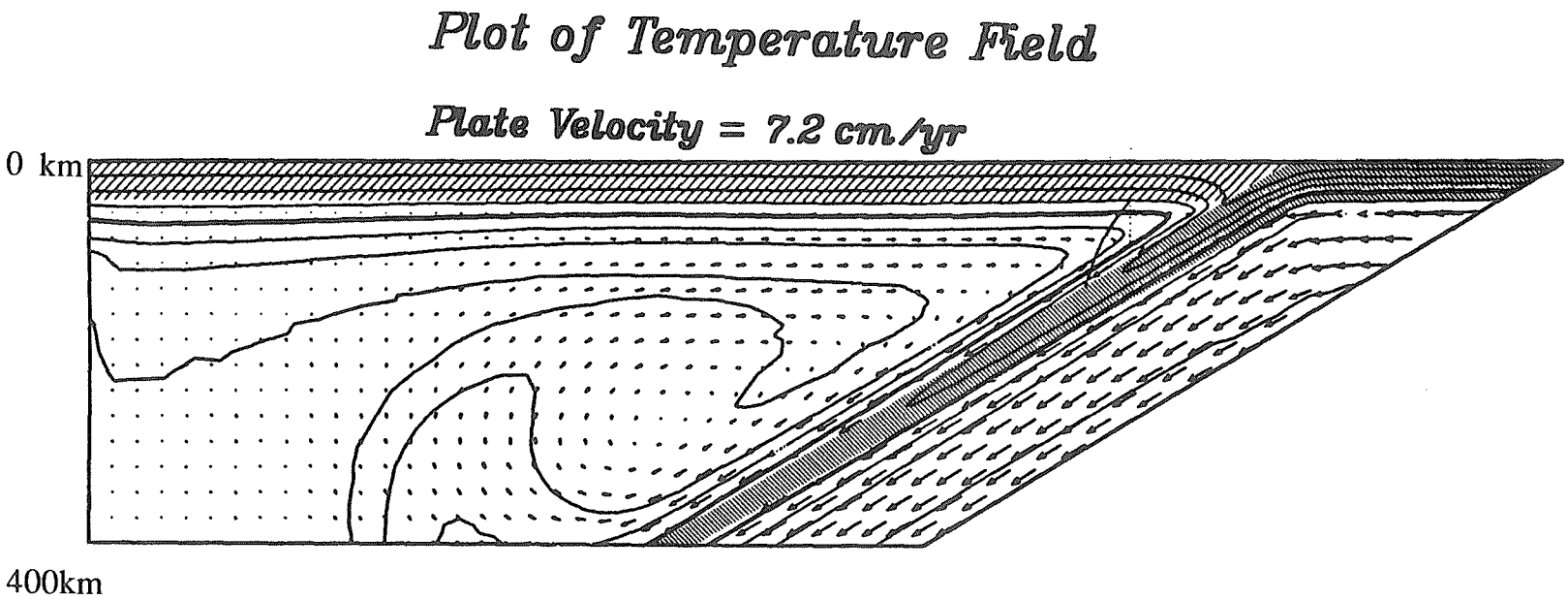


Figure 2.7d Identical to 2.7c but presents the whole model.

of ≈ 100 km in the wedge and a depth of ≈ 80 km in the slab mark the maximum depth of amphibole stability in both regimes, i.e., amphibole is unstable deeper than these lines. The vertical curve in the wedge, which turns horizontal at a shallow depth (70-80 km), corresponds to the damp solidus of peridotite. We have not drawn its continuation beyond the depth where amphiboles are unstable in the mantle, since by definition damp peridotite (amphibolitized peridotite) no longer exists. The kink in the double lined curve marks the point where the damp peridotite solidus and the amphibole stability curve intersect. Provided sufficient water is present to make the mantle damp, peridotite will melt everywhere below this line, i.e., it is the distribution of water that localizes the melting. The temperatures are dimensionalized assuming that T=1.0 equals 1325°C at a depth of 100 km on the side of the overriding plate. Figures 2.7a and b are for a 60° dipping slab while figures 2.7 c and d are for a 30° dipping slab. From the two figures one can see that we achieve sufficient pressures and temperatures to a) dehydrate amphiboles in the slab, b) melt the mantle if it is damp; definitely if it is wet, c) dehydrate amphibole in the mantle at 100 km, but d) not to melt a damp slab. Extensive melting of the slab will be even harder in the Earth since we have ignored the endothermic behavior of slab dehydration.

Since the dehydration of amphibole in the mantle occurs at ≈ 100 km and the slab's thermal boundary layer is thin at this depth, *Tatsumi* [1989] argues melting initiates very close to the slab surface due to released water lowering the solidus, and that the resulting melts rise vertically. This is his explanation for the nearly constant depth of the Wadati-Benioff zone below the volcanic front [*Gill*, 1981, (124±38km); *Tatsumi*, 1986, (112±19km)]. Below we discuss a distinct variation involving two novel ideas, the first involves lateral transport of water into the wedge and the second involves the focusing of melt towards the wedge corner due to the propagation direction of cracks. They lead to similar depths of the

underlying plate below the volcanic front, but suggest a slightly wider range as observed.

The actual temperature at the slab surface, as expected, is largely independent of velocity and dip. The slab heats up slightly more at lower velocities so that mechanisms that depend on temperature to dehydrate the hydrated minerals in the slab will lead to different depths of release of water depending on the dip and velocity of the plate. There is no correlation of either the volume of magma production or the depth of the Wadati-Benioff zone below the volcanic front with the convergence velocity. This is consistent with a dehydration reaction that is controlled by pressure rather than temperature, as the phase diagrams suggest for amphibole in a cold slab or wedge.

2.3.a.1 Water transport mechanism

From figures 2.7 a,b,c and d it can be seen that andesites cannot be generated by extensive melting of the slab, which requires temperatures in excess of 1250°C. Melting of the continental or the arc crust might be feasible if they extend down to where the isotherms are $\approx$ 1000°C. However, there is no petrological evidence that the bulk of the extruded basalts and andesites are recycled crust. There are many mantle geochemical signatures, and it would be impossible to start an island arc with such a mechanism. Crustal melting is observed in the Andes [*Hildreth and Moorbath*, 1988], but it is considered a secondary response to melts rising from a deeper primary source. It is seen from the thermal field that if there were no volatiles present, the only other way melt could be produced in this model is for the mantle to be very hot and for there to be some melt everywhere. There is no evidence for extensive melting and it is most unlikely to lead to the sharp volcanic front found in SZ. By elimination, the only other possibility is that volatiles from the slab migrate into hot regions of the mantle wedge,

lowering the solidus and causing melting.

2.3.a.1.a Evidence for importance of water in subduction zones

We have argued from our thermal models that the presence of lavas at island arcs can only be explained if water lowers the wedge solidus. The importance of sediments and water in subduction zones is hinted at further in a wide range of observations. These include ^{10}Be [*Tera et al.*, 1986; *Tatsumi and Isoyama*, 1988], rare gas systematics [*Staudacher and Allegre*, 1988], and Pb isotope arguments [*Davidson* , 1987]. Less sophisticated but more general observations of SZV indicate that they are generally more hydrous, frequently have amphibole phenocrysts, and are more likely to be explosive than most other types of volcanics. The continuous suites stretching from basalts across andesites to rhyolite at SZ suggest that along this fractionating path we are avoiding the low pressure olivine-plagioclase-clinopyroxene and the high pressure olivine-clinopyroxene-orthopyroxene thermal divides. Avoiding the thermal divide could be achieved by anhydrous fractionation at pressures <10 kbar and $> \approx 1$kbar or with the presence of water [*Wyllie*, 1979]. Water lowers the thermal stability of plagioclase more than pyroxene, and of pyroxene more than olivine. Mineral stability in andesites suggest water contents of 2 wt% [*Gill*, p.194, 1981], as does comparing eruption temperatures to the liquidi of andesites [*Gill*, p.64, 1981].

2.3.a.1.b Oxygen fugacity and volatile speciation

The oxygen fugacity (fO_2) of the wedge is important in deciding the speciation of the volatile component. If the conditions were more reducing than the Magnetite-Wustite buffer, then CH_4 would be the principal component of any fluid present and there would be very little, if any, water present. SZV commonly show high fO_2 but there is a wide range. The increased fO_2 could again reflect a

hydrous component from the slab. *Bonatti and Michael* [1989] have shown SZ peridotites to be the most depleted while *Haggerty and Tompkins* [1983] suggested that more depleted rocks have a lower fO_2. This suggests that the mantle wedge with no volatiles has a low fO_2. Similarly one would expect the harzburgite of the descending slab to have a low fO_2. From above we can see many lines of evidence suggesting a water source from the oceanic basalt and/or downgoing sediments. Water migrating from the slab would lead to high fO_2 in its neighborhood. *Mattioli et al.* [1989] from a study of published analyses of coexisting minerals in spinel peridotites suggest that there is a large variation in the fO_2 of SZ but that it is generally higher and close to the Fayalite-Quartz-Magnetite buffer. We can understand not only the high fO_2 but also the wide range discovered due to the various sources and their varying fO_2. Given a sufficient flux of volatiles it is probable that this flux controls the fO_2. Given the above estimates of fO_2, it is very likely that most H will be in H_2O while most C will be in CO_2. By water we mean H_2O rather than liquid water; in fact under mantle conditions water is above its critical point and is a fluid.

2.3.a.1.c How does the water leave the slab and set up a porous network?

We shall arrange this section following the path of a volume of water from the slab to the wedge, into melt, and its transport in the melt to the base of the lithosphere.

The miniscule volumes of water produced by dehydration of individual amphiboles will be unable to set up a sufficiently long crack to propagate large distances, unless these small volumes of water can interconnect. Hence the most probable means of propagation, at least initially, is by porous flow. The exact phase in which water leaves the slab to enter the mantle is unknown. Below we

present two possibilities. The first involves a silica-rich hydrous phase, while the second involves a water-rich melt.

We shall show that they are expected to lead to similar conditions. This is because silicic melts that are produced will be very hydrous (≈25 wt% H_2O) while the hydrous fluids that are produced can be expected to be silica-rich (≈20 wt%). Since their volumes are small in comparison to the total magma they produce, it will be difficult to differentiate between them. *Eggler* [1987a] argues against a large component of the source arising from a hydrous fluid and its dissolved solutes. He argues that MORB and SZV have the same K/H_2O ratios while at most only 20 wt% K_2O can dissolve in water from the slab. His argument does not restrict the mobile phase from being a hydrous melt, but neither does his argument restrict a hydrous phase unless the source of both MORB and SZV have the same K/H_2O ratio and the extraction of K and H_2O are the same in both environments. Since both K and H_2O are highly incompatible and have similar partition coefficients, only very small degrees of melting would separate the two components. Geochemically, identifying the mobile phase is important but in identifying the source region of SZM it may be irrelevant.

Let us consider the water leaving the slab in a silicic melt. At water saturated conditions it is found that the solidus of basalt/eclogite at 20-30 kbars is in the range 600—750°C. The thermal model of an old oceanic plate SZ gives sub-solidus temperatures at the slab-wedge interface. This interface is the focus of many heat sources and sinks that have been ignored. These include frictional heating on the thrust zone, advection of water up along the decollement, dehydration and melting. The first two could locally lead to temperature increases of up to 200°C. Dehydration is an endothermic process, while if there is melting it would be limited by the latent heat of fusion. Therefore in a SZ where the oceanic crust is young or frictional heating is high, it is conceivable that small

volumes of very silicic, hydrous melts (20-30 wt% H_2O) are produced. Basaltic melts have been shown to form an interconnected network with olivine [*Bulau and Waff*, 1979,1982; *Cooper and Kohlestedt*, 1982,1986]. The dihedral angle of siliceous melts in an eclogitic assemblage has not been evaluated. From the observation that there are frequently large garnet crystals, and that garnet seems to be immersed in pools of melt in highly molten systems, a suggestion has been made that basaltic melts will have a dihedral angle much less than 60° and possibly might wet garnet completely [*personal communication, Watson*,1989]. We therefore suggest that since basalt can equally interconnect through olivine, a melt can interconnect through a garnet-rich eclogite slab. The melts would migrate the short distance from their source to the slab-wedge interface where they must interact with peridotite if they are to migrate through by porous flow. To establish a porous network the melt and matrix must reach textural equilibrium. This will of necessity involve an approach towards chemical equilibrium. The work of *Wyllie et al.* [1989] implies that assimilation of peridotite by the silicic melt will cause crystallization. As its viscosity increases and the melt fraction decreases, it will stop migrating. It will assimilate more peridotite until it is solid. The water in the melt will be incorporated into amphiboles in the mantle wedge. This is not too dissimilar to the hybridization mechanism of *Nicholls and Ringwood* [1973].

Now let us consider the migration of water from the slab to the wedge in a silica-rich hydrous phase. There are also no experiments for the textural equilibrium of water in an eclogitic equilibrium. The only experiments on the textural equilibrium of water have involved olivine and quartz, by *Watson and Brenan* [1987], and *Watson et al.* [1989]. Their work suggests that water stays at 4 grain corners in olivine at low temperature and low pressure ($<$ 10Kbar and 1000°C). Water in equilibrium with eclogite will have a larger volume of dissolved solutes, (especially alkalis) which will make the hydrous fluid more similar to the matrix

and potentially allow lower dihedral angles. We suggest that the behavior of a solute-rich hydrous phase will be intermediate between pure water and dry melt. This is argued on the grounds that the primary controlling factor in deciding the dihedral angle between a melt or fluid and its matrix is not the temperature and pressure but rather the composition of the fluid or melt and the solid phases. For example, *Watson and Brenan* [1987], and *Watson et al.* [1989] have found that fluids have higher dihedral angles at higher CO_2 contents, lower temperatures and pressures, all conditions of lower solute solubility in the fluid phase. *Watson and Brenan* [1987] added NaCl to water in textural experiments with quartz and found an appreciable decrease in the dihedral angle. These experiments suggest that there is an inverse monotonic relationship between solute content and dihedral angle. Water at mantle temperatures and pressures will dissolve large amounts of silica [*Schnieder and Eggler*,1986;*Eggler*,1987b; *Ryabchikov et al.*,1982] and hence it seems likely that we might get closer to the situation of basalt that has a dihedral angle of less than 60° (the critical angle of the dihedral value below which interconnection is achieved, at all porosities, in an isotropic medium). Workers have shown that at high pressures the solubility of water in melts is very high [*Wyllie*, 1979], hence the first melts are extremely hydrous. In fact at higher temperatures and pressures (10 and 20 kbar, at 1200°C) *Watson et al.* [1989] have found that water will interconnect in olivine, i.e., conditions of higher solute solubility. Following the speculation regarding the wetting of garnets when discussing basaltic melts, it is argued that a water-rich fluid phase would interconnect in the slab. This would lead to rapid transport of water to the face of the slab. Since the slab interface is cold at 50 km depth, there would be insufficient dissolved solids in the fluid to allow the fluid interconnection to extend into the predominantly olivine mantle, i.e., following from our speculation that fluids can wet garnet better than olivine. After sufficient height of water collects in the slab, it can migrate into the

mantle by hydraulic fracture, since the difference between the lithostatic and hydrostatic gradient exceeds the strength of the rock. If $\Delta\rho = 2\ 10^3 kgm^{-3}$ and $h \approx 5$ km then $\Delta\rho gh \approx 1$ kbar; estimates of strength range widely but *Spera* [1987] quotes 0.5 kbar. If the dehydration kinetics were sufficiently rapid then it is possible that the stresses resulting from the volume change could help initiate cracks. The direction of fracture will be normal to the least compressive stress in the upward direction, this direction is into the mantle wedge. Such a crack, having a very limited source of fluid, will not propagate very far. The water so released would react with the peridotite, (though less pervasively than if the water was carried by a melt phase in a porous network) to produce hydrous phases including amphiboles. In the region of the slab / overriding plate interface where the thrust earthquakes occur (down to $\approx$ 60 km max.), the interface may be permeable and a conduit for water. Hence much of the water released at this depth might be expected to migrate back up the thrust zone to the surface. On the other hand a lot of the water is released deeper where there are no longer large earthquakes at the surface of the oceanic crust. It is likely that the deformation at the surface of the slab occurs by ductile flow and hence there is no reason to expect this to be a zone of high permeability.

2.3.a.1.d Lateral transport mechanism for water across mantle wedge

The water has now entered the mantle either as a hydrous silicic melt or a silica-rich hydrous fluid, the result of which has been the formation of amphiboles in the mantle wedge. The amphiboles are then carried by the induced mantle flow until they reach a depth at which they become unstable and release their water. Forced by the lack of appropriate experiments, an estimate is made of the dihedral angle of a hydrous fluid with peridotite at around 800°C and 30 kbar. We will assume, as before, that the dihedral angle is controlled by how similar the

fluid is to the matrix and that this is primarily a function of its solute content. Using data from *Eggler* [1987b] we estimate the solute content of fluids to be around 20 wt%, (between curves 4 and 5 in their figure 3). A similar estimate from *Ryabchikov et al.* [1982] would be 4 wt% at 20 kbar and 900°C but 35 wt% at 20 kbar and 1100°C, and in comparison with *Nakamura and Kushiro* [1974] we suggest that at the higher pressure but lower temperature a value of 20 wt% is not unreasonable. Rather than stating a linear relationship between dihedral angles we shall assume that the underlying physical property, surface tension, is better characterized by a linear relationship between the pure volatile and dry melt end-members. This leads to an estimate of $\gamma_{ol/fl}/\gamma_{ol/ol} \approx 1.74$, which entails a dihedral angle of $\approx 60°$ where $\gamma_{a/b}$ is the surface tension between phases a and b. Dihedral angles are very sensitive to only small changes in the value of the surface tension so this is all very speculative but it does tend to suggest that a free hydrous fluid phase could set up an interconnected network at textural equilibrium, at high pressures ($>$ 30 kbar), and even at low temperatures ($\approx$800°C). Hence it rises vertically until it is again fixed into an amphibole and the process is repeated. This is illustrated in cartoon fashion in figure 2.8. It can be seen from figure 2.8 that the net effect of this process is to transport all the water released laterally across the mantle wedge at the dehydration depth to the amphibole-saturated solidus. It is seen that this is a robust transport mechanism provided that a) the water leaves the slab to enter the wedge, b) that it rises vertically at a velocity greater than the vertical component of the induced flow carrying it down, c) that there is sufficient unhydrated mantle to carry the flux of water and d) that the direction of mantle flow is down and horizontally in the direction away from the slab. The last condition can hold for large distances (e.g., 20 km for 60° dip), the exact range depends on the flow. In appendix 2.B we present a simple analysis of the water transport mechanism in the mantle assuming that the

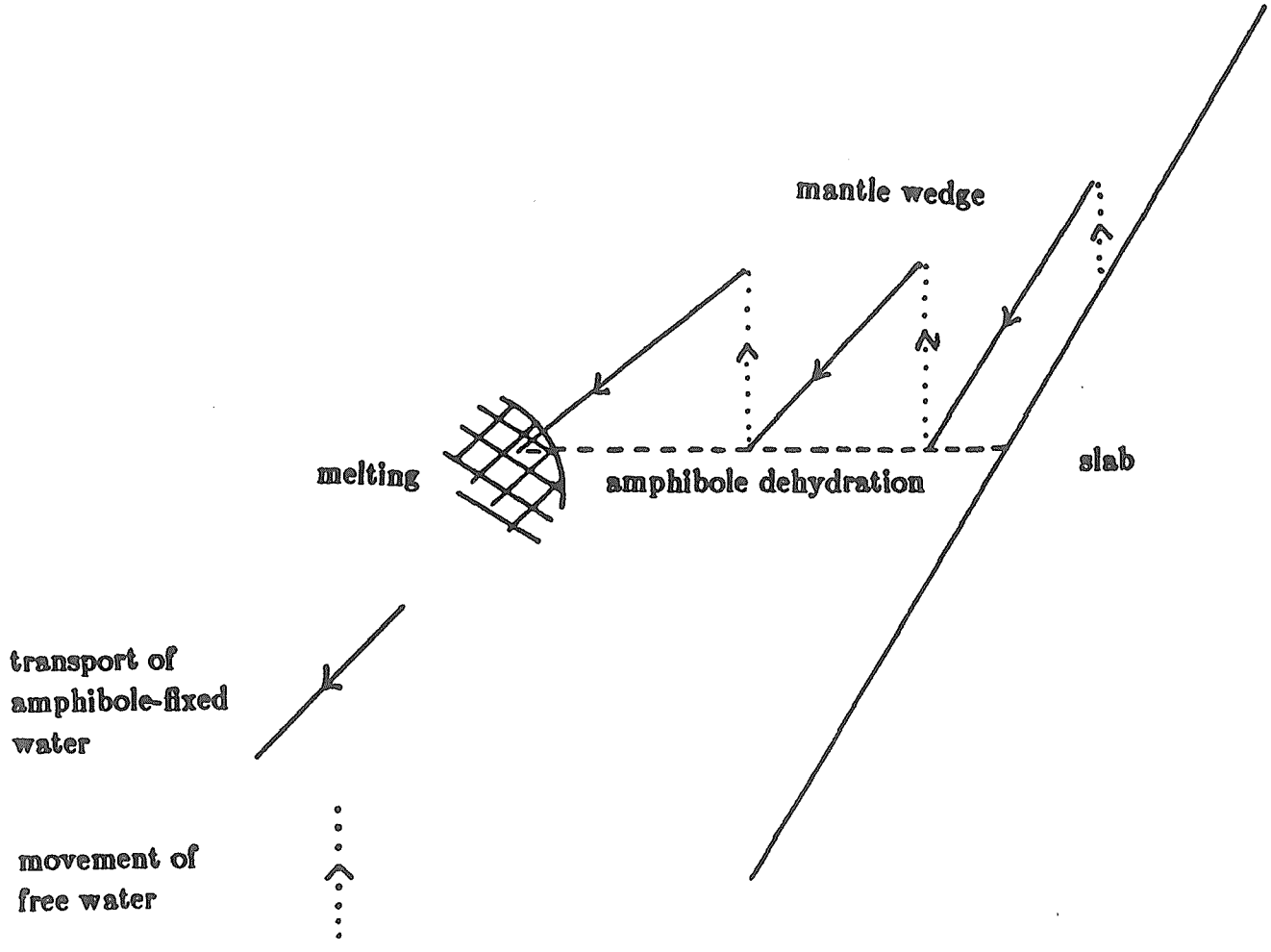


Figure 2.8 Cartoon of water transport mechanism. We are schematically following the path of one molecule of water.

upward transport of the water is by porous flow. This leads to an estimate of 0.003 wt% free water as well as 0.4 wt% water in amphibole being transported laterally at a velocity of $\approx$ 2 cm/yr. If the upward transport mechanism is very rapid e.g., hydraulic fracturing, then there is no change in the amount of water transported laterally by the amphibole but the amount of free water would be much smaller. It is seen that this mechanism will also process any other water generally present in the upper mantle and brought into the wedge corner by the slab induced flow.

Note that *Ahrens and Schubert* [1975] suggest that the basalt - eclogite transformation is kinetically delayed. If the temperature is very low at the slab wedge interface then the dehydration reactions could equally be delayed until a deeper depth in the slab and mantle. It is very difficult to get pargasite to dehydrate in the laboratory at low temperatures e.g., 700°C, due to kinetic reasons. Specifically it is believed to be delayed by the difficulty in nucleating garnet [*Holloway,personal communication*,1987]. Provided that the water is free to rise vertically in the wedge before it reaches the stability field of a DHMS e.g., Phase A at $\approx$ 200 km [*Liu* 1987], then this point is irrelevant.

As opposed to DHMS, which are unstable above $\approx$200 km, phlogopite is stable and might remain stable deep enough into the mantle to overlap with the stability of the DHMS. Its effect is predicted to be limited due to the lack of potassium in the mantle, especially in suboceanic mantle. Most of the K is likely to be brought into the wedge by the mobile phase migrating from the slab. Since the hydrous phase is unlikely to have more than 20 wt% K, then it is unlikely that more than 10 wt% water is incorporated in phlogopite. *Green* [1973] suggests that phlogopite could possibly make up to 1.6 wt% of mantle, without incorporation of K. This is to be compared to his estimate of 30 wt% amphibole. It is probable that the subduction zone wedge is depleted in K relative to the

MORB source before it is metasomatized by the melt/fluid phase from the slab. Phlogopite contains ≈5 wt% H_2O, while it is estimated that the amphibolitized mantle can hold 0.4 wt% H_2O. At the very most, 25% of the H_2O could be incorporated in the phlogopite. This is very much an upper bound, and since the upper stability limit of phlogopite is unknown, it is unclear whether this water is released as a free phase to return to the surface or whether it is carried deep enough down into the stability fields of the DHMS to be taken out of the SZM source region. *Liu* [1989] suggests that there is a region of the mantle in which water can exist as a free phase, but he suggests that this might not be the case at subduction zones, since water might be stabilized in other hydrated phases, including Mg-chloritoid, Mg-pumpellyite, and Mg-staurolite, if these phases are stable in the peridotite composition. *Liu* [1989] bases his discussion on the thermal model of *Hasebe et al.* [1970]. As discussed earlier, the model of *Hasebe et al.* [1970] involves excessive heating along the leading edge and no induced flow in the mantle wedge. Our model would not lead Liu to the same conclusion. Rather, we conclude that water can exist as a free phase in the suboceanic mantle adjacent to the slab (≈ 1000°C) from around 100 km, (the breakdown of amphibole) down to a depth of at least 250 km (below which the hydrated mineral Phase A is stable).

The presence of CO_2 reduces the solubility of solutes in hydrous fluids, and increases the temperature of the volatile saturated solidus. *Watson et al.* [1989] find that the dihedral angles of CO_2 rich fluids with olivine are $> 60°$. Hence, the presence of large volumes of CO_2 would limit the mobility of water and reduce its impact in reducing the solidus. Much less CO_2 enters the mantle wedge compared to H_2O. First, there is no equivalent to hydrothermal circulation, though there is hydrothermal carbonation largely as a result of the precipitation of calcite. Second, little of the seafloor sediments containing limestone etc. are expected to proceed down the thrust zone past the accretionary prism. Third, calcite breaks

down below 40 bars at relatively high temperatures. Conditions that are unlikely to be produced in the subducting crust. Above 40 bars calcite does not break down but melts. This occurs only at high temperatures (order 1300°C) [*Huang et al.*, 1980], hence it is probable that calcite survives through the shallow mantle to be returned to the deep mantle.

2.3.a.2 Water induced melting

As mentioned previously the largest variation in the thermal models is produced by varying the convergence velocity. Slow convergence velocities lead to very cold wedges suggesting that this induced flow mechanism would not be successful in explaining the whole range of arc volcanism. *Otsuki* [1989] suggests that the actual convergence velocity for all plates with Wadatii-Benioff zones extending below 400 km is 7.2 cm yr^{-1}. If this is true then there is no reason why the mantle wedge cannot melt. The heat is brought in by the induced flow and the solidus is lowered by the migration of water from the slab. Note for slabs with no extensions below 200 km *Otsuki* [1989] suggests velocities of 3.4 cm yr^{-1}. Our numerical model cannot apply to this case since our slabs penetrate 400 km, but even a velocity of 3.4 cm yr^{-1} is sufficient to obtain reasonably high temperatures in our numerical model. Otsuki's observation is related to trench migration, "roll-back" of the subduction hinge, back arc spreading and arc compression / extension. It is these velocities that change the widely varying global convergence velocities to this high constant local convergence velocity. Trench migration relative to the underlying deep mantle has the additional effect of necessitating large scale global mantle flows that will be supplemented on the local induced corner flow.

In the above section we discussed the transport mechanism for water and showed that water could allow melting to initiate at a depth of 100 km. Water is highly incompatible and enters the melt. If we consider that the system is in

steady state, then we can equate the water content of the melt to the porosity at each depth. This is outlined in appendix 2.C on water induced melting. There are two end member models for the melting. First, the solid matrix might be fixed and the matrix becomes more infertile as time proceeds. Second, the matrix could be renewed continuously by fresh, fertile mantle entering the source region as the residue leaves. The degree of melting as a function of temperature, pressure, and water content, could be considered. Unfortunately in the first scenario we do not know the total water content in the system while in the second we cannot define a degree of melting, since the solid matrix is not fixed.

More simplistically this whole region can be considered as one box at one temperature, pressure, and water content, and the degree of melting evaluated. This of course is a gross simplification but gives us a zeroth order estimation of the amount of melt produced. Consider T $\approx$ 1200°C, P = 30 kbar and water content of peridotite =0.5 wt%, then from *Green* [1973] we estimate a degree of melting of order 10-15%. This would suggest an average water content in the magmas of 3-5 wt%. This is much higher than for the magmas that reach the surface (0.5-1 wt%). If these figures are valid, one possible explanation is that all the melt at the source does not make it to the crust but is stranded in the mantle. Melts of higher water content will generally be cooler and as they rise will get saturated in volatiles leading to exsolution and rapid solidification. A lower bound for the amount of magma that is stranded in the mantle can be estimated if all the excess water is taken away by water saturated melts (25 wt% H_2O) leaving behind 1 wt% H_2O in the remaining magma. 8% of magma solidifies if the original water content is 3 wt%. Of course the upper bound is that nearly all the melt is stranded. Water solubility in melts remains quite high until very shallow depths e.g. 10% at 5kbars [*Green*,1973], hence it is difficult for the magmas to become saturated in the mantle.

We should like to caution that before estimates like the above can be refined, we shall need to consider melts at the surface as integrals of melt from a wide range of depths, temperatures and degrees of melting. It is important to consider the "shape" of the melting region [*O'Hara* 1985], the dynamic nature of melting [*Williams and Gill* 1989], and the integration of melts and magma chamber processes [*Cox* 1988; *Langmuir* 1989]. All we wish to suggest is that some of the melt generated in the source region does not reach the crust.

2.3.a.2.a Comparison of water and magma production rates

Since we are proposing that andesites derive from an H_2O-fluxed mantle, it is important that there is sufficient water. In appendix 2.B, estimates of water and lava production are assessed to consider the extent of mantle hydration. Estimates of oceanic crust hydration range from 2 - 6 wt%. We conservatively estimate that only 10% of this water is released into the asthenosphere if the water released along the thrust zone can proceed back along the fault zone to the surface and contribute to the large flow of water observed along the decollement [*Vrolijk et al.*, 1988]. This depends critically on the thrust zone not being "tight," and also that there is sufficient permeability in the oceanic crust for the water to reach the fault zone. It is more likely that the oceanic crust will undergo metamorphism from the greenschist facies to the amphibolite facies without losing much, if any, of its volatiles. Hence, by using 10% of the estimate of bound water subducted, we are deriving a very conservative lower bound estimate on the volume of water subducted. For a 7km thick slab descending at 7cm/yr, of density 3 $10^3 kgm^{-3}$, with 2 wt% H_2O, we estimate that 3 10^3kg of water enters the wedge per year per metre along strike of the trench. Results of ODP drilling [*Becker et al.*, 1989] suggest that the literature estimates of 2 wt% H_2O might be too large. (see appendix 2.B for discussion of this point.)

Reymer and Schubert [1984] suggest that igneous additions at island arcs during the Phanerozoic have averaged 1.1 km^3 per year, while *Brown and Musset* [1981] have estimated 0.5 km^3 per year. Given a total arc length of 3.6 10^4 km, and a density of 2.8 10^3 kgm^{-3}, we get from 2.5-5 10^4kg of crust added per year per metre along strike. These lavas are estimated to have from 0.5-2 wt% H_2O in the mantle, this suggests that the lavas return 1.25-10 10^2kg of H_2O to the crust per year per metre along strike. *Ito et al.* [1983] similarly estimated that much more water is subducted than is returned by the SZM to the crust.

Where does the water go? We speculate that very hydrous lavas are also produced in the SZ environment, and that they do not reach the crust. This conclusion was also reached in the previous section. These lavas in large part do not reach the crust because their hydrous nature implies lower initial temperatures and hence higher viscosity [*Sykes and Holloway*, 1987]. Also after rising a certain distance they become water saturated, begin to exsolve water, and become more viscous. They stagnate and solidify in the mantle. Many of these magmas will be in the lithospheric part of the mantle. Some of these magmas will pond at the base of the crust and after a sufficient volume has collected could become diapirically unstable and become the parents of tonalitic plutons and ultimately granites [*Presnall and Bateman*,1973; *Bohlen and Mezger*, 1989]. These very water-rich magmas will be reworked with inputs of more melt and as the lithosphere slowly deforms. Hence these might be the mantle precursors to the activity observed in the field by *Saleeby* [1989]. The most water-rich (>4 wt%) melts will be stranded below the crust. Lavas with larger water contents probably do not even reach the mechanical lithosphere but are swept back into the wedge to be melted again [*Arculus and Powell*, 1986].

As previously discussed some water might be returned to the deep mantle by phlogopite and DHMS. Some water is also held in nominally anhydrous phases

e.g., garnet, [*Aines and Rossman*, 1984], olivine [*Miller et al.*, 1987] and pyroxene. Since MORB and OIB melt under essentially anhydrous conditions (<0.1% H_2O), it is improbable that large amounts of water are held in a deep reservoir that does not allow sampling by ridges and plumes in a vigorously convecting mantle. *Ahrens* [1989] suggests that the Earth after accretion and the large Moon-forming impact had only few present-day oceans' worth of water in the mantle. The sub SZ mantle lithosphere, and ultimately the crust, seems a more reasonable reservoir than the deep mantle.

As we commented earlier, we discovered that younger, thinner plates are heated up more and have higher temperatures at their interfaces with the wedge. This fact combined with the importance of the dehydration and release of water from the slab might be partly responsible for the explanation of two different observations. *Nur and Ben-Avraham* [1986] observed that there was no volcanism corresponding to the path of oblique ridge subduction. *Abbott and Lyle* [1984] suggested that younger plates are hotter and dehydrate faster, with the water leaving vertically immediately. Similarly, we find that younger plates dehydrate at shallower depths but are not so confident in predicting the vertical escape of water. If the thrust zone is not "tight" then water may escape and this would explain the observations of *Nur and Ben-Avraham* [1986], since little or no water would remain to enter the asthenosphere and lower the solidus. In another observation that might be influenced by the dehydration reactions, *Ruff and Kanamori* [1987] find that the maximum magnitude of subduction zone thrust earthquakes increases for younger, more rapidly subducting slabs. In addition to the explanation of *Ruff and Kanamori* [1987] that this could be a function of the slab buoyancy, it could be related to the fact that the magnitude of such earthquakes has been shown to correlate with the "width" W (distance down dip) of the thrust zone. Possibly M_w is proportional to W^3 [*Byrne et al.*, 1988]. The release of water

could increase the width, since the transition from brittle to ductile behavior moves deeper as the water pressure increases [*Dahlen and Suppe*, 1987].

2.3.a.3 Melt transport and segregation. Direction of fracture propagation.

The transport of magmas from the source region to magma chambers high in the crust must involve crack propagation if for no other reason than to avoid freezing [*Turcotte*, 1987]. To avoid freezing on their way through the crust, the cracks need to be wide. The aspect ratio of cracks W/L is $\approx \sigma/G$ where W is the width, L the length, G the shear modulus of the melt, and σ is the available stress, and since the available stress is much lower than G, the cracks must also be long. In the wedge, the temperature difference between melt and country rock is not as high as in the cold lithosphere, hence the cracks need not be quite as wide or as long. For long cracks, it is argued that elastic forces are insignificant compared to buoyancy forces in defining their width and propagation speed [*Turcotte*, 1988; *Lister*, 1989]. These are argued to be controlled by the fluid-dynamics of the melt away from the crack tip. By contrast, elastic forces rather than buoyancy should control the direction of propagation, since it is sensitive to the local structure and stresses near the advancing crack tip. This has yet to be rigorously demonstrated, and little work has been done on finite, melt filled, propagating cracks in material with principal stress axes non-parallel to the direction of gravity. But if it is true, then the direction of crack propagation will be perpendicular to the direction of the least compressive stress, as shown by *Tsunakawa* [1983], provided the magma pressure is greater than a critical pressure. For a corner flow regime, this results in the favored direction of crack propagation being part of a logarithmic spiral (though we only get a very short nearly linear segment) towards the corner. For details see Appendix 2.D.

We assume two different regimes. First, inside the mechanical lithosphere where we assume that the least compressive stress is in the horizontal plane. In this regime the cracks propagate vertically. Second, in the asthenosphere where we have argued above that the least compressive stress will be aligned so that cracks are focussed towards the corner. In the real Earth of course, there will be a more gradual transition between these two regimes, but given the high temperature gradients and the exponential temperature dependence of viscosity the transition region might be very narrow. In the following discussion we are assuming the base of the mechanical lithosphere as defined for large scale horizontal plate movements is also the depth at which this transition in the orientation of stresses occurs in the lithosphere. This is unclear. If it's shallower, then melts can be focussed from further out in the mantle wedge than assumed in the discussion below.

Does this mechanism, of lateral transport of water, followed by stress controlled fracture propagation, lead to the volcanic front being located $\approx 124 \pm 34$ km [*Gill*, 1981] above the Wadati-Benioff zone? If the mechanical lithosphere (in which the stress regime is uniform laterally and controlled by the plate forces at a distance) is $\approx$ 100 km thick, then as argued by *Marsh* [1979] and *Spiegelman and McKenzie* [1987], the melts can be generated or focussed into the corner and rise vertically from this point, automatically satisfying the depth constraint. As argued in the Method section above, it seems unlikely that the mechanical lithosphere is this thick.

At different dips we find that the thickness of the thermal boundary layer is unchanged (provided that the convergence velocity is similar) but the horizontal distance to be traversed to reach the solidus will be proportional to $\mathrm{cosec}\theta_b$, where θ_b is the dip angle of the plate. Propagating cracks that are reversed from the correct point (i.e., the base of the mechanical lithosphere at a depth of 60 km

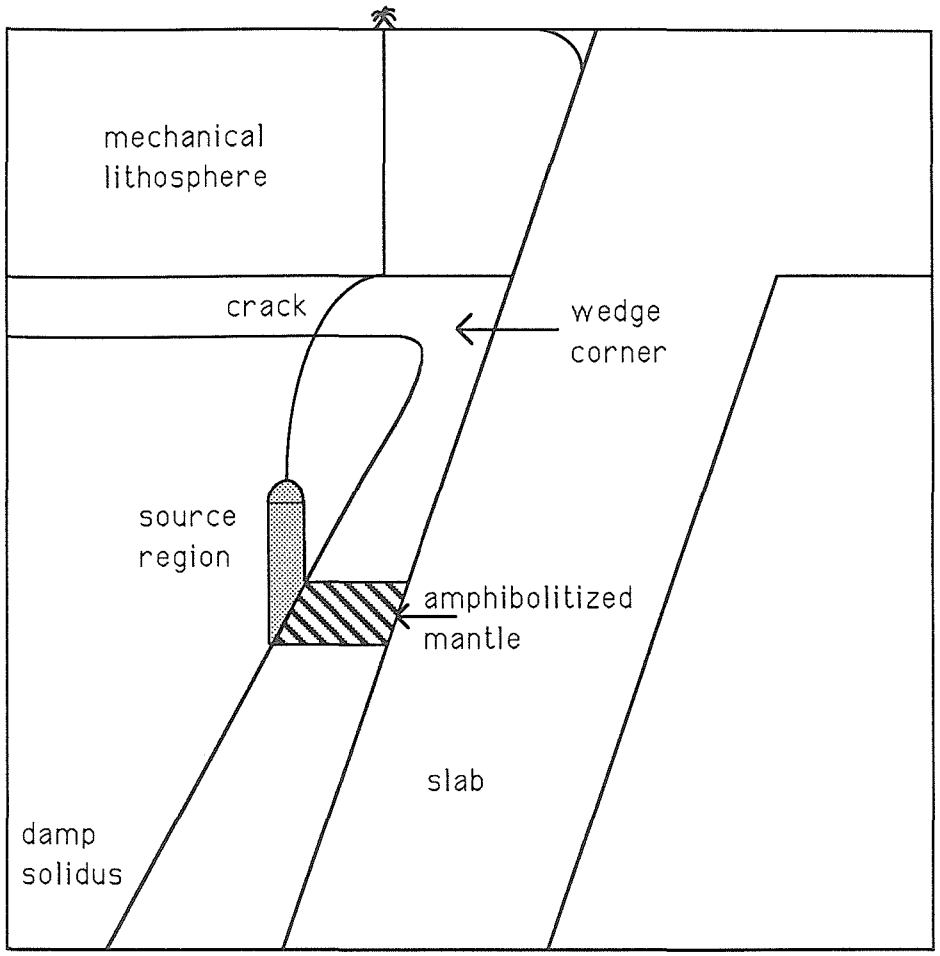


Figure 2.9 Cartoon of geometry of subduction zone source region.

above the slab, equivalently a distance ≈ 60 Cot $\theta_b$km from the wedge corner) are seen, in figures 2.7a and 2.7b to intersect the 1200°C isotherm for the 60° dipping slab while for the 30° slab in figures 2.7c and 2.7d it needs to be displaced $\approx$ 30 km laterally. This leaves the volcanic front $30\sin\theta_b$ or 15 km higher above the subducting plate, well within the variation quoted by *Gill* [1981].

The maximum lateral displacement of the crack occurs at an angle of 45° for slabs that have dips $> 45°$, otherwise the maximum lateral displacement of the crack from the vertical occurs at the slab. Note that this displacement is proportional to $\text{Cot}\theta_b \times \exp(\pi/4 - \theta_b)$ for $\theta_b > \pi/4$, while it is otherwise proportional to $\text{Cot}\theta_b$. As can be seen from figure 2.10, this is roughly parallel to $\text{Cosec}\theta_b$. This leads us to expect the depth of the Benioff zone below the volcanic front to be a weak function of dip, with one exception. The exception is for SZ with very steep dips $> 80°$, where it is expected that the volcanic front will be further from the underlying slab. This is actually observed [*Gill*, p.14, 1981], but it is frequently suggested to be a bias resulting from the larger errors involved in estimating the vertical position due to the steep dips [*Tatsumi*, 1986].

The maximum horizontal displacement of the crack's propagation path away from the vertical occurs near the maximum temperature for this end-point fixed to the overriding plate's mechanical lithosphere. To satisfy the depth of slab below the volcanic front ($\approx$110km - assuming that the Benioff Zone is located some 10 km below the surface of the oceanic crust at this depth) we suggest that the fractures propagate out from this depth of maximum temperature. This will also be the region of maximum melt production. This suggests that the water does not, as a rule, continue up to the shallow damp solidus, but is rapidly removed (in the melt) by cracks to the base of the mechanical lithosphere. With no water, the solidus of the wedge shallower than the region of maximum melt production will be at or close to the anhydrous solidus.

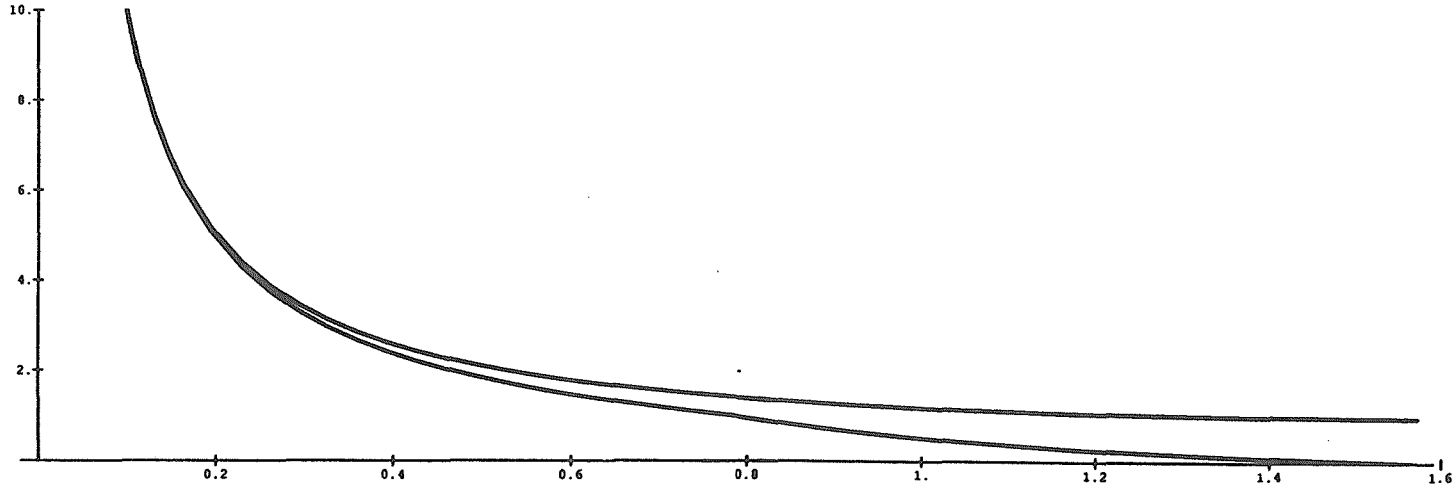


Figure 2.10 The lower line is a plot of $\mathrm{Cot}(\theta)\ \exp(\theta-\pi/4)$ from θ to $\pi/4$ and $\cot(\theta)$ from $\pi/4$ to $\pi/2$. This function is proportional to the maximum horizontal displacement of a crack for a wedge with dip angle θ. The upper line is a plot of $\mathrm{cosec}(\theta)$. This function is proportional to the horizontal width of the boundary layer. As one can see, except at very large dip angles these two are very similar.

For the 60° dipping slab, we conceive that water initiates melting at a depth of around 100 km depth and 20 km away from the slab. This water is then carried by melts to higher, hotter regions in the wedge where it induces ever greater degrees of melting. But on reaching the hottest isotherm we suggest that the melting cannot advance vertically since there is insufficient water in the melt. At the colder temperature the wet solidus is only attained at a higher water content. The melts are not able to migrate further and start to stagnate and crystallize. Crystallization allows some of the melt to progress slightly further since it achieves higher water contents, and the melt exsolves a vapor phase that could help in initiating cracks. Hence this freezing region provides a lid to the source region against which the melts accumulate and from which cracks propagate to the base of the mechanical lithosphere. This might seem slightly improbable, since for cracks to initiate the pressure must develop on a timescale fast compared to the Maxwell time of the rock, otherwise the surrounding rock could relieve the pressure by deforming viscously. Also we need to produce pressures large enough to overcome the strength of the material, estimated by *Spera* [1987] to be around 0.5kbar. The compacting matrix can produce a pressure difference $\approx \Delta\rho g L$, where L is the the compaction length. $L = a\sqrt{(\varsigma+4\eta/3)/(b\eta_l)}$, where a is the grain size, ς and η are the bulk and the shear viscosities of the solid respectively, while b is the constant in the law relating permeability k, to the porosity f, $k = a^2 f^n/b$, and η_l is the viscosity of the liquid. For matrix viscosities of 10^{19}Pas, a grain size of 10^{-3}m, a liquid viscosity of 10 Pas, and $b = 10^3$, one gets a compaction length of around 30 km. Rock strength of 0.5kbar would require a compaction length of $\approx$ 16 km, to produce sufficient pressure to fracture and initiate a crack, assuming a $\Delta\rho \approx 4\ 10^2 kgm^{-3}$. The pressure could develop suddenly. It could be the result of rapid interconnection across the whole source region as it crosses the percolation threshold. This leads to rapid migration of a pulse of melt, which will be

brought to rest by the cap. This rapid interconnection combined with the exsolution of a vapor phase might occur sufficiently quickly to induce cracks even in this relatively hot rock $\approx$ 1200°C. Whether or not a percolation phenomenon is responsible for generating the cracks, there is evidence for cracking of hot material, see *Nicolas* [1986] and *Sleep* [1988].

The above models predict the generation of melt but have not considered its potential contribution to the dynamics, due to its buoyancy. In the following section, we shall illustrate how this effect can generate a local flow reversal and extend the above discussion.

2.3.b Flow reversal in wedge

2.3.b.1 Flow reversal : basics.

Higher temperatures can be generated in the wedge by invoking a mechanism involving a reversal of the flow in the wedge leading to melting by adiabatic decompression [*Ida,* 1983, 1987; *Nye and Reid,* 1986; *DeBari et al.*, 1987; *Plank and Langmuir,* 1988; *Davies and Stevenson,* 1988; *Turcotte (personal communication);* *Tatsumi et al.*, 1983]. This mechanism has been investigated for the case of a spatially fixed buoyant body force in two dimensions. We have considered a case where the flow reversal is an internal feature of the system, superposed on a conventional external flow. This may be generated by the buoyancy arising from the extensive hydration of the mantle wedge and the small degrees of melting at the wet and damp solidi; melting introduces buoyant residue and melt [*Scott and Stevenson,* 1989]. This large cylindrical diapir rises vertically until the degree of melting increases beyond the point of interconnection. Then we get segregation of the melt from its source and extraction by small cracks through the mantle wedge [*Nicolas,* 1986] as described above. A stationary body force might not be an

unreasonable description of such a system since the cracks transport the melt rapidly and hence their contribution to the buoyancy force is minimal. Therefore the source of buoyancy is solely the source region, which is reasonably fixed since it is related to the water flux.

Given our linear rheology, it would be possible to solve the two problems separately, i.e., the corner flow and the buoyancy induced flows with their respective boundary conditions and superpose the two resultant flow fields. The corner flow problem can be solved analytically but the buoyancy induced flow would be tedious and the resulting insight would be minimal.

A suite of numerical examples were investigated where the magnitude of the buoyancy force is characterized by the buoyancy number mentioned in the Method section above. Some resulting velocity fields are illustrated in figures 2.11 , 2.12 and 2.13. We observe that for these models we can reverse the wedge flow some distance above the location of the source of buoyancy but the flow is nearly all downwards immediately below the source of buoyancy. In this model we do not see the 'updraft' that *Ida* [1983, 1987] claims for his buoyancy driven reversal. In figure 2.11(a) we present the whole model (N.B. 700km deep) and the box represents the magnified regions shown in the other figures. In figure 2.11(b) we present the corner flow with no source of buoyancy. Figures 2.11(c) and (d) illustrate the difference a factor of 2 in the buoyancy makes. In the upper two plots of figure 2.12, we see that the resultant flow pattern away from the buoyant region is approximately equivalent. This emphasizes that it is the total integrated buoyancy that is important. Total integrated buoyancy is the product of the volume and buoyancy of the buoyant region. Comparing the upper right plot with the one on the lower left we see that the location of the buoyancy force relative to the plate is important. The buoyancy force is practically the same but the reversal is more intense for the source of buoyancy placed further away from the slab. It can

Figure 2.11 Four velocity plots. (a) represents a box 700km deep in which the following models were ran. The small box represents the position of the magnified views presented in the next three figures. The shaded area represents the mechanical lithosphere, i.e. 40km thick. (b) Represents the velocity field with no buoyancy force present. The bottom of this blow up is at a depth of ≈ 140km while the top of the box is at ≈60km. (c) Is with buoyancy in the shaded box of buoyancy number of 10^6. Note that there is only a very weak localized return flow. (d) Same as (c) but Buoynacy number is increased to 2 10^6. Notice that the flow reversal is much more vigorous.

Figure 2.12. Same convention as figure 2.11.Four plots of the velocity field in a mantle wedge with a localized source of buoyancy. The localized buoyancy force is in the shaded region. The descending slab is dashed. (a) and (b) have the same total buoyancy and the similarity shows that this is more important than its concentration. Figures (b) and (c) illustrate the difference that the position of the buoyancy force relative to the slab has a large effect. (d) Illustrates the much weaker flow reversal with half the buoyancy.

Figure 2.13 Same convention as Figure 2.11. Four plots of velocity field given a localized source of buoyancy. Both plots show the flow of the wedge if the plate and wedge are decoupled by a weak zone down to 200km. In (a) and (b) the weak layer has a viscosity a factor of ten lower while in plots (b) and (c) the weak layer is a factor of 100 less viscous. (b) and (d) are the same as (a) and (c) but are at a lower magnification.

Figure 2.11 *Plots of Velocity Field*

Plate velocity = 7cm/yr

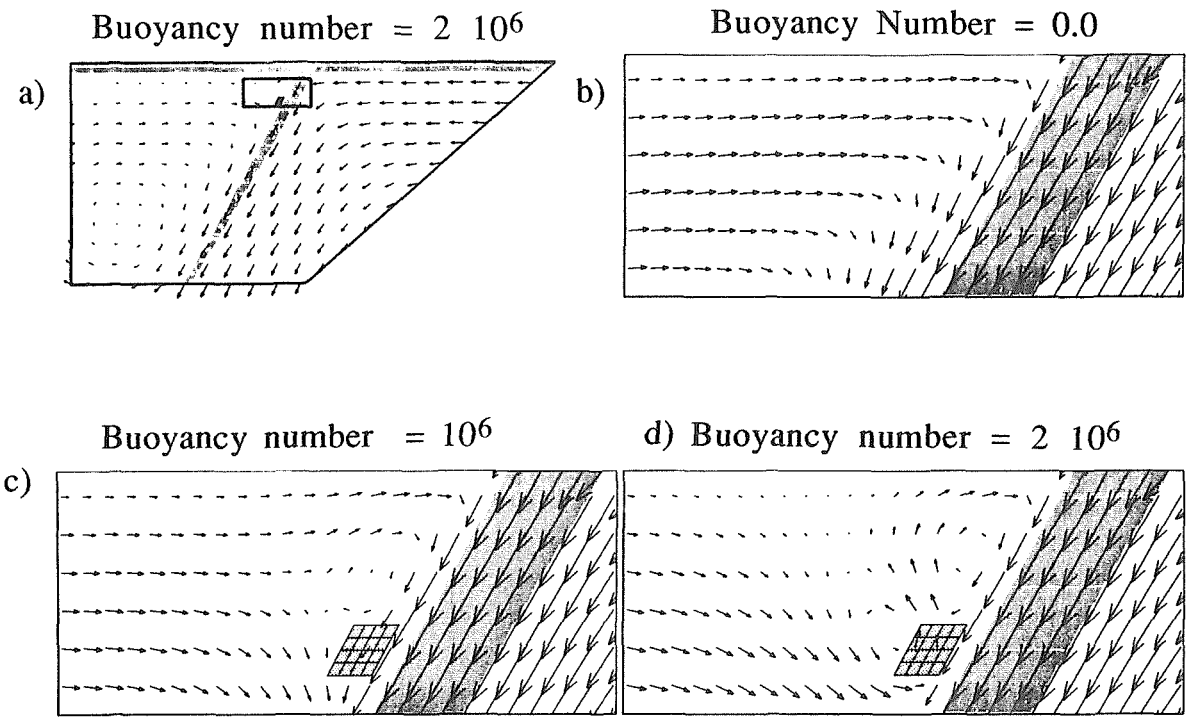
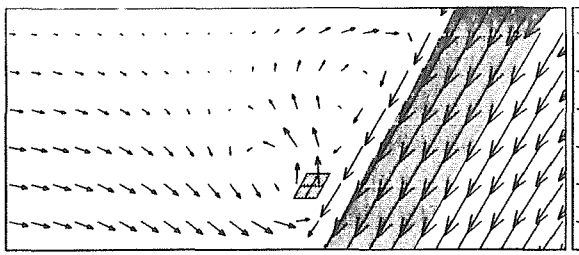


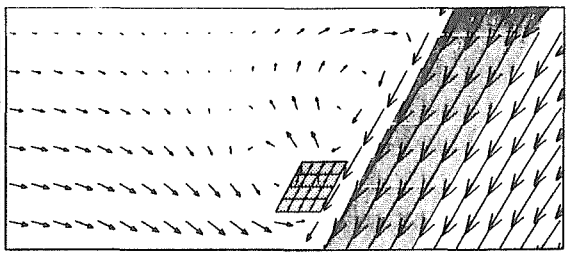
Figure 2.12 *Plots of Velocity Field*

Plate velocity = 7 cm/yr

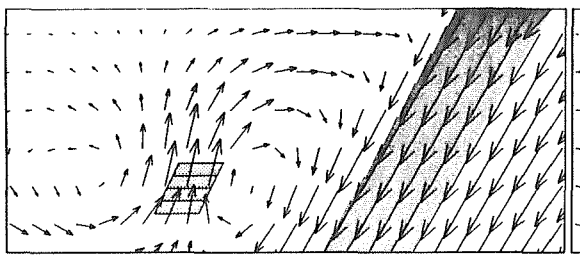
a) Buoyancy number = 8 10^6



b) Buoyancy number = 2 10^6



c) Buoyancy number = 2 10^6



d) Buoyancy number = 10^6

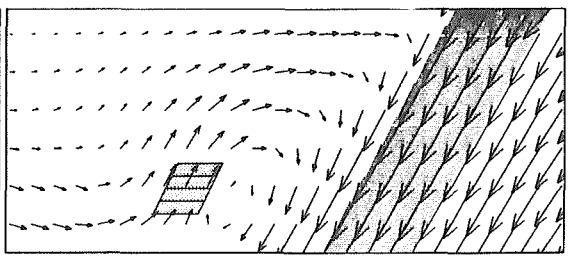
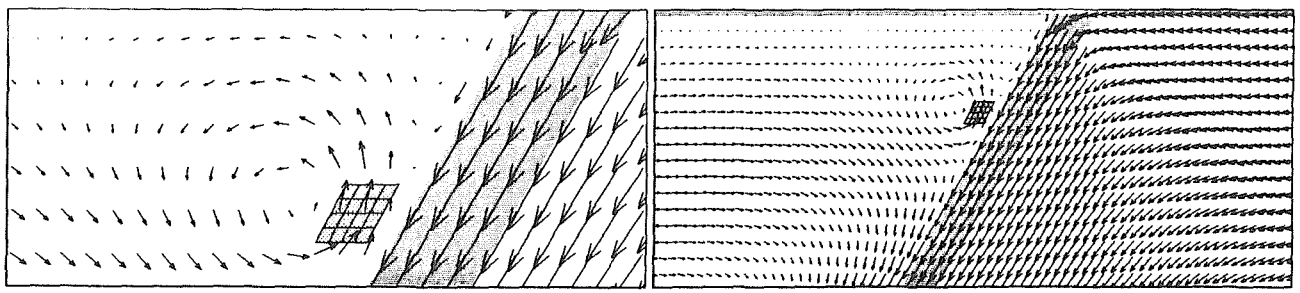


Figure 2.13 *Plots of Velocity Field*

Plate velocity = 7cm/yr.

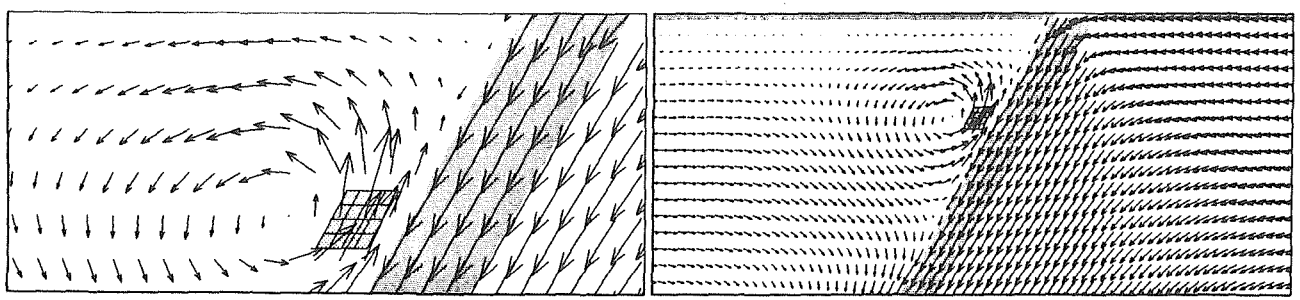
a) Buoyancy number = 2 10^6

b) Buoyancy number = 2 10^6



c) Buoyancy number = 2 10^6

d) Buoyancy number = 2 10^6



be shown that the critical value of the buoyancy to produce a reversal is such that the upward velocity of the buoyant region as predicted for a Stokes cylinder is greater than the induced plate flow at that point. The velocity of a buoyant rigid cylinder in an infinite medium is given by *Landau and Lifschitz* p.68 [1959] as

$$v = \frac{R^2 \Delta\rho g(0.5 - \gamma - \log(vR\rho/4\eta))}{4\eta} \qquad (6)$$

where γ is Euler's constant ≈ 0.577, η is the viscosity of the matrix through which the cylinder flows and R is the radius of the cylinder. Note that the velocity is also on the right-hand side, but being in a logarithmic term it is a very weak dependence.

In comparing the two lower diagrams in figure 2.12 it is seen that the critical buoyancy number for this distribution of buoyancy is slightly $< 10^6$. If the diapir is cylinder-like of radius 10 km and the buoyancy is ≈ 10 kgm^{-3} then a critical buoyancy number of order 10^6 is achieved with a local wedge viscosity of $\leq 7\ 10^{18}$Pas. A buoyancy of 10kgm^{-3} is equivalent to 2 wt% melt, if $\Delta\rho$ between the matrix ($\approx$3.34 10^3kgm^{-3} [*Klein and Langmuir* 1987]) and melt ($\approx$ 2.95 10^3kgm^{-3} [*Rigden et al.* 1984]) is $\approx$5 10^2kgm^{-3}. *Sato et al.* [1988a] obtained a large increase in electrical conductivity when 2 wt% melt was present. They believe that the large increase in conductivity occurs as a result of interconnection. Estimates of the largest seismic compressional velocity variations between the wedge and the center of the plate suggest anomalies of $\approx$10% at a depth of 100 km [*Anderson*, 1988]. If the temperature contrast between the molten region and the center of the plate is 800°C, and $\partial V_p/\partial T$ = -4.0 10^{-4} kms^{-1}°C^{-1} [*Anderson et al.*, 1968], then it suggests that 5 to 7% of the contrast has to be explained by the presence of melt. *Mavko* [1980] suggests that 3 wt% melt organized in a tube geometry, (which is closer than films or spheres to the geometry expected from melt texture experiments) can explain anomalies of this size. Similarly the

laboratory results of *Sato et al.* [1989] suggest that this size anomaly would correspond to 2 wt% melt. *Sato et al.* [1989] reinterpret the seismic work of *Suyehiro and Sacks* [1979] below Japan suggesting that very little melt is present, since they only found anomalies of $\approx 6\%$. This could locally be the case; i.e., the degree of melting along strike could vary, but we believe that there might be other explanations. It is possible that the seismic energy from the Benioff zone earthquakes, (that are the sources for this work) did not sample the center of the slab, hence they might not see the full anomaly. Remember, the slab is an anti-wave guide. Further, we argue that the source mantle in a SZ is hydrated and infertile rather than consisting of fertile peridotites on which the laboratory experiments were conducted. Similarly, estimates of melt content can be made from the attenuation structure of SZ. The low Qs are consistent with 2 wt% melt but the Q structures of mantle wedges are poorly resolved and the laboratory extrapolation of ultrasonic results [*Sato and Sacks*, 1989], depends on the same mechanism of attenuation operating at both of these widely differing frequencies. The choice of 10 km radius for the source is given by the intersection of the extent of hydration with the wet solidus isotherm, and the height of the source, given by the height that the maximum temperature occurs above the amphibole dehydration depth. These dimensions are poorly predicted and to some extent arbitrary but the equivalent radius cannot be wrong by more than a factor of 2. This is in contrast to the uncertainty in predicting the appropriate viscosity, discussed next.

The efficiency of the buoyancy force depends critically upon the wedge viscosity. Assuming that the buoyancy is due to the extensive hydration of the mantle and more concentrated regions of partial melting, one might argue that the effective viscosity in the adjacent regions could vary from 10^{19}—10^{21}Pas, depending largely upon how close the rock is to its melting temperature. A viscosity of 10^{21}Pas requires implausible degrees of hydration and melting, while as shown

above a viscosity of $< 10^{19}$Pas would require 1-2% melt. *Rydelek and Sacks* [1988] estimate the viscosity of the asthenosphere beneath the Japanese island arc to be 10^{18}Pas from the diffusion of a stress pulse from the overriding lithosphere to the subducting slab to explain a 27 yr delay in the triggering of large subduction events by events in the overriding lithosphere. *Thatcher et al.* [1980] estimate its viscosity to be 10^{19}Pas from the relaxation of the deformation following the Sanriku earthquake of 1896. Peridotite at its solidus might have a viscosity closer to 10^{18}Pas [*Craig and McKenzie,*1986] and hence if much of the wedge is close to the solidus it is possible that the volume averaged viscosity of the wedge is close to 10^{20}Pas. Other estimates of the asthenospheric viscosity include: *Wdowinski et al.* [1989] $<10^{20}$Pas from deformation of Andes and Aegean, *Cathles* estimates from post-glacial rebound $< 10^{21}$Pas for the whole asthenosphere as opposed to 4 10^{19} Pas for the uppermost asthenosphere. In a constant viscosity wedge, it is only possible to produce a flow reversal if the average viscosity $\approx 10^{19}$Pas.

2.3.b.2 Decoupling of wedge

From the estimates presented above it is probable that the viscosity of the wedge will be too large to allow flow reversal given the magnitude of buoyancy forces available. Flow reversal though could be achieved if the mantle wedge flow is largely decoupled from the slab. This can be achieved by the presence of a weak layer near or at the surface of the slab. In figure 2.13 we show the results of buoyancy-induced reversal with a thin layer in the wedge on top of the slab with a viscosity one tenth and one hundreth the viscosity of the rest of the wedge, extending to a depth of 200 km. The lower diagrams with the more decoupled wedge show substantially higher reversal velocities. Possible mechanisms for generating a weak zone include phase changes (dehydration, hydration, basalt-eclogite), presence of water and sediments, and the generation of shear zones.

Even though friction along a mega-shear might not produce melting, it can potentially produce low viscosities because a rock's strength drops dramatically way before it reaches its melting temperature.

A slab decoupled from the wedge by a layer of viscosity 10^{20}Pas will be important for modeling the migration of seismicity along the slab [*Dmwoska et al.*,1988] and from the slab to the continental overriding lithosphere and vice-versa [*Rydelek and Sacks*,1988]. Unfortunately, the wedge viscosity is but one component in the migration of stresses and cannot be easily disentangled from other components. Moreover the observations of correlations between the slab and overriding lithosphere seismicity are preliminary. As these studies progress they will help constrain the coupling of the slab and wedge.

Evidence for potential decoupling and weakness can be found in the work of *Raleigh and Paterson* [1965] on the deformation of serpentine, and *Heard and Rubey* [1966] on the tectonic implications of the dehydration of gypsum. Also the work of *Yuen et al.* [1978] and *Yuen and Schubert* [1979] suggests that there could be a shear zone, since even though the increase in temperature is unlikely to be dramatic, there would be a substantial reduction in viscosity. This would be more than sufficient for the wedge to become decoupled.

2.3.b.3 Dimensionality of problem

Two comments need to be made regarding the above discussion. First, the dimensionality of the problem needs consideration, and second, other sources of buoyancy may be present. Even though the problem seems 2D with plates extending into a homogeneous mantle, we observe localized volcanic edifices at the surface. Also the volcanic rocks show geochemical and isotopic variations. Similarly there are variations of other geophysical variables along the arc. Hence it is conceivable that the partial melting process has a large degree of along arc

variation.

If so, and if they are well separated, then it becomes more difficult for individual diapirs to reverse the flow. This can be seen by calculating the radius required for a diapir to move at 5 cm/yr through a matrix of viscosity 10^{19}Pas. It is $\approx$ 40 km, i.e., a factor of 16 more buoyancy. Hence if the sources of buoyancy are localized along arc then they will lead to weaker wedge flow reversal and less melting by adiabatic decompression. The variation in degree of melting along arc need not be large or localized, to produce the localized edifices. This localization could be more a reflection of the segregation process than the distribution of melting [*Stevenson*, 1989]. If the variation in the amount of melt present is small and smooth, then the process can be conceived of as being 2D rather than 3D, since the diapirs would interact and add to each others' flow and reduce the drag. Due to the large Prandtl number of the mantle, the influence of each diapir will extend very far, hence a large interaction is to be expected. The localization probably is a sensitive function of porosity.

The large cylindrical diapir will be unstable along strike and might be expected to develop more spherically shaped diapirs. From *Kerr and Lister* [1988] it is found that the spacing of instabilities of a cylindrical diapir is independent of viscosity contrast between the diapir and its surrounding and that it is in a constant ratio to the radius of the cylindrical diapir. It must be pointed out that a partially molten system only behaves as a viscous diapir as modeled by *Kerr and Lister* [1988] as long as there is no interconnection. Hence this instability might only control an early diapir before it starts rising rapidly or developing extensive interconnection, but the deformations developed during this phase might lead to regions of larger adiabatic uprise and to the higher probability of fracture developing at the peaks of these instabilities.

We propose that rapid melt transport processes (Darcy flow on obtaining interconnection, and crack propagation on achieving extraction) short-circuit the diapir phase of melt transport and prevent it from developing and actually rising any appreciable distance. Hence, even though the equations of *Kerr and Lister* [1988] do not describe the whole system, the along arc spacing could be controlled by this early stage of the system. Another explanation of the spacing of a partially molten system might involve a later stage in the system when we have a situation as described by *Stevenson* [1988], where the spacing is controlled by the distance of the molten region from the nearest boundary, which is the base of the lithosphere or the free surface. To get a spacing of 70 km one would require a diapir of radius 8km from the work of *Kerr and Lister* [1988]. In their theory, closer spacing arises from thinner cylinders. This would tend to suggest that volcanos closer together would have smaller potential sources and either smaller or less active volcanic centers.

2.3.b.4 Other sources of buoyancy

Other sources of buoyancy that have not been explicitly addressed here include the intrinsic buoyancy of residue in the solid wedge (which is a function of the degree of melting) and the thermal buoyancy due to the thermal gradients. An indication of the potential contribution of these and other effects to the buoyancy balance is presented in table 2.2.

One can see that there are many potentially significant components contributing to the buoyancy of the subduction zone wedge. These figures suggest that if the hydration of the mantle wedge is as extensive as suggested in our previous discussion, it might be as significant as the combined buoyancy of melt and residue. One needs to consider the rheology of the surrounding matrix in deciding the efficiency of the various sources of buoyancy in generating flow reversal, and in

Table 2.2. SOURCES OF BUOYANCY

source of buoyancy	buoyancy per m along slab $\Delta\rho/\rho$ %	Comments
melt	0.4	f=2%,$\Delta\rho$ 0.7g/cm^{-3}
water	0.06	w=0.1%,$\Delta\rho$ 2.3g/cm^{-3}
hydrated mantle	0.5	h=10%,$\Delta\rho$ 0.14g/cm^{-3}
temperature	1.2	$\Delta T = 400°C$
mantle residue	0.1	F=10%,$\Delta\rho$ 0.04g/cm^{-3}
		F=degree of melting

the case of hydration the viscosity of the surrounding matrix is expected to be generally much higher than near the partially molten regions.

The residue is less dense than the fertile source rock and hence is a source of buoyancy. Given that it is lighter, it tends to stay near the surface and the source region. Therefore the lateral density gradients that it develops locally cannot be sharp though the total buoyancy could be large due to its extent. Thermal buoyancy is difficult to incorporate in this kinematic model since the major temperature contrast (slab to wedge) drives the subduction to some unknown extent [*Carlson*, 1983; *Davies*, 1984; *Richter*, 1977; and *Spence*, 1987]. Therefore the kinematic boundary conditions include it to an equally unknown extent. Since there are very large viscosity contrasts, the effects of the unaccounted temperature contrasts would be difficult to model numerically. The buoyancy of the residue will tend to reverse the flow, while the effect of the unaccounted thermal buoyancy is unknown. Ultimately the residue is carried by the induced flow down and out of the wedge corner and it is mixed back into the mantle by the global flow.

2.3.b.5 Need for time dependence

If flow reversal is allowed to develop to a steady state temperature field then it is actually colder than a model without flow reversal since there will be closed streamlines in the model with flow reversal. This is because the closed streamlines in the wedge corner are continually cooled by the downgoing slab. If on the other hand the mechanism is allowed to be time-dependent though, hotter temperatures can be periodically produced. To demonstrate the initial increase in temperature followed by the long term decrease due to a permanent constant buoyancy force, we ran the following model. We took the result of a steady state calculation of a model without any sources of buoyancy and used it as the initial conditions for a time dependent calculation using an explicit version of the implicit code. In the

model there was no decoupling and there was a buoyancy number of 10^6 extending over an area of approximately $100km^2$. From figure 2.14 we see that after 0.5Ma the temperatures are hotter than those due solely to corner flow but by 1.0Ma the temperatures are colder. Hence the flow should oscillate with a half period of $\approx$ 0.75Ma. This mechanism will develop a time dependence, since as material goes along a closed streamline in the wedge corner it will cool down each time it comes past newly subducted cold slab and produce less melt and hence the driving force for the reversal is reduced. Equally the fertility of the matrix decreases in periods of flow reversal since the same material continues to melt by adiabatic decompression during each loop. This also makes it harder for the reversal to persist, but as the reversal dies the infertile matrix is carried away from the source region by the induced flow and replaced with more fertile material. It is significant that the residue of the melting is lighter than fertile mantle and tends to remain behind in the wedge. As the process continues one can imagine the wedge becoming more refactory until it produces insufficient melt to give the system the necessary buoyancy to generate flow reversal. This process is illustrated in the cartoons of figure 2.15. Therefore we expect that time-averaged, the source region of SZM (before metasomatism by the mobile slab phase) is infertile, depleted and refractory compared to the MORB source region.

The lifetimes of individual volcanos are very short, up to 10Ma, but most are probably much shorter, around 100 kyrs-1Ma. *Ida* [1981] quotes a study in which *Tsunakawa* [1982] claims to have evidence for alternation in the stress field in the Japan arc at a time scale of a few Ma. There is also some estimate on the time scale for variations from the volumes of lavas erupted as a function of time e.g., *Kennet et al,* [1977] estimates a few Ma, as do *McBirney et al.*, [1974] from studies of the Cascades in Oregon.

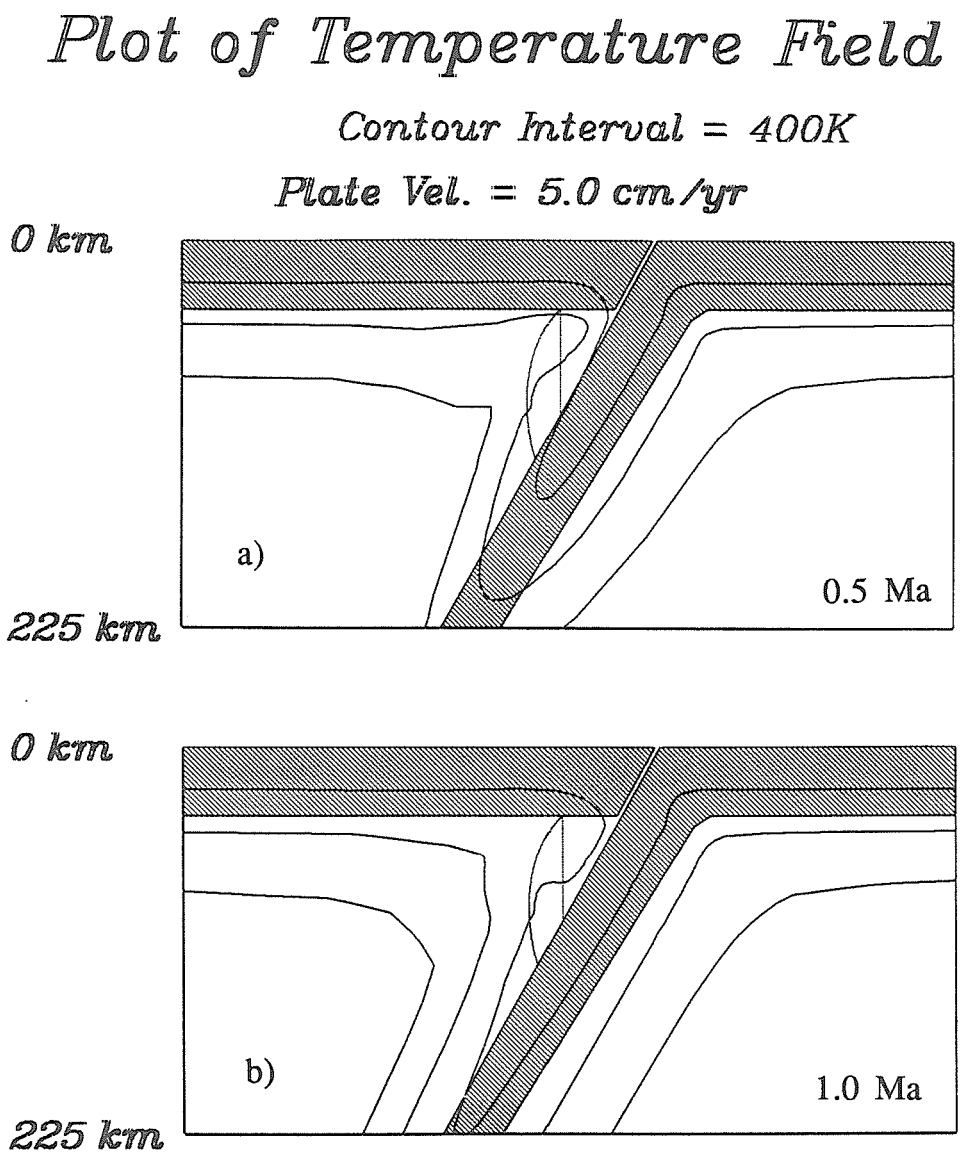


Figure 2.14 Temperature contours of models with localized source of buoyancy. They are at around 0.5Ma (a) and 1.0Ma (b) after the introduction of a localized source of buoyancy. The thermal field at 0Ma was the steady-state temperature field with no buoyancy force. Notice that the temperature increases until 0.5Ma but by 1.0Ma the mantle wedge is cooler.

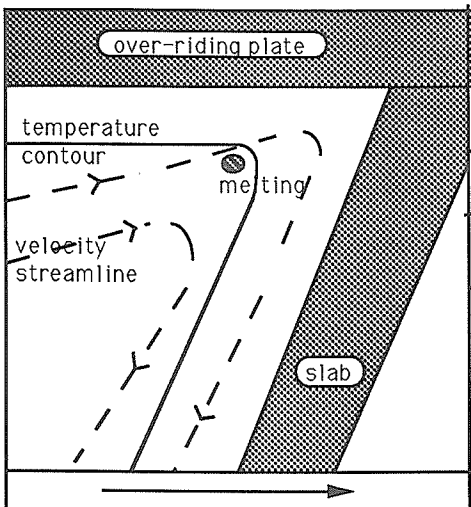


Figure 2.15a Cartoon of time cycle of model. Flow brings hot, fertile material into wedge corner; we have melting. Both melt and residue are buoyant.

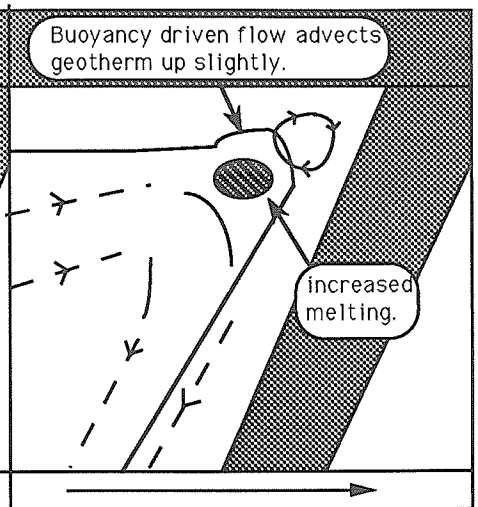


Figure 2.15b Melt and residue provide buoyancy which starts to locally reverse the flow, this raises the geotherms locally, and leads to a closed streamline.

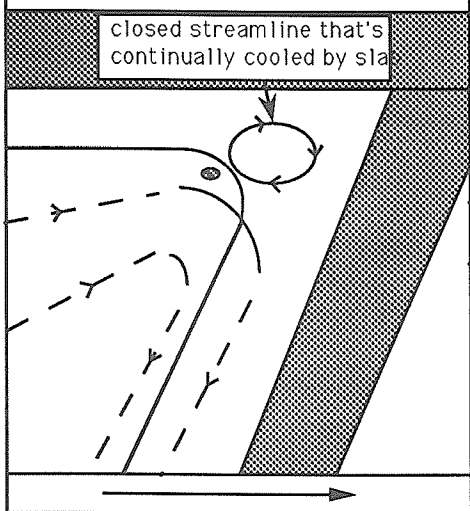


Figure 2.15c Corner gets cold and full of infertile material. Therefore melting decreases and the buoyancy force wanes.

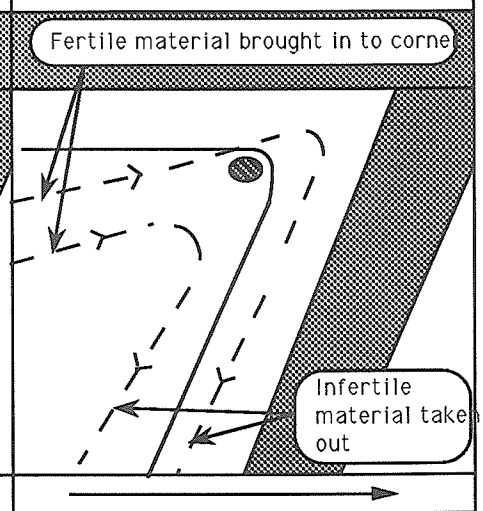


Figure 2.15d Plate driven flow becomes dominant, fertile material brought in and corner heated up. Melting increases and cycle repeats.

2.3.c Petrological and geochemical implications of model

The transport of a hydrous phase into the mantle wedge can potentially explain many of the distinctive features of SZV. The SZM have high water contents (2-5 wt%), hydrous phenocrysts including amphiboles, and high chlorine contents. Isotopically, SZM have higher $^{87}Sr/^{86}Sr$ than MORB, while the $^{143}Nd/^{144}Nd$ are similar (i.e., SZV are above the "mantle array"). Seawater has a high concentration of radiogenic Sr but negligible Nd. Hence, it is probable that the oceanic crust is enriched in radiogenic Sr. The hydrous phase will pass through any sediment subducted and will leach any component with a high solubility out from the sediment. $^{207}Pb/^{204}Pb$ versus $^{206}Pb/^{204}Pb$ is high compared to other oceanic basalts. Sediments have large Pb concentrations and high $^{207}Pb/^{204}Pb$ compared to $^{206}Pb/^{204}Pb$. ^{10}Be has only been found in SZV, but isn't ubiquitous. ^{10}Be has a short half-life and is generated solely on the surface of the Earth. Hence, its presence has been interpreted as a reflection of the recent input of sediment into the source. *Tatsumi and Isoyama* [1988] have demonstrated that ^{10}Be is highly soluble in water and could be leached from the sediment by a hydrous fluid dehydrating from the underlying basalt and migrating through the sediment and onto the source region. The $^{3}He/^{4}He$ are $\approx$ 6.5-7.0. and are slightly lower than MORB $\approx$ 8. They have generally been interpreted as due to addition of crustal ^{4}He, from down-going sediment or from assimilation in passage through crust.

Large Ion Lithophile Elements (LILE) are enriched relative to Rare Earth Elements (REE), while High Field Strength Elements (HFSE) are depleted relative to REE. Hence, the most diagnostic geochemical characteristic of arc magmas is $Ba/Th > 450$. The high LILE content can be attributed to fluids/melts from the slab carrying these elements in solution. It has been discovered that LILE are much more soluble in water at high pressure than HFSE [*Tatsumi et al.*,1986].

Island arc tholeiites are depleted in Light (L) REE while calc-alkaline SZV are relatively enriched in LREE and P. Hence the REE pattern of SZV is not unique, but *Davidson* [1987] pointed out the interesting observation that the REE as a whole are depleted in SZV relative to Mid-Ocean Ridge Basalts (MORB). This suggests either that the wedge is highly depleted or that SZV reflects a higher degree of melting. Rather than requiring a mechanism to decrease the HFSE, we instead propose that the mantle is originally depleted and is enriched with elements with high mobilities in the hydrous phase. The generally hydrous nature of the SZV and the high chlorine contents can also be explained by the incorporation of seawater.

The oxygen isotopes are generally $\delta^{18}O \approx 5.5\text{-}9.0$ ‰, with the higher values signifying interaction with material that has been recently in the near surface environment (sediments, continental crust or oceanic crust). The majority of the values are 5.5 - 7.0 ‰ i.e., similar to MORB. *Gregory and Taylor* [1981] discovered that the upper levels (pillow lavas and diabase dike complex) of the Oman Ophiolite had very high values $\delta^{18}O > 5.7$ ‰ and up to 12.0 ‰, while the lower gabbroic levels had low $\delta^{18}O$ of < 5.7 ‰ down to 4.0 ‰. Hence, if it is argued that the oceanic crust is extensively melted to provide the SZV source, one would expect the upper layers to be preferentially melted and for the $\delta^{18}O$ to be > 5.5 ‰. Since most values are similar to MORB, this is an argument against the slab being the sole source of the volcanics. For the hydrogen isotopes the δD values are generally between -40 and -80 ‰; spanning a similar range as Oceanic Island Basalt (OIB) and MORB. While the $\delta^{13}C$ vary from -2 to -12. The $\delta^{13}C$ has been interpreted as reflecting the depth of outgassing due to the pressure sensitive solubility of carbon in silicate melts [*Javoy et al.*,1989].

Other features of the data have been explained as resulting from the slower ascent to the surface compared to MORB. These include assimilation in passage

through crust and increased fractionation. A distinctive feature of SZV is the high normative silica of some of the volcanics with a tendency for volcanics above thicker crust to have higher normative silica, e.g., the andesitic volcanism of the Andes. Various suggestions have been advanced, including more silicic parent magmas due to unique conditions at the source (possibly hydrous, high fO_2, high alkalies), or from increased fractional crystallization as a result of slower ascent. Slower ascent could be a result of a) higher viscosity, (due to high normative silica [chicken and egg argument] and/or lower temperature), b) more compressive stress regime, or c) a density barrier e.g., a thick crust. Geochemically, SZM are characterized by suppression of Fe-enriching processes. The suppression of iron enrichment is thought to be due to the fractionation of magnetite [*Gill*,1983], and/or hornblende [*Grove and Baker*,1984]. Our model does not address these upper level processes, but the above suggestions are reasonable explanations for the above features.

In evaluating the water budget, we mentioned the probability that very water rich lavas do not reach the mechanical lithosphere but are swept back into the source region. This might explain the decoupling of incompatible elements from major elements required by the high concentration of incompatible elements in the crust relative to its total volume. The very hydrous melts will be the result of small degrees of melting, allowing fractionation of the more incompatible elements. This effect though is magnified since these melts can undergo multiple freeze-melting episodes before they enter melts that exit the convecting wedge to the mechanical lithosphere. *O'Nions and McKenzie* [1987] suggested that the decoupling could be achieved by extracting very small degrees of melting ($<1\%$) at the periphery of the melting zone i.e. extracting the more incompatible elements from a proportionally larger volume of the mantle compared to the major elements. It has been argued that such small degrees of melt can be extracted since

the dihedral angle of basalt with olivine is $<60°$; but the work of *Fujii et al.* [1986] and the electrical conductivity observation of *Sato et al.* [1988a] mentioned above suggests that there could be a threshold and it might be 2%. If it is true, then repeated small degrees of melting ($\approx$ 2%) in SZ could still accomplish the decoupling, whereas the $<1\%$ amounts of melt required to produce the same decoupling in one pass melting could not be extracted from the source region. This might contribute to the fact that SZV have higher levels of incompatible elements relative to MORB. The flow geometry in SZ is conducive to repeat melting of the same material while the diverging flow at ridges allows only single pass melting.

The flow reversal mechanism can allow for higher temperatures but not as high as required by petrology from analysis of primitive lavas, e.g., *Tatsumi et al.* [1983] (1400°C) and *Nye and Reid* [1986] (1600°C). These experiments are meaningful only if they have correctly identified the primary magma, and then only if the melting process can be thought of as batch equilibrium melting at a single temperature and pressure. Since the erupted lavas are most probably a reflection of an integration of melt generated at very different temperatures and depths, one suspects that it is very difficult to disentangle the source conditions. Since we argue that the most hydrous melts from the source region don't reach the surface and are not integrated into the "primitive" lavas of *Tatsumi et al.* [1983] and *Nye and Reid* [1986] we suspect the values quoted and would argue for lower values.

A lower bound on the temperature of magmas in the source region can be estimated from the eruption temperatures. The eruption temperatures for andesites range from 1050—1100°C [*Gill* 1981]; since the water saturated solidus is 900—1000°C this is a strong argument for melting at a higher temperature e.g., the amphibole buffered solidus. Eruption temperatures from MOR seem less variable and are $\approx$ 1200°C. Assuming a source depth of $\approx$ 80 km for SZM and

adiabatic temperature gradient for the magmas of 1°C/km, we estimate a minimum temperature at the source of 1130—1180°C. Given the prediction of *McKenzie and Bickle* [1988] that the MORB source region is at ≈1320°C, then they lose 80°C in addition to that due to adiabatic decompression (assumed source depth ≈ 40 km). SZM have a much harder path to the surface, hence SZM would be expected to lose at least this much to their surroundings i.e., a minimum of 1210—1270°C for SZM at their source. Temperatures predicted for our model are around 1200°C and due to its cyclical behavior will show higher temperatures during the later stages and in general should show a wider range of temperatures. These eruption temperature arguments are only suggestive since we do not know the path history of magmas from their source region to the surface, including how long they reside in magma chambers. The temperatures are buffered by crystallization; SZM contain more crystals than MORB.

There is further evidence for a depleted source from the observation of *Davidson* [1987] that SZV are depleted in REE relative to N-MORB and the observation of *Bonatti et al.* [1989] that the most depleted peridotites come from the SZ environment. Also *Falloon et al.* [1988], state that their experimental study suggests that the source of island arcs basalts is more refractory.

This mechanism suggests that the first melts produced in a young SZ, (before there has been sufficient subduction to cool down the wedge) will have very high water contents since the hotter fertile wedge will allow larger volumes of water saturated melts to be generated, increasing their probability of reaching the surface. They would be expected to appear closer to the trench. These might be equated with Boninites [*Dobson and O'Neil*,1987]. If *Otsuki* [1989] is incorrect and there are cases of slow relative convergence and we were to invoke the buoyancy flow reversal mechanism to explain the melting, then we would require that the source of buoyancy be deep and away from the slab. Unless the buoyancy pump

is placed in a hot region of the mantle we will not raise hot material since this mechanism does not develop an 'updraft' as discussed earlier.

During the reversal stage of the periodic cycle the water will still leave the slab and will be transported laterally by the mechanism described earlier, since it takes up only a small proportion of the wedge i.e., is localized to $\approx$ 1km width near the slab and $\approx$ 5-10 km in depth across the wedge (Appendix 2.B). Hence it is not much affected by the decoupling, which extends more than a few km into the wedge. Equally, it is not greatly affected by the reversal, since as mentioned a reversal does not develop much of an "updraft."

2.4 Conclusions

1. Induced corner flow cannot generate extensive partial melting of the subducted slab and hence produce high-Al basalts from an eclogitic source as proposed by *Brophy and Marsh* [1986].

2. Induced corner flow can laterally transport water large distances fixed in amphibole near the dehydration depth of amphibole in the mantle.

3. This water will only produce melting if the local relative velocity of wedge and plate is $\approx > 6 \pm 2$cm/yr.

4. Balancing water input into the SZ with volume and water content of SZM entering the crust leads us to suggest that much more magma is generated in the SZ but that it is very hydrous and is arrested in the lithospheric mantle. Some of these water-rich frozen melts might become diapirically unstable and produce tonalitic plutons and possibly the granites seen in continental convergent margins. The most hydrous magmas are the result of small degrees of melting. These will not reach the mechanical lithosphere, but will be swept back into the molten zone and undergo further melting. This leads to a process where the same material can be melted repeatedly allowing large fractionations and a decoupling of

incompatible elements from the major elements.

5. It is suggested that the magmas are transported to the base of the lithosphere by cracks whose direction of propagation is controlled by the least compressive stress. This leads to a mechanism that helps in focusing the melts at the volcanic front.

6. Due to the buoyancy of the depleted residue it is suggested that the fertility of the SZV source is variable, and that the time-averaged source is more depleted and refractory than the source of MORB since the residue is not cleared out of the source region as efficiently as in the MOR environment. Hence, for the major elements we prefer an N-MORB to an OIB source. The metasomatism due to the mobile phase from the slab will complicate the issue for minor incompatible elements.

7. The temperature field and melt production of a SZ will be modulated by a time-dependent flow reversal leading to slightly higher temperatures as well as time dependence with a period of $\approx$ 1.5Ma.

REFERENCES

Abbott D. H., and M. Lyle, Age of oceanic plates and volatile recycling, *Geophys. Res. Lett.*, **11**, 951-954, 1984.

Ahrens T. J., and G. Schubert, Gabbro-Eclogite Reaction Rate and Its Geophysical Significance. *Rev. Geophys. Space Phys.*, **13**, 1975,

Aines R. D., and G. R. Rossman, The water content of mantle garnets, *Geology*, **12**, 720-723, 1984.

Alt J. C., J. Honnorez, C. Laverne, and R. Emmermann, Hydrothermal Alteration of 1km section through the upper Oceanic crust, Deep Sea Drilling Project Hole 504B : Mineralogy, Chemistry and Evolution of Seawater-basalt Interactions, *J. Geophys. Res.*, **91**, 10309-10336, 1986.

Anderson, D. L., Thermally induced phase changes, lateral heterogeneity of the mantle, continental roots and deep slab anomalies, *J. Geophys. Res.*, **92**, 13968-13980, 1987.

Anderson, D. L., Temperature and Pressure derivatives of Elastic Constants with application to the Mantle. *J. Geophys. Res.*, **93**, 4688-4700, 1988.

Anderson, D. L., Where on Earth is the Crust? *Physics Today*, **42**, 38-46, 1989.

Anderson, D. L., and J. Bass, Transition region of the Earth's upper mantle, *Nature*, **320**, 321-328, 1986.

Anderson, O. L., E. Schreiber, R. C. Liebermann, and N. Soga, Some elastic constant data on Minerals relevant to Geophysics, *Rev. of Geophys.*, **6**, 491-498, 1968.

Anderson, R. N., S. Uyeda, and A. Miyashiro, Geophysical and Geochemical Constraints at Converging Plate Boundaries - Part 1. Dehydration in the Downgoing Slab, *Geophys. J. Roy. Astr. Soc.*, **44**, 333-357, 1976.

Anderson, R. N., S. E. Delong, and W. M. Schwarz, Thermal Model for

Subduction with dehydration in the downgoing slab, *J. Geol.*, **86**, 731-739, 1978.

Anderson, R. N., S. E. Delong and W. M. Schwarz, Dehydration, Asthenospheric Convection and Seismicity in Subduction Zones, *J. Geol.*, **88**, 445-451, 1980.

Andrews, D. J., and N. H. Sleep, Numerical Modelling of Tectonic Flow behind Island Arcs, *Geophys. J. Roy. Astr. Soc.*, **38**, 237-251, 1974.

Arculus, R. J., and R. Powell, Source Component Mixing in the regions of Arc Magma Generation, *J. Geophys. Res.* , **91**, 5913-5926, 1986.

Astiz, L., T. Lay and H. Kanamori, Large intermediate-depth earthquakes and the subduction process, *Phys. Earth Planet. Int.*, **53**, 80-166, 1988.

Bargen, N. von, and H. S. Waff, Permeabilities, Interfacial Areas, and Curvatures of Partially Molten Systems : Results of Numerical Computations of Equilibrium Microstructures, *Geophys. J. Roy. Astr. Soc.*, **91**, 9261-9276, 1986.

BVTP, Basaltic Volcanism Study Project, Basaltic Volcanism on the Terrestial Planets, *Pergamon*, New York, 1981.

Batchelor, G. K., An introduction to Fluid Dynamics, Cambridge University Press, Cambridge, 1967.

Becker, K., H. Sakai et al., Drilling deep into young oceanic crust at hole 504b, Costa Rica Rift, *Rev. Geophys.*, **27**, 79-102, 1989.

Bergman, E. A., and S. C. Solomon, Source mechanisms of earthquakes near Mid-ocean ridges from body waveform inversion : Implications for the early evolution of oceanic lithosphere, *J. Geophys. Res.* , **89**, 11415-11441, 1984.

Beukel, J. van den, and R. Wortel, Temperatures and shear stresses in the upper part of a Subduction Zone, *Geophys. Res. Lett.*, **14**, 1057-1060, 1987.

Bodine, J. H., M. S. Steckler and A. B. Watts, Observation of Flexure and the

Rheology of the Oceanic Lithosphere, *J. Geophys. Res.*,**86**,3695-3707,1981.

Bodri, L., and B. Bodri, Numerical investigation of tectonic flow in island-arc areas, *Tectonophysics*, **50**, 163-175, 1978.

Bohlen, S. R., and K. Mezger, Origin of Granulite Terranes and the formation of the lowermost continental crust, *Science*, **244**, 326-329, 1989.

Bonatti, E., and P. J. Michael, Mantle peridotites from continental rifts to ocean basins to subduction zones, *Earth Planet. Sci. Lett.*, **91**, 297-311, 1989.

Brophy, J. G., Basalt Convection and Plagioclase Retention: A model for the Generation of High-Alumina Arc Basalt, *J. Geol.*, **97**, 319-329, 1989.

Brophy, J. G., and B. D. Marsh, On the Origin of High-Alumina Arc Basalts and the Mechanics of Melt Extraction, *J. Petrol.*, **27**, 763-789, 1986.

Brown, G. C., and A. E. Mussett, *The Inaccessible Earth*, George Allen and Unwin, 1981.

Byrne, D. E., D. M. Davis and L. R. Sykes, Mechanics of the shallow region of subduction zones and the loci and maximum size of thrust Earthquakes, *submitted to Tectonics*, 1987.

Carlson, R. L., Plate motions, boundary forces, and horizontal temperature gradients: implications for the driving mechanism, *Tectonophysics*, **99**,149-164,1983.

Carslaw, H. S. and J. C. Jaeger, *Conduction of heat in solids*,(2^{nd} edition) Oxford University Press, Oxford, 1959.

Cathles, L. M., *The viscosity of the Earth's Mantle*, Princeton University Press, Princeton, N.J., 1975.

Cloos, M., Thermal evolution of convergent plate margins: thermal modeling and re-evaluation of isotopic Ar-ages for Blueschists in the Franciscan complex of California, *Tectonics*,**4**,421-438,1985.

Cooper, R. F., and D. L. Kohlstedt, Interfacial energies in the olivine-basalt system, *High Pressure Research in Geophysics, Adv. Earth Planet. Sci.*,**v12**, edited by S. Akimoto and M. H. Manghnani, Center for Academic Publications, Tokyo, 217-228, 1982.

Cox, K. G., Numerical Modeling of a Randomized RTF Magma Chamber : A Comparison with Continental Flood Basalt Sequences, *J. Petrol.* ,**29**,681-697,1988.

Craig, C. H., and McKenzie D. P., The existence of a thin low-viscosity layer beneath the lithosphere, *Earth Planet. Sci. Lett.*,**78**,420-426,1986.

Crawford, A. J., T. J. Falloon and S. Eggins, The origin of high-alumina basalts, *Contrib. Mineral. Petrol.* ,**97**,417-430,1987.

Creager, K. C., and T. H. Jordan, Slab penetration into the lower mantle beneath the Mariana and other island arcs of the northwest Pacific, *J. Geophys. Res.*, **91**, 3573-3580, 1986.

Crough, S. T., Thermal origin of mid-plate hot-spot swells. *Geophys. J. Roy. Astr. Soc.* ,**55**,451-469,1978.

Crough, S. T., and Thompson, Numerical and Approximate solutions for lithospheric thickening and thinning, *Earth Planet. Sci. Lett.*,**31**,397-402,1976.

Dahlen, F. A., and J. Suppe, Mechanics, growth, and erosion of mountain belts, *Geol. Soc. Am. Special Paper* , **218**, 1987.

Davidson, J. P., Crustal contamination versus subduction zone enrichment: Examples from the Lesser Antilles and implications for mantle source compositions of island arc volcanic rocks, *Geochimica et Cosmochimica Acta*, **51**, 2185-2198, 1987.

Davies, G. F., Geophysical and isotopic constraints on mantle convection: an interim synthesis, *J. Geophys. Res.* ,**89**,6017-6040,1984.

Davies, J. H., and D. J. Stevenson, Flow Direction of Mantle in Subduction Zone Wedge, *Trans. Am. Geophys. Un. EOS*, **69**, 1439, 1988.

DeBari, S. M., and R. G. Coleman, Examination of the deep levels of an island arc : evidence from the Tonsina Ultramafic-mafic assemblage, Tonsina, Alaska, *J. Geophys. Res.* ,**94**,4373-4391,1989.

DeBari, S. M., S. M. Kay and R. W. Kay, Temperatures from tectonized ultramafic xenoliths, *J. Geol.* ,**95**,329-341,1987.

Delany, J. M., and H. C. Helgeson, Calculation of the thermodynamic consequences of dehydration in subducting oceanic crust to 100kb and 780°C, *Am. J. Science* ,**278**,638-686,1978.

Detrick, R. S., and S. T. Crough, Island subsidence, hot spots and lithospheric thinning, *J. Geophys. Res.* ,**83**,1236-1244,1978.

Dmwoska, R., J. R. Rice, L. C. Lovison and D. Josell, Stress transfer and seismic phenomena in coupled subduction zones during the earthquake cycle, *J. Geophys. Res.*, **93**, 7869-7884, 1988.

Dobson, P. F., and J. R. O'Neil, Stable isotope compositions and water contents of boninite series volcanic rocks from Chichi-jima, Bonin Islands, Japan, *Earth Planet. Sci. Lett.* , **82**,75-86,1987.

Eggler, D. H., Influence of H_2O and CO_2 on melt and fluid chemistry in subduction zones, *Crust/Mantle Recycling at Convergence Zones* , Kluwer Academic Publishers, ed. S. R. Hart and L. Gulen, NATO ASI series. 97-104, 1987a.

Eggler, D. H., Solubility of Major and Trace Elements in Mantle Metasomatic Fluids: Experimental Constraints, *Mantle Metasomatism* , Academic Press,ed. M. A. Menzies and C. J. Hawkesworth,21-41,1987b.

Eggler, D. H., Water-saturated and undersaturated melting relations in a Paricutin andesite and an estimate of water content in the natural magma,

Cont. Mineral. Petrol. , **34**,261-271,1972.

Falloon, T. J., and D. H. Green, Glass inclusions in magnesian olivine phenocrysts from Tonga: evidence for highly refractory parental magmas in the Tongan Arc, *Earth Planet. Sci. Lett.*, **81**,95-103,1986.

Falloon, T. J., and D. H. Green, The solidus of carbonated, fertile peridotite. Submitted to *Earth Planet. Sci. Lett.*, 1989.

Falloon, T. J., D. H. Green, Hatton C. J. and Harris K. L., Anhydrous partial melting of a fertile and depleted peridotite from 2 to 30 kbar and apllication to basalt peterogenesis, *J. Petrol.*,**29**,1257-1282, 1988.

Fujii, N., K. Osamura and E. Takahashi, Effect of water Saturation on the distribution of partial melt in the olivine-pyroxene-plagioclase system, *J. Geophys. Res.* ,**91**,9253-9260,1986.

Fujisawa, H., N. Fujii, H. Mizutami, H. Kanamori and S. Akimoto, Thermal diffusivity of Mg_2SiO_4,Fe_2SiO_4, and NaCl at high pressure and temperature, *J. Geophys. Res.* ,**73**,4727-4733,1968.

Furlong, K. P., D. S. Chapman and P. W. Alfeld, Thermal modeling of the geometry of subduction with implications for the tectonics of the overriding plate, *J. Geophys. Res.*,**87**,1786-1802,1982.

Furukawa, Y. and S. Uyeda, Thermal state under the Tohoko Arc with consideration of crustal heat generation, *Tectonophysics*,**164**,175-187,1989.

Fyfe, W. S., N. J. Price and A. B. Thompson, *Fluids in The Earth's Crust*, Elsevier, Amsterdam, 1978.

Giardini, D. and J. H. Woodhouse, Deep seismicity and modes of deformation in Tonga subduction zone, *Nature*, **307**, 505-509, 1984.

Gill, J., Orogenic Andesites and Plate Tectonics, *Springer-Verlag* , , 1981.

Goto, K., H. Hamaguchi, and Z. Suzuki, Earthquake generating stresses in a descending slab, *Tectonophysics*, **112**, 111-128, 1985.

Green, D. H., Magmatic activity as the major process in the chemical evolution of the Earth's crust and mantle, *Tectonophysics*, **13**, 47-71, 1972.

Green, D. H., Contrasted melting relations in a Pyrolite upper mantle under mid-oceanic ridge, stable crust and island arc enviroments, *Tectonophysics*, **17**, 285-297, 1973.

Green, D. H., Experimental testing of "equilibrium" partial melting of peridotite under water-saturated, high pressure conditions, *Canad. Mineral.*, **14**, 255-268, 1976.

Green, D. H. and N. J. Pearson, Ti-rich accessory phase saturation in hydrous mafic-felsic compositions at high P, T. *Chem. Geol.* , **54**, 185-201, 1986.

Gregory, R. T., and H. Taylor, An Oxygen isotope profile in a section of Cretaceous Oceanic Crust, Samail Ophiolite, Oman : Evidence for $\delta^{18}O$ Buffering of the Oceans by Deep (>5km) Seawater-Hydrothermal Circulation at Mid-Ocean Ridges, *J. Geophys. Res.*, **86**, 2737-2755, 1981.

Griggs, D. T., The sinking lithosphere and the focal mechanism of deep earthquakes, *The nature of the Solid Earth* , , ed. Robertson E. C., J. F. Hays, and Knopoff L., McGraw-Hill, p361-384, 1972.

Grove, T. L., and M. B. Baker, Phase equilibrium controls on the tholeiitic versus calc-alkaline differentiation trends, *J. Geophys. Res.* , **89**, 3253-3274, 1984.

Grove, T. L., and R. J. Kinzler, Petrogenesis of Andesites, *Ann. Rev. Earth Planet. Sci.* , **14**, 417-454, 1986.

Gurnis, M., A Reassessment of the heat transport by variable viscosity convection with plates and lids, *Geophys. Res. Lett.* , **16**, 179-182, 1989.

Gust, D. A., and M. R. Perfit, Phase relations of a high-Mg basalt from the Aleutian island arcs: implications for primary island arc basalts and high-Al basalts, *Contrib. Mineral. Petrol.* , **97**, 7-18, 1987.

Haggerty, S. E., and L. A. Tompkins, Redox state of Earth's upper mantle from kimberlitic ilmenites, *Nature*, **303**, 295-300, 1983.

Hasebe, Ḱ., N. Fujii, and S. Uyeda, Thermal processes under Island Arcs, *Tectonophysics*, **10**, 335-355, 1970.

Heard, H., and W. Rubey, Tectonic Implications of Gypsum Dehydration, *Geol. Soc. Am. Bull.*, **77**, 741-760, 1966.

Hess, P. C., *Origins of Igneous Rocks.* Harvard University Press, Cambridge, Massachusetts, 1989.

Hildreth, W., and S. Moorbath, Crustal contribution to arc magmatism in the Andes of Central Chile, *Contrib. Mineral. Phys.*, **98**, 455-489, 1988.

Honda, S., Thermal Structure Beneath Tohoku, Northeast Japan - A Case Study for Understanding the detailed thermal structure of the subduction zone, *Tectonophysics* , **112**, 69-102, 1985.

Houseman, G., The dependence of convection planform on mode of heating, *Nature*, **332**, 346-349, 1988.

Hsui, A. T., B. D. Marsh and M. N. Toksoz, On the melting of the subducted oceanic crust: Effects of Subduction Induced mantle flow, *Tectonophysics* , **99**, 207-220, 1983.

Hughes, T. J. R., W. K. Liu and A. Brooks, Finite Element Analysis of Incompressible Viscous Flows by the Penalty Function Formulation, *J. Computational Phys.*, **30**, 1-60, 1979.

Ida, Y., Thermal and Mechanical Processes Producing Arc Volcanism and Back-Arc Spreading, *Arc Volcanism : Physics and Tectonics.*, , ed. D. Shimozuru and I. Yokoyama, Terra Scientific Publishing Company (TERRAPUB), Tokyo, 165-175, 1983.

Ida, Y., Convection in the mantle wedge above the slab and tectonic processes in subduction zones, *J. Geophys. Res.*, **88**, 7449-7456, 1983.

Ida, Y., Structure of the mantle wedge and volcanic activities in the island arcs, *High Pressure Research in Mineral Physics*, 473-480, 1987. editors Manghnani M. H. and Y. Syono. by Terra Scientific Publishing Company (TERRAPUB), Tokyo / American Geophysical Union, Washington, D. C., 1987.

Isacks, B. L., and M. Barazangi, Geometry of Benioff zones, lateral segmentation and downwards bending of the subducted lithosphere, *in Island Arcs, Deep Sea Trenches and Back-Arc Basins*, Maurice Ewing Ser. 1, edited by M. Talwani and W.C. Pitman III, 99-114, AGU, Washington, D.C., 1977.

Ito, E. I., D. M. Harris, and A. T. Anderson Jr., Alteration of Oceanic Crust and geologic cycling of chlorine and water, *Geochim. Cosmochim. Acta*, **47**, 1613-1624, 1983.

Jacques, A. L., and D. H. Green, Anhydrous melting of peridotite at 0-15kb pressure and the genesis of tholeiitic basalts, *Contrib. Mineral. Petrol.*,**73**,287-310, 1980.

James, D. E., Andean crustal and upper mantle structure, *J. Geophys. Res.*,**76**,3246-3271, 1971.

Jarrard, R. D., Relations Among Subduction Parameters, *Rev. Geophys.* ,**24**,217-284, 1986.

Javoy, M. F., F. Pineau and P. Agrinier, Volatiles and Stable isotopes in recycling, *Crust/Mantle Recycling at Convergence Zones* , Kluwer Academic Publishers, ed. S. R. Hart and L. Gulen, NATO ASI series. 121-138, 1987.

Jeanloz, R., and A. B. Thompson, Phase transitions and mantle discontinuities, *Rev. Geophys. Space Phys.*,**21**,51-74, 1983.

Jeanloz, R., and S. Morris, Temperature distribution in the crust and mantle, *Ann. Rev. Earth Planet. Sci.*,**14**,377-415, 1986.

Johnston, A. D., Anhydrous P-T phase relations of near-primary high alumina

basalt from the South Sandwich Islands, *Contrib. Mineral Petrol.*,**92**,368-382, 1986.

Kawakatsu, H., Double seismic zones; kinematics, *J. Geophys. Res.*, **91**, 4811-4825, 1986.

Kay, R. W., Volcanic Arc magmas : Implications of a melting-mixing model for element recycling in the crust-upper mantle system, *J. Geol.*, **88**, 497-522, 1980.

Kay, R. W., and S. M. Kay, Something to do with island arcs, *Crust/Mantle Recycling at Convergence Zones* , Kluwer Academic Publishers, ed. S. R. Hart and L. Gulen, NATO ASI series. 1987.

Kay, S. M., and R. W. Kay, Aleutian Magmas in space and time, Decade of North American Geology, *in press Geol. Soc. Am.*, , 1988.

Kennett, J. P., A. R. McBirney, and R. C. Thunnell, Episodes of Cenozoic Volcanism in the circum-Pacific region, *J. Volcanol. Geotherm. Res.*, **2**, 145-163, 1977.

Kerr, R. C., and J. R. Lister, Island Arc and mid-ocean ridge volcanism, modelled by diaparism from linear source region, *Earth Planet. Sci. Lett.*, **88**, 143-152, 1988.

Klein, E. M., and C. H. Langmuir, Global correlations of ocean ridge basalt chemistry with axial depth and crustal thickness, *J. Geophys. Res.* ,**92**,8089-8115, 1987.

Landau, L. D., and E. M. Lifshitz, Fluid Mechanics, Course of Theoretical Physics, Volume 6. *Pergamon Press*,Oxford,pp536,1959.

Langmuir, C. H., Geochemical consequences of *in situ* crystallization, *Nature*, **340**, 199-205, 1989.

Lister, J. R., Buoyancy-driven fluid fracture: The effects of material toughness and of low-viscosity precursors. *preprint*, 1989.

Lister, J. R., Buoyancy-driven fluid fracture: similarity solutions for the horizontal and vertical propagation of fluid-filled cracks, *preprint*, 1989.

Liu, L., Effects of H_2O on the phase behaviour of the forsterite-enstatite system at high pressures and temperatures and implications for the Earth, *Phys. Earth Planet. Inter.*,**49**,142-167, 1987.

Liu, L., Water, low-velocity zone and the descending lithosphere, *Tectonophysics*,**164**,41-48, 1989.

Maaloe, S., and T. S. Petersen, Petrogenesis of oceanic andesite, *J. Geophys. Res.* , **86**, 10273-10286, 1981.

Marsh, B. D., Island-Arc Volcanism, *Am. Scientist* ,**67**,161-172,1979.

Mattioli, G. S., M. B. Baker, M. J. Rutter, and E. M. Stolper, Upper mantle oxygen fugacity and its relationship to metasomatism, *preprint* ,,1989.

Mavko, G. M., Velocity and Attenuation in Partially Molten Rocks, *J. Geophys. Res.* ,**85**,5173-5189,1980.

McBirney, A. R., J. F. Sutter, H. R. Naslund, K. G. Sutton, and C. M. White, Episodic volcanism in the central Oregon Cascade range. *Geology*,**2**,585-589,1974.

McKenzie, D. P., Speculations on the consequences and causes of plate motions, *Geophys. J. Roy. Astr. Soc.*, **18**, 1-32, 1969.

McKenzie, D. P., The Generation and Compaction of Partially Molten Rock. *J. Petrol.* ,**25**,713-765,1984.

McKenzie, D. P., The extraction of magma from the crust and mantle. *Earth Planet. Sci. Lett.*,**74**,81-91,1985.

McKenzie, D. P., and M. J. Bickle, The Volume and Composition of melt generated by extension of the lithosphere, *J. Petrol.* ,**29**,625-680,1988.

Miller, G. H., G. R. Rossman, and G. E. Harlow, The Natural Occurence of Hydroxide in Olivine, *Phys. Chem. Minerals*,**14**,461-472, 1987.

Minear, J. W., and M. N. Toksoz, Thermal regime of a downgoing Slab and New Global Tectonics, *J. Geophys. Res.*, **75**, 1397-1419, 1970.

Morris, J. D., and S. R. Hart, Isotopic and incompatible element constraints on the genesis of island arc volcanics, *Geochim. Cosmochim. Acta.*,**47**,2015-2030,1983.

Morton, J. L., and N. H. Sleep, A mid-ocean ridge thermal model : Constraints on the volume of axial hydrothermal heat flux, *J. Geophys. Res.*,**90**,11345-11353,1985.

Murase, T., and I. Kushiro, Compressional Wave Velocity in Partially Molten Peridotite at High Pressures, *Carnegie Inst. Wash. Yr. Book*,**78**,559-562, 1979.

Murase, T., and I. Kushiro, Shear Wave Velocity in Partially Molten Peridotite at High Pressures, *Carnegie Inst. Wash. Yr. Book*,**79**,307-310,1980.

Mutter, J. C., W. R. Buck, and C. M. Zehnder, Convective partial melting 1. A model for the formation of thick basaltic sequences during the initiation of spreading, *J. Geophys. Res.* ,**93**,1031-1048,1988.

Myers, J. D., B. D. Marsh, and A. K. Sinha, Primitive high-Al basalts in Aleutians, *Contrib. Mineral. Petrol.* ,**91**,221-234,1985.

Mysen, B. O., and Boettcher A. L., Melting of a Hydrous Mantle: I. Phase Relations of Natural Peridotite at High Pressures and Temperatures with Controlled Activities of Water, Carbon Dioxide and Hydrogen, *J. Petrol.* ,**16**,520-548,1975.

Nagao, T., and S. Uyeda, Heat flow measurements in the northern part of Honshu, Northeast Japan using shallow holes, *Tectonophysics* ,**164**,301-314,1989.

Nakamura, Y., and I. Kushiro, Composition of gas phase in Mg_2SiO_4-SiO_2-H_2O at 15 kbar, *Carnegie Inst. Washington Yearbook* ,**73**,255-258,1974.

Nicholls, I. A., and A. E. Ringwood, Effect of Water on Olivine Stability in Tholeiites and the Production of Si O_2-Saturated Magmas in the Island Arc Environment, *J.Geol.*, **81**, 285-300, 1973.

Nicolas, A., A Melt extraction model based on structural studies in mantle peridotites, *J. Petrol.*,, **27**, 999-1022, 1986.

Nur, A., and Z. Ben-Avraham, Volcanic gaps due to oblique subduction of aseismic ridges, *Tectonophysics*, **99**, 355-362, 1983.

Nye, C. J., and M. R. Reid, Geochemistry of primary and least fractionated lavas from Okmok volcano, central Aleutians: Implications for arc magmagenesis, *J. Geophys. Res.*, **91**, 10271-10287, 1986.

Olafsson, M., and D. H. Eggler, Phase relations of amphibole, amphibole-carbonate, and phlogopite-carbonate peridotite: petrologic constraints on the asthenosphere, *Earth Planet. Sci. Lett.*,**64**, 305-315, 1983.

O'Hara, M. J., The importance of the "shape" of the melting regime during partial melting of the mantle, *Nature*,**314**,58-62,1985.

O'Nions, R. K., and D. P. McKenzie, Melting and continent Generation, *Earth Planet. Sci. Lett.* ,**90**,449-456,1988.

Otsuki, K., Empirical relationships among the convergence rate of plates, roll-back rate of trench axis and island-arc tectonics: "laws of convergence rate of plates," *Tectonphysics*,**159**,73-94,1989.

Oxburgh, E. R., and Turcotte D. L., The Physico-Chemical Behaviour of the descending Lithosphere, *Tectonphysics*,**32**,107-128,1976.

Peacock, S. M., Thermal effects of metamorphic fluids in subduction zones, *Geology*,**15**,1057-1060,1987.

Perfit, M. R., D. A. Gust, A. E. Bence, R. J. Arculus, and S. R. Taylor, Chemical characteristics of island-arc basalts : implications for mantle sources, *Chem. Geol.*,**30**,227-256,1980.

Plank, T., and Langmuir C. H., An Evaluation of the global variations in the major element chemistry of arc basalts, *Earth Planet. Sci. Lett.*,**90**,349-370,1988.

Poreda, R., and H. Craig, Helium isotope ratios in circum-Pacific volcanic arcs, *Nature* ,**338**,473-478,1989.

Presnall, D. C., and P. C. Bateman, Fusion relations in the system $NaAlSi_3O_8$–$CaAl_2O_8$–$KAlSi_3O_8$–SiO_2–H_2O and the generation of granitic magma in the Sierra Nevada batholith, *Geol. Soc. Am. Bull.*,**84**,3181-3202,1973.

Rabinowicz M. , B. Lago, and C. Froidevaux, Thermal transfer between the continental asthenosphere and the oceanic subducting lithosphere : its effect on subcontinental convevction, *J. Geophys. Res.*, **85**, 1839-1853, 1980.

Raleigh, and M. Paterson, Experimental deformation of Serpentinite and its tectonic implications, *J .Geophys. Res.*, **70**, 3965-3985, 1965.

Reymer, A., and G. Schubert, Phanerozoic addition rates to the continental crust and crustal growth, *Tectonics* , **3**, 63-77, 1984.

Richter, F. M., On the driving mechanism of plate tectonics, *Tectonophysics*, **38**, 61-88, 1977.

Rigden, S. M., T. J. Ahrens, and E. M. Stolper, Densities of liquid silicates at High Pressures, *Science*, **226**, 1071-1074, 1984.

Rigden, S. M., T. J. Ahrens, and E. M. Stolper, Shock Compression of Molten Silicate: Results for a model basaltic composition, *J. Geophys. Res.*, **93**, 367-382, 1988.

Ringwood, A. E., The petrological evolution of island arc systems, *J. Geol. Soc. London*,**130**,183-294,1974.

Rubin, K. H., and J. D. MacDougall, Ra-226 Excesses in mid-ocean ridge basalts and mantle melting, *Nature*,**335**,158-161,1988.

Ruff, L., and H. Kanamori, Seismic coupling and uncoupling at subduction zones, *Tectonophysics*,**99**,99-118,1987.

Ryabchikov, I. D., and A. L. Boettcher, Experimental evidence at high pressures for potassic metasomatism in the mantle of the Earth, *Am. Mineral.* ,**65**,915-919,1980.

Ryabchikov, I. D., W. Schreyer and K. Abraham, Compositions of aqueous fluids in equilibrium with pyroxenes and olivines at mantle pressures and temperatures, *Contrib. Mineral. Petrol.*,**79**,80-84,1982.

Rydelek, P. A., and I. S. Sacks, Asthenospheric viscosity inferred from correlated land-sea earthquakes in noth-east Japan, *Nature*,**336**,234-237,1988.

Saleeby, J. B., Tectonic significance of serpentinite mobility and ophiolitic melange. *Geol. Soc. Am.*,Special Paper **198**,153-168, 1984.

Saleeby, J. B., Progress in Tectonic and Petrogenetic Studies in an exposed cross-section of young (≈100Ma) Continental Crust, Southern Sierra Nevada, California, *in press. Exposed Cross Sections of the Continental Crust.*,, NATO Advanced Studies Institute, ed. M. H. Salisbury, D. Reidel Publishing Co., Dordrecht, Holland, 1989.

Sato, H., I. S. Sacks, T. Murase, G. Muncill, and H. Fukuyama, Attenuation of Compressional Waves in Peridotite Measured as a Function of Temperature at 200MPa, *Pageoph*,**128**,433-447, 1988a.

Sato, H., I. S. Sacks, T. Murase, and C. M. Scarfe, Thermal Structure of the Low Velocity Zone derivevd from Laboratory and Seismic Investigations, *Geophys. Res. Lett.*,**15**,1227-1230, 1988b.

Sato, H., I. S. Sacks, E. Takahashi, and C. M. Scarfe, Geotherms in the Pacific Ocean from laboratory and seismic attenuation studies, *Nature*,**336**,154-156, 1988c.

Sato, H., I. S. Sacks, and T. Murase, The use of laboratory velocity data for

estimating temperature and partial melt fraction in the low-velocity zone: comparison with heat flow and electrical conductivity studies, *J. Geophys. Res.*,**94**,5689-5704, 1989.

Sato, H., and I. S. Sacks, Anelasticity and Thermal Structure of the Oceanic Upper Mantle : Temperature Calibration with Heat Flow Data, *J. Geophys. Res.*,**94**,5705-5715, 1989.

Saunders, A. D., J. Tarney, and S. D. Weaver, Transverse geochemical variations across the Antartic peninsula: implications for the genesis of calc-alkaline magmas, *Earth Planet. Sci. Lett.*,**46**,344-360, 1980.

Schneider, M. E., and D. H. Eggler, Fluids in equilibrium with peridotite minerals: Implications for mantle metasomatism, *Geochim. Cosmochim. Acta*,**50**,711-724, 1986.

Scott, D. R., and D. J. Stevenson, A self-consistent model of melting, magma migration and buoyancy-driven circulation beneath mid-ocean ridges, *J. Geophys. Res.*, **94**, 2973-2988, 1989.

Sleep, N. H., Stress and Flow Beneath Island Arcs, *Geophys. J. Roy. astr. Soc.*,**42**,827-857, 1975.

Spence, D. A., and D. L. Turcotte, Magma-Driven Propagation of Cracks, *J. Geophys. Res.*, **90**, 575-580, 1985.

Spence, W., Slab pull and the seismotectonics of subducting lithosphere, *Rev. of Geophys.*, **25**, 55-69, 1987.

Spence, W., Sress origins and Earthquake Potential in Cascadia, *J. Geophys. Res.*, **94**, 3076-3088, 1989.

Spera, F. J., Dynamics of Translithospheric Migration of Metasomatic Fluid and Alkaline Magma, *Mantle Metasomatism* , Academic Press,ed. M. A. Menzies and C. J. Hawkesworth,1-20, 1987.

Spiegelman, M., and D. McKenzie, Simple 2-D models for melt extraction at

mid-ocean ridges and island arcs, *Earth Planet. Sci. Lett.*,**83**,137-152, 1987.

Staudacher, T., and C. J. Allegre, Recycling of Oceanic Crust and Sediments : the noble gas subduction barrier, *Earth Planet. Sci. Lett.*,**89**,173-183, 1988.

Stevenson, D. J., On the role of Surface Tension in the Migration of Melts and Fluids, *Geophys. Res. Lett.*,**13**,1149-1152, 1986.

Stevenson, D. J., Rayleigh-Taylor Instabilities in Partially Molten Rock, *Trans. Am. Geophys. Un. EOS*, **69**, 1404, 1988.

Stevenson, D. J., Spontaneous small-scale melt segregation in partial melts undergoing deformation, *Geophys. Res. Lett.*, **16**, 1067-1070, 1989.

Stolper, E., A phase diagram for Mid-Ocean Ridge Basalts: Preliminary Results and Implications for Peterogenesis, *Contrib. Mineral. Petrol.*,**74**,13-27, 1980.

Suyehiro, K., and I. S. Sacks, P- and S-wave velocity anomalies asociated with subducting lithosphere determined from travel-time residuals in the Japan region, *Bull. Seismol. Soc. Am.*,**69**,97-114, 1979.

Sydora, L. J., F. W. Jones, and R. St. J. Lambert, The thermal regime of the descending lithosphere : The effect of varying angle and rate of subduction, *Can. J. Earth Sci.*,**15**,626-641, 1978.

Sykes, M. L., and J. R. Holloway, Evolution of granitic magmas during ascent: A phase equilibrium model, *Magamtic Processes: Physiochemical Principles*, The Geochemical Society, Special Publication No. 1. , Ed. B. O. Mysen, 447-461, 1987.

Tatsumi, Y., Origin of high-magnesium andesites in the Setouchi volcanic belt, southwest Japan. II: Melting phase relationship at high pressures, *Earth Planet. Sci. Lett.*,**60**,305-317, 1982.

Tatsumi, Y., Migration of Fluid Phases and Genesis of Basalt Magmas in Subduction Zones, *J. Geophys. Res.* ,**94**,4697-4707, 1989.

Tatsumi, Y., Formation of the Volcanic front in Subduction Zones, *Geophys. Res. Lett.*,**13**,717-720, 1986.

Tatsumi, Y., and H. Isoyama, Transportation of Beryllium with H_2O at high pressures: implications for magma genesis in subduction zones, *Geophys. Res. Lett.*,**15**,180-183, 1988.

Tatsumi, Y., D. L. Hamilton, and R. W. Nesbitt, Chemical Characteristics of Fluid Phase released from subducted lithosphere and origin of Arc Magmas: Evidence from high-pressure experiments and natural rocks, *J. Volcanol. and Geotherm. Res.* ,**29**,293-309, 1986.

Tatsumi, Y., M. Sakuyama, H. Fukuyama, and I. Kushiro, Generation of Arc Basalt Magmas and Thermal Structure of the Mantle Wedge in Subduction Zones, *J. Geophys. Res.* ,**88**,5815-5825, 1983.

Tatsumi, Y., M. Sakuyama, H. Fukuyama, and I. Kushiro, Generation of Arc Basalt Magmas and Thermal Structure of the Mantle Wedge in Subduction Zones, *J. Geophys. Res.* ,**88**,5815-5825, 1983.

Tera, F., L. Brown, J. Morris, I. S. Sacks, J. Klein, and R. Middleton, Sediment incorporation in island-arc magmas: inferences from Be^{10}, *Geochim. Cosmochim. Acta*, **50**, 535-550, 1986.

Thatcher, W., T. Matsuda, T. Kato, J. B. Rundle, Lithospheric Loading by the 1896 Riku-u Earthquake, Northern Japan : Implications for Plate Flexure and Asthenospheric Rheology, *J. Geophys. Res.* ,**85**,6429-6435, 1980.

Tichelaar, B. W., and Ruff, L., Variability in depth of seismic coupling along the Chilean Subduction Zone, *Trans. Am. Geophys. Un. EOS* ,**70**,398, 1989.

Toksoz, M. N., and A. T. Hsui, Numerical studies of back-arc convection and the formation of marginal basins, *Tectonophys.*, **50**, 177-196, 1978.

Toksoz, M. N., and A. T. Hsui, A review of the thermo-mechanical structures

at convergent plate boundaries and their implications for crust/mantle recycling, *Crust/Mantle Recycling at Convergence Zones* , Kluwer Academic Publishers, ed. S. R. Hart and L. Gulen, NATO ASI series. 75-80, 1987.

Toksoz, N., J. W. Minear, and B. R. Julian, Temperature field and geophysical effects of a down-going slab, *J. Geophys. Res.*, **76**, 1113-1138, 1971.

Toramaru, A., and N. Fujii, Connectivity of Melt Phase in a Partially Molten Peridotite, *J. Geophys. Res.* ,**91**,9239-9252, 1986.

Tovish, A., G. Schubert, and B. P. Luyendyk, Mantle Flow Pressure and the angle of Subduction : Non-Newtonian Corner Flows, *J. Geophys. Res.* ,**83**,5892-5898, 1978.

Tritton, D. J., Physical Fluid Dynamics, *Van Nostrand Reinhold Company*,,New York, 1977.

Tsunakawa, H., Ancient Stress Field in Japan Based on K-Ar ages of Neogene Dikes and boninite, Ph. D. thesis, University of Tokyo, 1982.

Tsunakawa, H., Simple two-dimensional model of propagation of magma-filled cracks, *J. Volcanol. Geotherm. Res.*,**16**,335-343, 1983.

Turcotte, D., Physics of magma segregation processes. *Magmatic Processes: Physiochemical Principles*,Editor B. O. Mysen, The Geochemical Society, Special Publication No. 1,69-74,1987.

Turcotte, D. L., A heat pipe Mechanism for Volcanism and Tectonics on Venus. *J. Geophys. Res.*,**94**,2779-2785, 1989.

Turcotte, D., and G. Schubert, A Fluid theory for the deep structure of Dip-Slip Fault Zones, *Phys. Earth Planet. Int.* ,**1**,381-386, 1968.

Turcotte, D., and G. Schubert, Frictional Heating of the descending Lithosphere, *J. Geophys. Res.* ,**78**,5876-5886, 1973.

Turcotte, D., and G. Schubert, Geodynamics: Applications of continuum physics to geological problems,

John Wiley and Sons., 1982.

Uyeda, S., Subduction Zones : An introduction to comparitive Subductology, *Tectonophysics.*, **81**, 133-159, 1982.

Van der Molen, I., and M. S. Paterson, Experimental deformation of partially molten granite, *Contrib. Mineral. Petrol.*,, **70**, 299-318, 1979.

Vassiliou, M. S., B. H. Hager, and A. Raefsky, The distribution of earthquakes with depth and stress in subducting slabs, *J. Geodynamics*, **1**, 11-28, 1984.

Vidale, J. E., Waveform effects of a high-velocity subducted slab, *Geophys. Res. Lett.*, **14**, 542-545, 1987.

Vrolijk, P., G. Myers, and J.C. Moore, Warm fluid migration along tectonic melanges in the Kodiak accretionary complex, Alaska. *J. Geophys. Res.*, **93**, 10313-10324, 1988.

Waff, H. S., and J. R. Bulau, Equilibrium fluid distribution in an ultramafic partial melt under hydrostatic stress conditions, *J. Geophys. Res.*, **84**, 6109-6114, 1979.

Waff, H. S., and J. R. Bulau, Experimental determination of near-equilibrium textures in partially molten silicates at high pressures, *High Pressure Research in Geophysics, Adv. Earth Planet. Sci.*, **v12**, edited by S.Akimoto and M. H. Manghnani,pp229-236, Center for Academic Publications, Tokyo, 1982.

Wallace, M. E., and Green, D. H., An experimental determination of primary carbonatite composition, *Nature*, **335**, 343-345, 1988.

Watson, B. E., J. M. Brenan, and D. R. Baker, Distribution of Fluids, *Continental Lithospheric Mantle.*, in press, 1989.

Watson, B. E., and J. M. Brenan, Fluids in the Lithosphere, Part 1 : Experimentally-determined wetting characteristics of CO_2 - H_2O fluids and their implications for Fluid Transport, Host-Rock Physical Properties and

Fluid Inclusion Formation, *Earth Planet. Sci. Lett.*, **85**, 497-515, 1987.

Watts, A. B., J. H. Bodine, and M. S. Steckler, Observations of Flexure and State of Stress in Oceanic Lithosphere, *J. Geophys. Res.* ,**85**,6369-6376, 1980.

Wdowinski, S., R. J. O'Connell, and P. England, Continuum Models of continental deformation above subduction zones: Application to the Andes and the Aegean, *J. Geophys. Res.*, **94**,10331-10346, 1989.

Weidner, D. J., and E. Ito, Mineral Physics Constraints on a uniform mantle composition, *High Pressure Research in Mineral Physics*, Edited by Manghnani M. H. and Y. Syono. Terra Scientific Publishing Company, Tokyo / American Geophysical Union, Washington D.C., 439-446, 1987.

Wiens, D. A., and S. Stein, Age dependence of oceanic intraplate seismicity and implications for lithospheric evolution. *J. Geophys. Res.*, **88**, 6455-6468, 1983.

Wiens, D. A., and S. Stein, Intraplate seismicity and stresses in young oceanic lithosphere, *J. Geophys. Res.*, **89**, 11442-11464, 1984.

Williams, R. W., and J. B. Gill, Effects of partial melting on the uranium decay series, *Geochim. Cosmochim. Acta.*, **53**, 1607-1619, 1989.

Wyllie, P. J., Magmas and volatile components, *Am. Mineralogist,*,**64**,469-500,1979.

Wyllie, P. J., Subduction products according to experimental prediction, *Geol. Soc. Am. Bull.*,,**93**,468-476,1982.

Wyllie, P. J., Magma Genesis, Plate Tectonics, and Chemical Differentiation, *Rev. Geophys.* ,**26**,370-404,1988.

Wyllie, P. J., M. R. Carroll, A. D. Johnston, M. J. Rutter, T. Sekine and S. R. Van Der Laan, Interactions among magmas and rocks in subduction zone regions: experimental studies from slab to mantle to crust, *Eur. J.*

Mineral.,,1,165-179, 1989.

Yoder, H. S., *Generation of Basaltic Magma,* Washington D. C.: National Academy of Sciences. 1968.

Yuen, D. A., L. Fleitout, G. Schubert and C. Froidevaux, Shear Deformation zones along major transform faults and subducting slabs. *Geophys. J. R. Astr. Soc.*,,54,93-119, 1978.

Yuen, D. A., and G. Schubert, On the stability field of frictionally heated shear flows in the asthenosphere, *Geophys. J. R. Astr. Soc.*,,57,189-207, 1979.

Zhou, H-W., D. L. Anderson, and R. W. Clayton, Modeling of residual spheres for subduction zone earthquakes in the northwest Pacific, submitted to *J. Geophys. Res.*, 1988.

Appendix 2.A

ANALYTIC APPROXIMATIONS

Thinnest thickness of thermal boundary layer. (Initial thickness).

The initial thickness of the thermal boundary layer in the mantle occurs at the point where the flow begins to descend near the wedge corner. Assuming that this thickness is a balance between horizontal advection into the corner and horizontal conduction into the slab we get

$$v_x \partial T/\partial x = \kappa \partial^2 T/\partial y^2 \qquad \text{A1}$$

If we assume that v_x is constant and equal to the plate velocity ≈ 7 cm/yr, then the initial thickness is estimated to be around 500m. A more reasonable assumption is that the flow decreases as it approaches the corner, if we assume that it decreases linearly from the plate velocity to zero over 5km then we find that an estimate of the thickness is $\approx$ 5km.

Analytic Model

The model considers a slab descending at a constant velocity into a wedge at an angle θ. The x axis is perpendicular to the slab and the y axis is parallel to the slab. The flow is assumed to be parallel to the slab and equal to the velocity of the slab everywhere. We solve for the steady-state thermal regime, and the temperature gradient is assumed to be much stronger perpendicular to the slab than parallel to it. Hence we are left to solve

$$v \partial T/\partial y = \kappa \partial^2 T/\partial x^2 \qquad \text{A2}$$

We solve the above equation for the following initial condition at the surface, y = 0; where the temperature in the mantle away from the slab is constant at

T_1, while there is a linear gradient across the slab from T = T_o for the slab upper surface to T = T_1 at its lower surface. We find that the solution is

$$T = T_1 + (T_1 - T_o)\left(\frac{(x-h)}{2h}[erf\,(b) - erf\,(a)] - \alpha\pi^{1/2}\frac{1}{h}[\exp(-b^2) - \exp(-a^2)\right) \quad A3$$

where $a = -\frac{x}{\alpha}$; $b = -\frac{(x-h)}{\alpha}$; $\alpha = \left(\frac{4\kappa y}{v}\right)^{1/2}$; T_1, is the mantle temperature, T_o is the temperature of the upper surface of the plate as it enters the asthenosphere, and h is the width of the slab.

Appendix 2.B

TRANSPORT OF WATER

Assume that water leaves the slab between the depths of 50 and 80km and enters the mantle, where it immediately reacts with the peridotite to form amphibole. This amphibole remains stable until it reaches a depth of $\approx$ 100km, where it breaks down and releases its water. It is argued that at this high pressure the water forms an interconnected network with the hydrated peridotite [*Watson et al.*, 1989]. Therefore it rises vertically by porous flow through this hydrated mantle until it reaches virgin mantle. Then it gets carried back down and further away from the slab. The net result of the two processes is a horizontal flux of water away from the slab.

We can evaluate the extent of the region that is hydrated and also the volume of free water by equating fluxes. The extent of the hydrated region and the volume of free water will vary as we move further away from the slab. We are interested in these values as the water reaches the potential source region so we shall use the values applicable to this region. The variables with the largest uncertainty are the horizontal component of the mantle wedge velocity and the viscosity of the water.

The horizontal flux carried by the mantle wedge in the amphibole and the free water must equal the flux out of the plate that enters the mantle wedge and also the amount of water that gets carried away by the magma, assuming that most of the volatiles return to the surface. Hence

$$\rho_m V_x \left(w + \left(f \rho_w / \rho_m \right) \right) h = F \qquad \text{B1}$$

Where

V_x is the horizontal component of the mantle wedge flow.

w is the weight fraction of H_2O in the hydrated peridotite.

f is the volume fraction of H_2O as a free super-critical vapor.

ρ_w is the density of the free super-critical vapor.

ρ_m is the density of the mantle wedge.

h is the depth over which the mantle is hydrated and free water is present.

F is the mass flux of water per unit length along arc through the system.

Also the vertical flux in the mantle wedge of water is zero, i.e. the amount of water carried down in the fixed amphibole must equal the volume of water carried up by the free mobile phase. Using the expression for the relationship between porosity and permeability, from *McKenzie* [1984], $k \alpha f^3$, we get

$$\rho_m V_z \left(w + \left(f \rho_w / \rho_m \right) \right) = \rho_w a^2 f^3 \Delta\rho g / b \eta \qquad \text{B2}$$

Where

V_z is the vertical component of the mantle wedge flow.

a is the grain size of the mantle wedge.

b is a constant in the permeability law $k = a^2 f^3 / b$, $\approx 10^3$.

$\Delta\rho$ is $\rho_m - \rho_w$.

g is the acceleration due to gravity.

η is the viscosity of the super-critical fluid.

If we know the composition of the mantle then we can predict what the value of w is. *Green* [1972] and *BVTP* [1981] estimated that w might be 0.4wt%. Here we are assuming that we are in the stability field of amphibole and that none of the limiting components of amphibole (Na,Ca,Al) are transported in by the hydrous fluid. As discussed in the main text it is highly probable that the fluid will have a substantial amount of dissolved alkalies. *Eggler* [1987] has mentioned that up to 20 wt% K could be dissolved. If it is similar for Na, and since there is one water

molecule for every Na atom, then we can expect a 10 - 20 wt% increase in the amount of amphibole formed, if Na was the limiting factor (actually it is just as probable to be limited by lack of Al and Ca). Hence 0.5wt% seems a reasonable upper bound for the amount of water in a wedge peridotite saturated with amphibole. Estimates of the viscosity of water at these temperatures and pressures e.g. *Spera* [1987], and *Ahrens and Schubert* [1975] range from 7 10^{-3} to 2 10^{-4}Pas. If the typical grain size is 10^{-3}m, then we can solve for f given that $\Delta\rho \approx 2.5$ 10^{3}kmm^{-3}, and $V_z \approx 10^{-9}$ms^{-1} (3cm/yr). We can ignore the free water component being carried down by the solid matrix to first order.

$$f = (V_z \eta w \rho_m / \rho_w a^2 C \Delta\rho g)^{1/3} \approx 2\ 10^{-3} \qquad \text{B3}$$

If we had used the permeability/porosity relationship of *von Bargen and Waff* [1986], k = 10^{-4} a^2 f^2 then we would have found f << 2 10^{-3}. N.B. at these low values of porosity this framework breaks down since the tubules forming the network would no longer be big enough to display fluid properties, but the conclusion that the system needs very little porosity to carry the upward flux to balance the downward flux carried in the hydrated matrix remains unchanged.

Since we now have an estimate of w, f, and V_z, given an estimate of F we can predict the extent of this mantle hydration, h, from equation 1. Various people have estimated the water content of the slab to be up to 3 wt%; many workers consider 2 wt% a generous maximum [*Peacock*, 1988; *Ito et al.*, 1983], while others estimate ≈ 3 wt% [*Anderson et al.*, 1976; *Fyfe et al.*, 1978]. If 10 % of this water is released between 50 and 80 km and enters the wedge, then 0.2% is a reasonable estimate of the water content of the slab that enters the wedge. This is assuming that the volatiles can escape during the metamorphism from greenschist to amphibolite facies and that the thrust zone is not a "tight" fault in the oil industry parlance, both are unlikely and hence this is a large underestimate. Then the flux F is

$$F = w_s \; V_z \; h_c \; \rho_p \qquad \text{B4}$$

where

w_s is the weight fraction H_2O in the slab that enters the wedge ($2\ 10^{-3}$).

V_z is the vertical velocity of the slab. $2\ 10^{-9} \mathrm{m} s^{-1}$(6cm/yr)

h_c is the thickness of the oceanic crust. 6 km.

ρ_p is the density of the subducting plate. $3.5\ 10^3\ \mathrm{kg} m^{-3}$.

We have assumed that most of the dehydration is caused by the amphibole going beyond its high pressure stability limit, so as the slab descends 6km in 0.1Ma, 6 km of slab is dehydrated in this time. Therefore we estimate F $\approx 4\ 10^{-5} \mathrm{kg} s^{-1}$ per m along arc.

If we believe that most of the water subducted returns to the surface in SZV then we can also estimate F. An estimate of lava production is found in the corner flow section, 2.5-5 $10^4 \mathrm{kg} yr^{-1}$ per m along the arc. It is estimated that SZV close to the surface have 2-5wt% H_2O, then this leads to an estimate of F $\approx$ 1.6-6 $10^{-6} \mathrm{kg} s^{-1}$ per m along arc. If these melts were the result of 10% degree of melting and 50% crystallization then we evaluate that the initial melts at the source had 0.1 wt % H_2O in their source. Note that the two estimates don't agree by an order of magnitude. This could be due to the fact that part of the water is returned to the deep mantle to be sampled at MOR and in OIB after a long residence time. It could also reflect the uncertainty in estimating the amount of melt produced and also the amount of water that the slab injects into the mantle wedge. *Becker et al.* [1989] report on the drilling of hole 504B by ODP (deepest penetration into oceanic crust by ODP hole). The initial results from the neutron activation tool indicate that alteration products are concentrated along contacts between relatively unaltered units of fairly homogeneous geochemistry. The

deepest 200m of sheeted dikes (1.5km below the surface) were only slightly altered while the alteration increased as one approached the surface. This has been suggested to be controlled by permeability. These results seem to suggest that the alteration is nowhere pervasive and that it is largely restricted to the upper 1.2km. This suggests the presence of possibly less H_2O than has been previously estimated. The hole did not reach the gabbros, but if the permeability argument is correct then they should be altered even less than the sheeted dikes. Most probably excess water is released by the slab relative to the water returned by the lavas to the crust. As discussed in the text, a probable resolution of the discrepancy involves the incorporation of water in melts that do not make it to the crust, but are ponded at the Moho; they enter the crust later as tonalites, etc. From the two estimates of water flux we suggest that the thickness of the hydrated mantle will be 0.5 to 5 km, if the horizontal component of the velocity is 1cm/yr. Hence, given either estimate we get reasonable degrees of hydration.

Equally this flux needs to be carried down from the slab to the 100km dehydration depth of mantle in the wedge. If we assume that the water is produced uniformly throughout the depth from 50 to 80 km then we estimate the thickness of hydrated layer parallel to the slab to be 0.1-1 km at a depth of 80km where the amphibole dehydration is complete.

Appendix 2.C

WATER INDUCED MELTING

We shall briefly discuss some problems arising from water induced melting. First, we shall assume that water is completely incompatible and all of it enters the melt. It is widely believed to have a distribution coefficient similar to K, i.e. ≈ 0.01. Hence, the only way water can go through the super-solidus region is in the melt phase. If the system is in steady state then this requires that the water flux is constant through the system. The product of the solubility of water in the melt and the melt flux has to equal the water flux.

$$F = wa^2 f^3 \Delta\rho g / b\eta \qquad \text{C1}$$

We can solve for the melt porosity cubed times the solubility. Equally we can limit the melt porosity since we need so much melt to generate the arc. Hence

$$F_m = a^2 f^3 \Delta\rho g / b\eta \qquad \text{C2}$$

Where F_m is the melt flux.

Using the estimates of $1 km^3$ per year for the whole island arc system (40,000km long) we get an estimate of f ≈ 0.08 if $\eta \approx 10$Pas. The final melt that leaves the system most probably can have no more than 1 wt % H_2O (2 wt % extreme), but of course lower down in the melting column the amount of melt might be lower and the amount of water higher. The very first melts will be saturated because we have an infinitesimal ($5\ 10^{-5}$%) amount of free water. At this depth the solubility of water in melt is ≈ 25 wt %. Hence we will produce only $\approx 2\ 10^{-4}$melt! The next melting will occur on the breakdown of a pargasitic amphibole. This will be at $\approx 1150^o$ C, and the melts will be undersaturated in water.

If we consider a time development of this system then we might better understand its possible behavior as a continuous steady state system. We get a very small volume of melt produced at the water saturated solidus, and as it migrates away it will remain molten since the matrix is hotter above it. It will lead to additional melting until the water content is lowered to the point that the liquidus of the melt is the same as the ambient temperature. But most of the melt is generated further into the wedge at the breakdown of amphibole. This material has a water content of 0.4-0.5 wt % H_2O, and is moving nearly parallel to the isotherms. The melt produced, though, migrates vertically and similar to the very small degree of melting initially produces further melting higher up until the water is diluted in the melt. Using such expressions in detail are difficult in this system since it isn't closed and we have both a flux of water and a flux of solid material.

Appendix 2.D

DIRECTION OF PROPAGATION OF FRACTURES

The direction of propagation of fractures is controlled by the least compressive stress. The direction of opening of the crack will be in this direction while the direction of propagation will always maintain this relationship. Conditions where this might not be true are mentioned in the text, and we caution that a full solution to the problem of a finite crack propagating in a non-uniformly stressed medium has not been solved. Corner flow can be solved exactly (see *Batchelor* , [1967]). The following derivation of the corner flow equations follows closely the derivation due to *McKenzie* [1969]. If Ψ is the stream function and hence the velocity is given by

$$(v_r, v_\theta) = \left(\frac{1}{r}\frac{\partial \Psi}{\partial \theta}, -\frac{\partial \Psi}{\partial r}\right) \qquad \text{D1}$$

then for the corner flow problem illustrated in figure 2.D.1 with an acute angle θ_b we find that

$$\Psi = \frac{rv\left[(\theta_b - \theta)\sin(\theta_b) - \theta_b\,\theta\sin(\theta_b - \theta)\right]}{\theta_b^2 - \sin^2\theta_b} \qquad \text{D2}$$

Similarly, given the velocity field and a constant viscosity rheology one can solve for the stress regime. The deviatoric stresses in general are as follows

$$\sigma_{rr} = 2\eta\frac{\partial v_r}{\partial r} \qquad \text{D3a}$$

$$\sigma_{\theta\theta} = 2\eta\left(\frac{1}{r}\frac{\partial v_\theta}{\partial \theta} + \frac{v_r}{r}\right) \qquad \text{D3b}$$

$$\sigma_{r\theta} = \eta\left(\frac{1}{r}\frac{\partial v_r}{\partial \theta} + \frac{\partial v_\theta}{\partial r} - \frac{v_\theta}{r}\right) \qquad \text{D3c}$$

For this specific case we find

$$\sigma_{rr}=\sigma_{\theta\theta}=0 \qquad \text{D4}$$

and

$$\sigma_{\theta r}=\frac{2v\eta[\theta_b\cos(\theta_b-\theta)-\sin\theta_b\cos\theta]}{r(\theta_b^2-\sin^2\theta_b)} \qquad \text{D4}$$

Hence we can see that the greatest shear stresses are parallel to r = constant, and the θ = constant directions. In general the axes that maximize the shear stress lie at 45^o to the principal axes. In the co-ordinates of the principal axes the shear stress is zero. We have used the sign convention that compression is +ve while dilatation is -ve. Hence the axes of maximum and minimum compression are at 45^o with the direction of least compressive stress oriented as shown in figure 2.D.2.

What is the direction of propagation of the fracture? We know the instantaneous direction of the dyke at each position and hence we need to define the equation of the path. From figure 2.D.3 we find that

$$dr/d\theta=r \qquad \text{D5}$$

hence

$$r=r_o\,exp\,[\theta-\theta_o] \qquad \text{D6}$$

a logarithmic spiral. In figure 2.D.4 we draw some representative examples of propagation directions for the ideal case of exact isoviscous corner flow. Some features of these directions include that they tend to concentrate near the wedge corner on the upper surface and that they allow material laterally far from the wedge corner to approach the wedge corner. In particular we have evaluated the paths for a 30^o and 60^o slab that ends up on the upper surface 60 km above the downgoing slab. This is the case for the volcanic front to be approximately 120 km above the Benioff zone or 100km above the upper surface of the slab, (i.e. 40 + 60km).

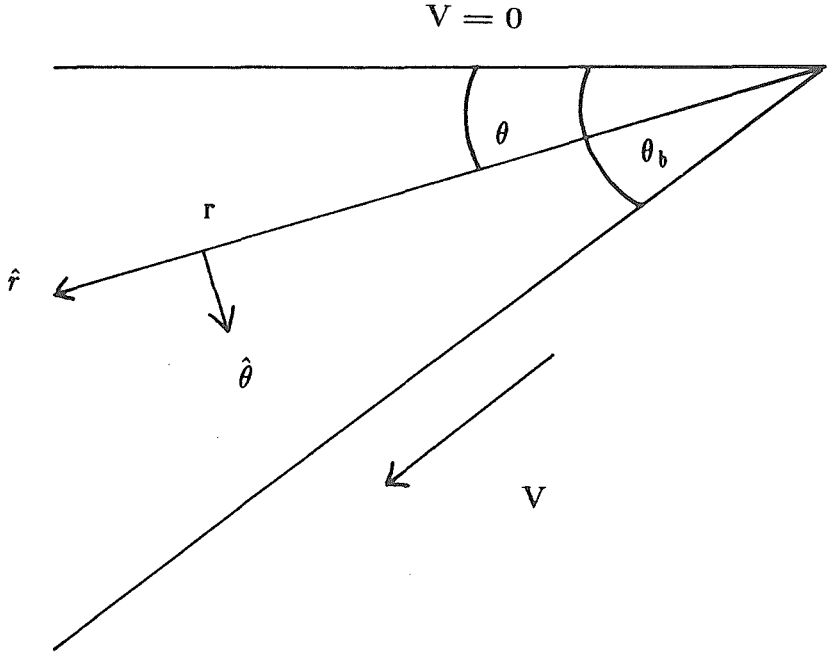


Figure 2.D.1 Figure showing co-ordinate system of Appendix 2.D.

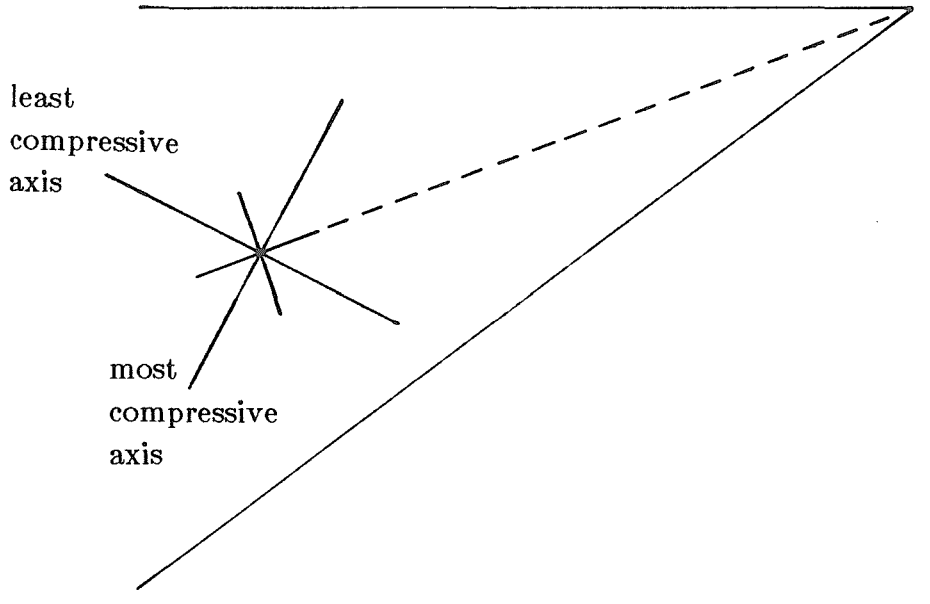


Figure 2.D.2 Figure illustrating relationship of principal stress axes in corner flow regime.

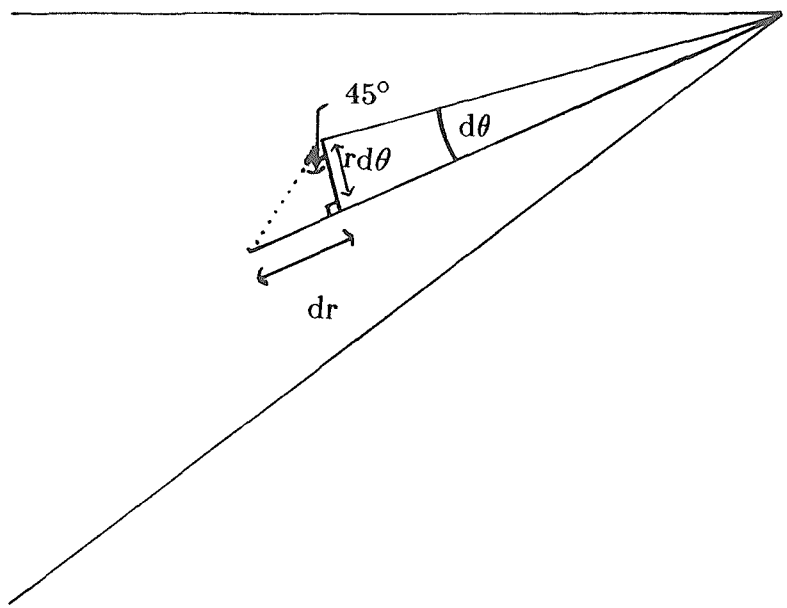


Figure 2.D.3 Figure to illustrate derivation of equation D5.

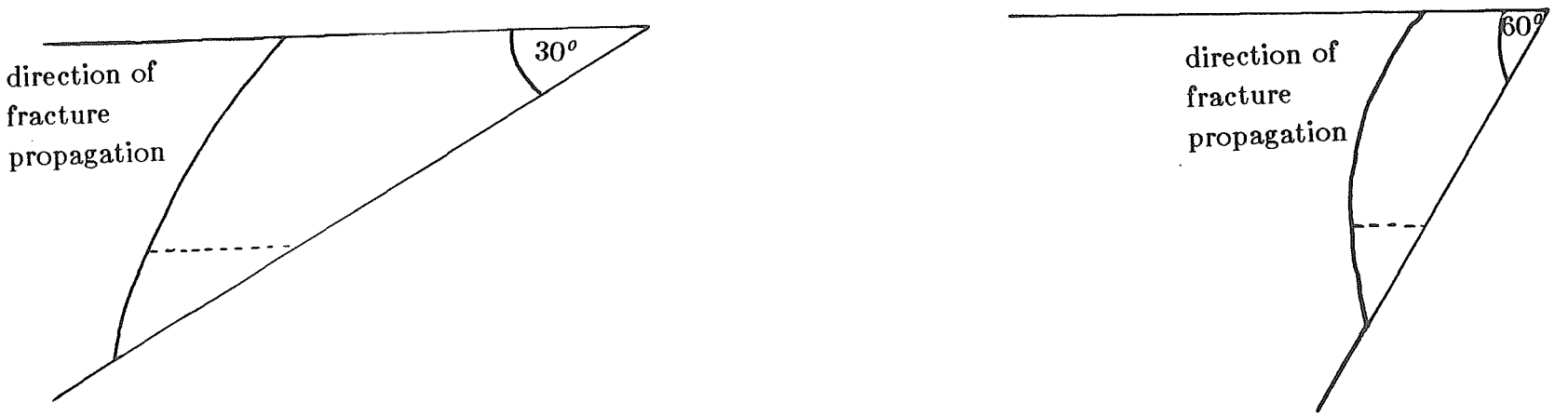


Figure 2.D.4 Figure to illustrate propagation direction of fractures for dip angles of 30° (a) and 60° (b).