UCRL-TR-209573

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Final report for "Development of generalized mapping tools to improve implementation of data driven computer simulations" (LDRD 04-ERD-083)

M.E. Pasyanos, A.L. Ramirez, G.A. Franz

February 10, 2005

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

FY04 LDRD Final Report
Development of generalized mapping tools to improve implementation of data driven computer simulations
LDRD Project Tracking Code: 04-ERD-083

Michael Pasyanos, Principal Investigator

Table of Contents

Table of Contents 3
Section 1: The Proposal 4
- Motivation 4
- Scope 4
- Approach 6
- Expected Results 6
- Personnel 7
- Work Plan 7
- Estimated Budget 8

Section 2: Tool Development 9
- Introduction 9
- Simple data transformations 9
- Advanced data transformations 18
- Modeling of realistic spatial distributions of a physical parameter 19
- Post-processing model transformations 21
- Probability distribution transformations 22
- Implemented transformations 22
- Applications 24

Section 3: Applications 26
- Abstract 26
- Introduction 26
- Methodology 26
- Region 27
- Base sampler 28
- Markov Chain Monte Carlo 30
- Data 34
- Results 36
- Posterior Analysis 41
- Conclusions 44

Section 4: Future Applications 45
Acknowledgements 46
References 46

Section 1: The Proposal

Motivation

Probabilistic inverse techniques, like the Markov Chain Monte Carlo (MCMC) algorithm, have had recent success in combining disparate data types into a consistent model. The Stochastic Engine (SE) initiative was a technique that developed this method and applied it to a number of earth science and national security applications. For instance, while the method was originally developed to solve ground flow problems (Aines et al.), it has also been applied to atmospheric modeling and engineering problems. The investigators of this proposal have applied the SE to regional-scale lithospheric earth models, which have applications to hazard analysis and nuclear explosion monitoring. While this broad applicability is appealing, tailoring the method for each application is inefficient and time-consuming.

Stochastic methods invert data by probabilistically sampling the model space and comparing observations predicted by the proposed model to observed data and preferentially accepting models that produce a good fit, generating a posterior distribution. In other words, the method "inverts" for a model or, more precisely, a distribution of models, by a series of forward calculations. While powerful, the technique is often challenging to implement, as the mapping from model space to data needs to be "customized" for each data type. For example, all proposed models might need to be transformed through sensitivity kernels from 3-D models to 2-D models in one step in order to compute path integrals, and transformed in a completely different manner in the next step.

We seek technical enhancements that widen the applicability of the Stochastic Engine by generalizing some aspects of the method (i.e. model-to-data transformation types, configuration, model representation). Initially, we wish to generalize the transformations that are necessary to match the observations to proposed models. These transformations are sufficiently general not to pertain to any single application. This is a new and innovative approach to the problem, providing a framework to increase the efficiency of its implementation. The overall goal is to reduce response time and make the approach as "plug-and-play" as possible, and will result in the rapid accumulation of new data types for a host of both earth science and non-earth science problems.

Scope

Implementation of the SE to wide-ranging applications is currently hampered by the need to customize the model-to-data mapping required for each data type. In order to make progress on this front, it will be necessary to first characterize the different transformation types, illustrated with specific examples. This will then allow us to generalize the implementation of each transformation into generic mapping classes.

By constructing classes, we will be providing a framework for a large range possible transformations that will be necessary for future data types. We will start by characterizing all possible transformations for three-dimensional (i.e. x, y, z or lat, lon, depth) but time-independent models. We will then consider transformations of time-dependent four-dimensional models (x, y, z, t). While the former is appropriate for earth science problems where the earth is essentially static, the latter is relevant to other applications like ground-water flow or atmospheric problems.

In one application, we have applied the SE probabilistic inverse technique to make high-resolution, regional seismic earth models applicable to problems of non-proliferation. The study area selected was the region of Eastern Asia in the vicinity of the Yellow Sea and Korean Peninsula (YSKP). This region is a high-priority area for nuclear explosion monitoring, where it is important to accurately characterize the seismic velocity structure.

As an example, one data type used in regional structural problems is a receiver function, which measures reverberations directly under a seismic receiver. As such, they are sensitive to the depth structure under a given point. For a given set of points, then, we are essentially collapsing the 3-D model along the z-axis to a series of 1-D models though depth-dependent sensitivity kernels (**Figure 1.1**). This is an extremely common transformation for surface observable data, either by itself or in combination with other transformations.

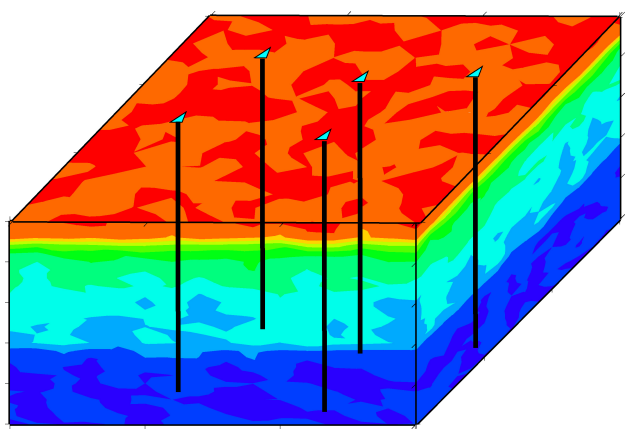

Figure 1.1. Example of one data type which is a series of depth-dependent-only sensitivity kernels.

In some cases, several transformations might be possible for the same data. Seismic waveforms are another data type that can be used to construct and test regional earth models. Depending on the approximation desired, the model transformation can either be 1-D (average structure along source-receiver path), 2-D (slice through the model along the path), or fully 3-D (which might only be a volume along the path).

By classifying the fundamental transformations, and combinations of transformations, new data types can be quickly assimilated into the problem. Furthermore, future applications can make use of these general transformations in order to quickly implement the SE technique. This will have broad applications to a wide variety of investigations.

Approach

Our approach is to answer a set of questions. First, does the approach of generalizing the mapping from model to data enhance the implementation of SE problems? We will answer this question by first performing a census of existing model transformations. These will be taken from several SE applications. By tabulating existing transformations, we will establish classes of these transformations. We will then build out the framework to include all relevant instances of each class.

Once these classes are established we will then ask how this generalization method improves the integration of new data and speed its operation. We will specifically test this for a relevant problem by selecting data sets which utilize the same model transformation. We will then assess the value of this approach for the integration of new data sets.

Finally, what are other possible generalizations (i.e. configuration, model representation) we can make to improve efficiency? While this cannot be fully investigated in the current proposal, this is clearly an opportunity to try and explore other strategies at improving implementation of the SE approach.

Expected Results

From this research, we will develop a framework for quickly and easily incorporating different data types into probabilistic inverse techniques like the SE. We will report on our findings of the generalization of the mappings, focusing on the ability of such an approach to successfully increase the efficiency of implementation. While the framework will be general and hence applicable to multiple products, we will also be providing instances of its application in order to deliver concrete examples of possible benefits. For a relevant earth science problem (a high-resolution regional seismic model of YSKP), we will provide results.

This project is highly relevant to the Energy and Environment (E&E) Directorate's strategic needs, which include data driven simulations of complex systems, particularly those that are relevant to earth science problems. In particular, this would help facilitate interaction with GEON (Keller, 2003), a consortium which provides cyberinfrastructure for the geosciences, and which LLNL is a team member. The E&E Directorate shares GEON's goals to provide infrastructure that manages, preserves, and accesses earth science data and fosters studies that address complex scientific problems. This proposal directly addresses this need.

Depending on the specific application, this research is also relevant to several overall missions of the lab. For example, the nuclear explosion monitoring application supports National Security efforts in Nonproliferation & Arms Control and Technologies for Monitoring. Other earth and atmospheric science applications support the lab's efforts in environmental technologies, such as environmental remediation and atmospheric release.

This research has several potential follow-ons. First, if this technique appears promising, we can prepare a new proposal to further develop this approach and develop other potential generalizations that can improve efficiency and increase response time. This proposal would continue to address data driven simulations of complex systems for a variety of applications.

Another potential follow-on is specific to the regional seismic model application. If we are able to incorporate enough data types to produce a high-resolution model, then we will be assessing the feasibility of using stochastic models to lower the nuclear monitoring threshold. In combination with correlation techniques which match new waveforms against template waveforms, we could significantly improve the monitoring threshold in regions of interest by using model-derived waveforms in lieu of empirical data in aseismic regions. This work would most likely be supported by agencies like the NNSA Office of Defense Nuclear Non-Proliferation.

Personnel

Michael Pasyanos is a seismologist/geophysicist with an expertise in seismic surface waves and tomographic inversion methods. His primary work at LLNL has been in the ground-based nuclear explosion monitoring (GNEM) group where he has established methods to improve the ability to discriminate between earthquakes and nuclear explosions using surface waves and geophysical models.

Abelardo Ramirez is a geophysicist with expertise in electrical and electromagnetic methods, tomographic and stochastic inversion methods. His primary work at LLNL has been in electrical geophysics where he has developed and used methods to map subsurface structures and plumes in support of a variety of LLNL programs.

Gregory "Al" Franz is a computer scientist. His primary work at LLNL is in the Earth Sciences group, within the Energy, Environment, Biology and Institutional Computing Division of the Computing Applications and Research Department (CAR-EEBI) where he has developed software and numerical methods to utilize the Stochastic Engine to propose models of the Earth's structure, based upon seismological observations.

Work Plan

FY04 (six-month project)

Perform a census of model transformations. Classify theoretical model transformation types. Develop a framework for set of transformations. For relevant problem, implement one common transformation on several data sets and compile results. Explore other strategies for improving SE approach.

Estimated Budget

Personnel	FY04 (FTE)
MP (9743)	0.25
AR (9880)	0.25
GF (9823)	0.50
Total wages	1.00*6 months = 0.50 FTE

Wages	$88 K
TID	$2 K

Total	$90 K

Section 2: Tool Development

Introduction

The Stochastic Engine (SE) is a data driven computer simulation tool for predicting the characteristics of complex systems. The SE integrates accurate simulators with the Monte Carlo Markov Chain (MCMC) approach (a stochastic inverse technique) to identify alternative models that are consistent with available data and ranks these alternatives according to their probabilities. Implementation of the SE is currently cumbersome owing to the need to customize the pre-processing and processing steps that are required for a specific application.

This project widens the applicability of the Stochastic Engine by generalizing some aspects of the method (i.e. model-to-data transformation types, configuration, model representation). We have generalized several of the transformations that are necessary to match the observations to proposed models. These transformations are sufficiently general not to pertain to any single application. This approach provides a framework that increases the efficiency of the SE implementation. The overall goal is to reduce response time and make the approach as "plug-and-play" as possible, and will result in the rapid accumulation of new data types for a host of both earth science and non-earth science problems.

When adapting the SE approach to a specific application, there are various pre-processing and processing steps that are typically needed to run a specific problem. Many of these steps are common to a wide variety of specific applications. Here we list and describe several data transformations that are common to a variety of subsurface inverse problems. A subset of these steps have been developed in a generalized form such that they could be used with little or no modifications in a wide variety of specific applications. This work was funded by the LDRD Program (tracking number 04-ERD-083).

Simple data transformations

The MCMC approach aims to search for subsurface models that are most consistent with available data. These models can then be used to calculate other parameters of interest such as seismic velocity. Some of the calculations require that the properties in a 3-D model be somehow "simplified" in order to be consistent with the physics being modeled. **Figures 2.1 – 2.10** shows some of the simple data transformations we have identified and in many cases have coded.

3-D model to 3-D volume: (refer to **Figure 2.1**). This transformation is simply a reduction in volume size with possible re-sampling. This is necessary when the forward problem requires a full 3-D volume, but not necessarily one that covers the complete 3-D model. Instead the codes might only require a single parameter (i.e. Vp only instead of Vp, Vs, density, Qp, Qs, etc.) or a small volume along the path, as would be required for finite difference codes. Examples include finite difference codes like E3D (Larsen and Schultz,

1995) or spectral element finite element codes like SPECFEM3D (Komatitsch and Tromp, 1999; Komatitsch et al., 2002) for waveform analysis. This is also required for finite difference travel time codes, like the algorithm originally developed by Vidale (1988) and further refined by Hole and Zelt (1995), which uses an approximation to the Eikonal equation to compute first arrival travel times through regularly gridded velocity structures.

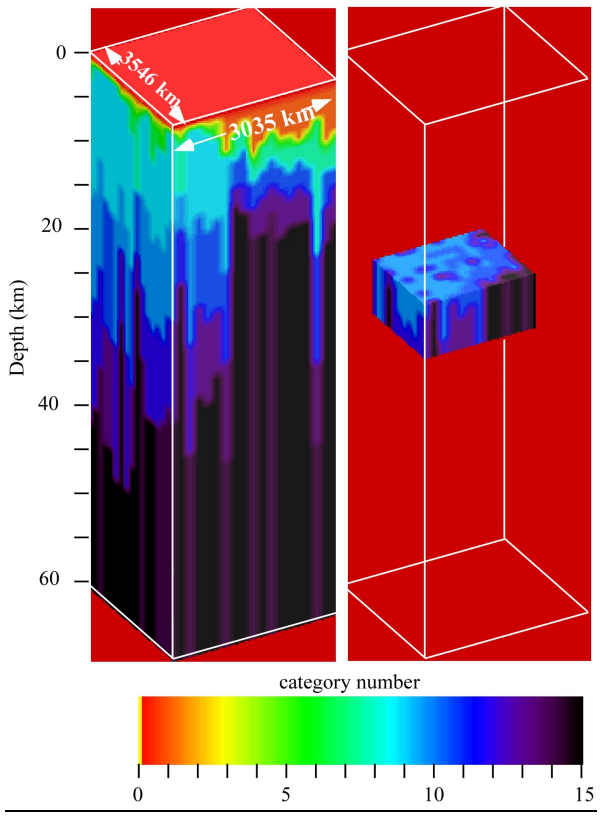

Figure 2.1. 3-D model to 3-D volume

3-D model to 2-D map: (refer to **Figure 2.2**). One such simplification (left side of **Figure 2.2**) involves the integration of properties along the vertical direction using depth-dependent sensitivity kernels as weighting functions; the integrated properties are displayed as a 2-D map. When the sensitivity kernel has the shape of a delta function with the maximum located at depth Z, this transformation returns the horizontal distribution of model values for this depth. The calculation of seismic surface wave group velocity or phase velocity is an application that uses this approach. This is also used for seismic, body wave travel time calculations to reduce a complex model to a series of 2-D surfaces.

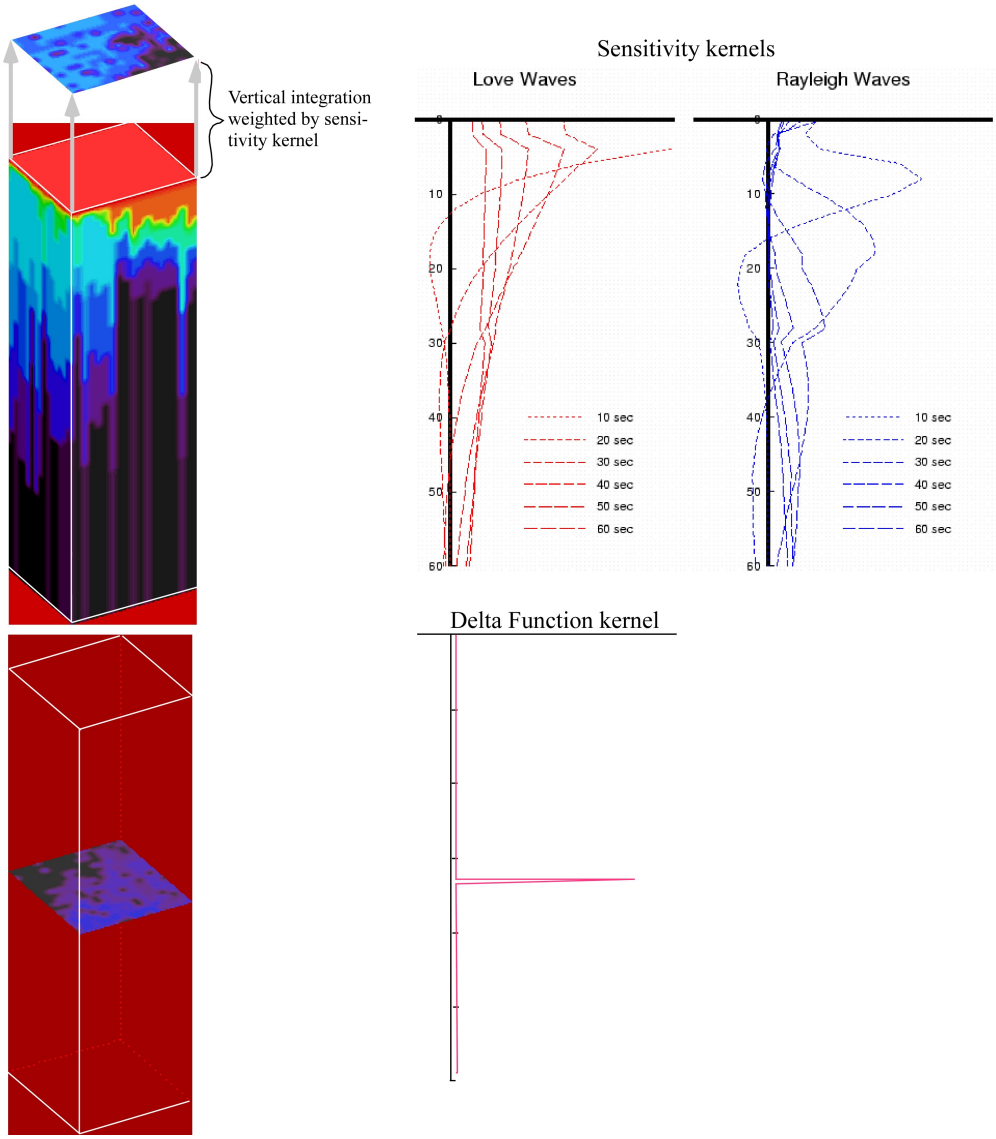

Figure 2.2. 3-D model to 2-D map

3-D model to 2-D cross-section: (refer to **Figure 2.3**). In this case, we need to extract a distribution of values along a 2-D vertical plane (say, from source to receiver). In some cases, the 2-D distribution of values is calculated by averaging the property along the normal to the plane of interest. Seismic waveform synthesis using a 2-D version of the 3-D finite difference programs like E3D (Larsen and Schultz, 1995) is one application that uses this approach, as would be programs that calculate travel times using a Gaussian beam technique like Xgbm (Davis and Henson, 1993).

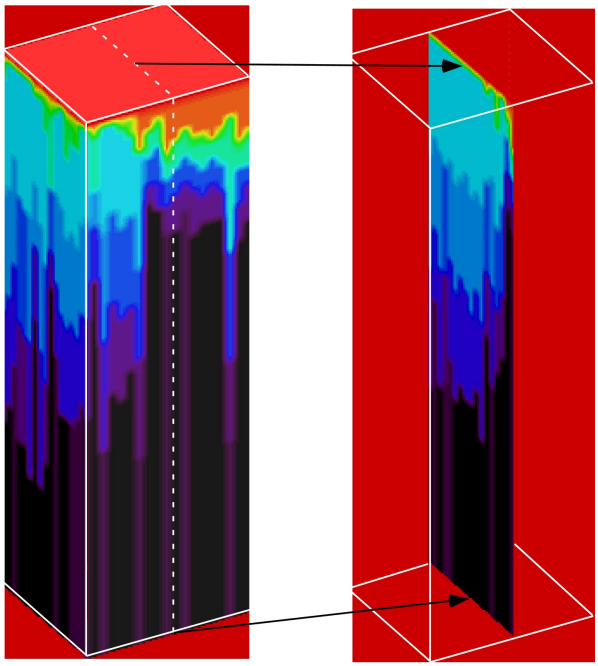

Figure 2.3. 3-D model to 2-D cross-section

3-D model to 1-D profile: (refer to **Figure 2.4**). This transformation consists of extracting a vertical profile of property values from the 3-D model. This approach is used when calculating teleseismic receiver functions, which are the reverberations and conversions of incident P-waves under a seismic station.

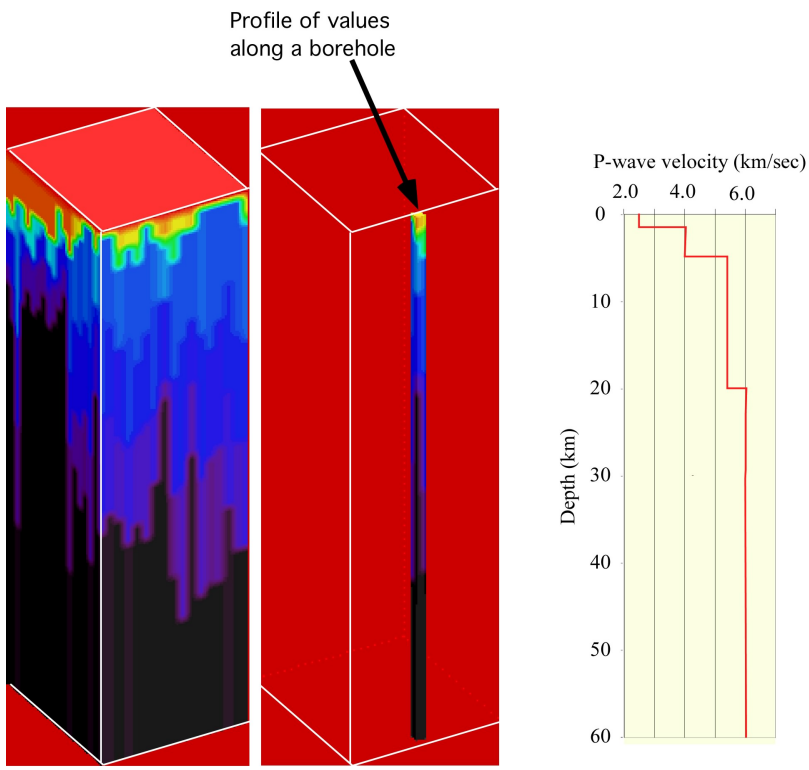

Figure 2.4. 3-D model to 1-D profile

3-D model to surface: (refer to **Figure 2.5**). This data transformation is necessary in order to extract some types of surface from a 3-D model. The surface could either be isocontours like iso-velocities (i.e. surface having 6.0 km/s P-wave velocity) or a surface that might represent physical units such as lithology. Examples of this would be surfaces of basin depth or Moho depth.

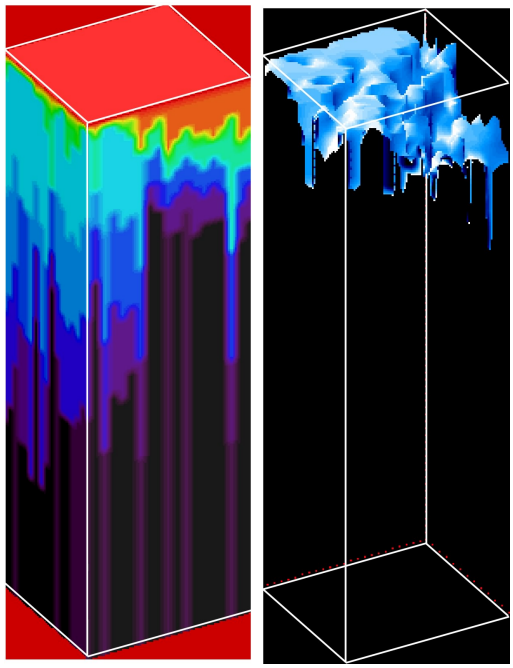

Figure 2.5. 3-D model to surface

2-D cross-section to 1-D profile: (refer to **Figure 2.6**). In this case, we extract a distribution of values along a 2-D vertical plane, and then average the values to produce a "layered-cake" model. Only the values within a given layer are averaged together in order to preserve the geological layering, layer boundaries and discontinuities. This approach is used when synthesizing waveforms using a simplified 1-D model like reflectivity (Kennett, 1985) or a Gaussian beam approach like Xgbm (Davis and Henson, 1993).

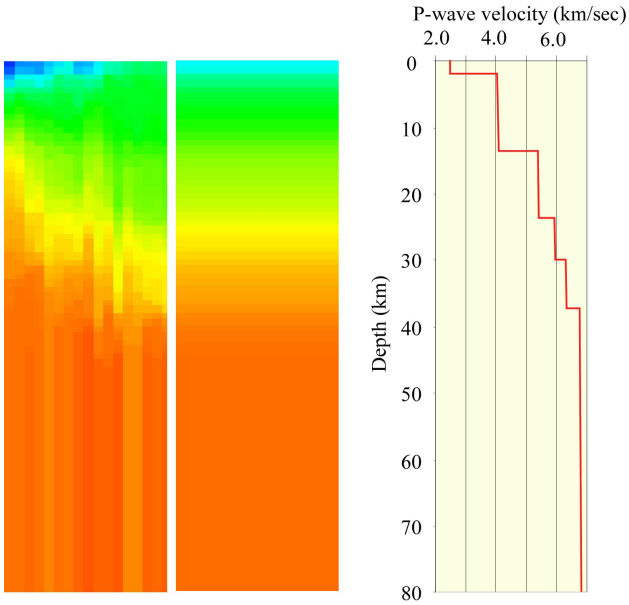

Figure 2.6. 2-D cross-section to 1-D profile

3-D model to voxel: (refer to **Figure 2.7**). This transformation consists of extracting a single voxel value from a 3-D model.

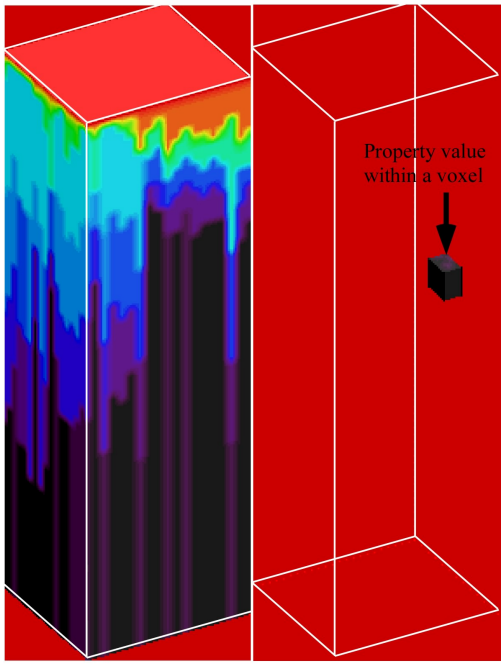

Figure 2.7. 3-D model to voxel

2-D model to pixel: This transformation consists of extracting a single pixel value from a 2-D model.

Integrate properties along a 2-D great circle path: (refer to **Figure 2.8**). This transformation integrates a property (such as slowness) along a great circle path. This transformation is useful when calculating seismic arrival times (which accumulates as travel time slowness), group times (which accumulates as group slowness), or amplitudes (which accumulates as attenuation).

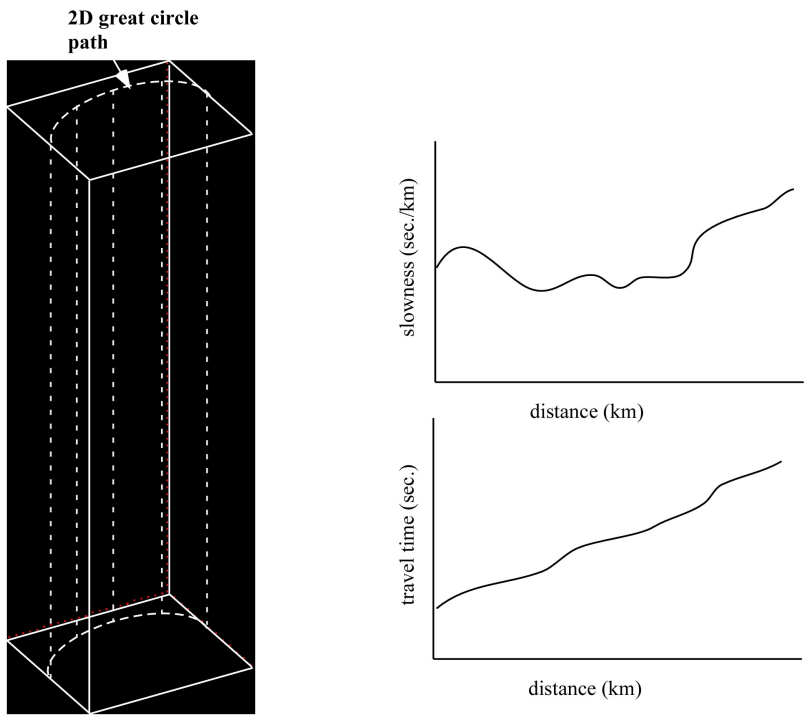

Figure 2.8. Integrate properties along a 2-D great circle path.

Integrate properties along a general 2-D path: (refer to **Figure 2.9**). This transformation is similar to the one above except that the path of interest follows a general route that is specified by the user. This transformation is useful when large contrasts in seismic properties are expected to result in significant ray bending where the ray is not expected to travel along the great circle path. Examples are the same as in the integration along the great circle path.

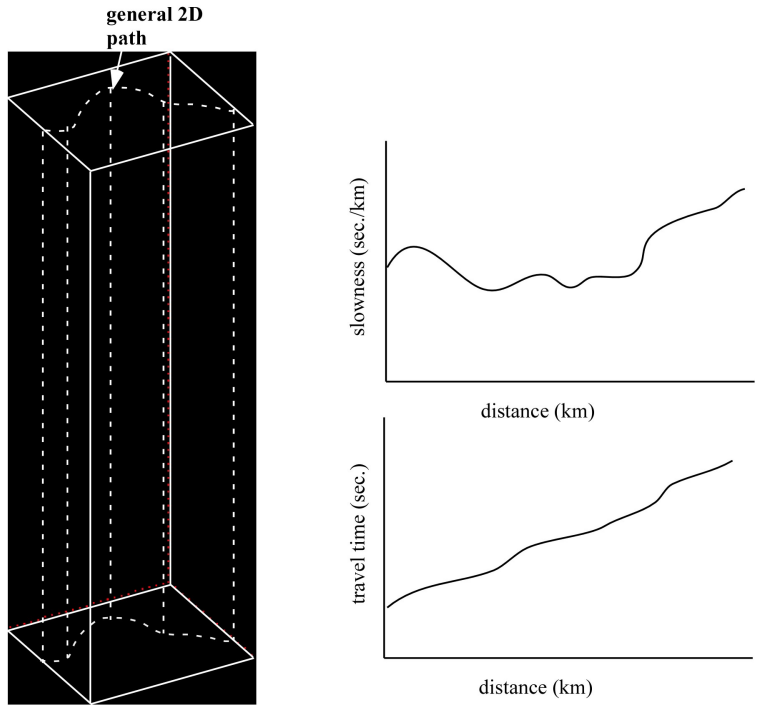

Figure 2.9. Integrate properties along a general 2-D or 3-D path.

Map category to physical data. (refer to **Figure 2.10**). We can use the lithology (the general physical characteristics of a rock) as the subsurface model of interest because it is typically correlated with many parameters of interest such as seismic velocity and density. In these cases, one can choose a categorical simulation approach where each category is associated with a discrete property value. For example, if the set of categories is $\{gravel, sand, clay\}$, then $gravel \rightarrow VP_1, sand \rightarrow VP_2, clay \rightarrow VP_3$, where the VP_i represent distinct p-wave velocity values or some other property. This mapping from category to physical parameter is one of the data transformations we could generalize in order to accelerate the implementation of specific applications of the MCMC approach.

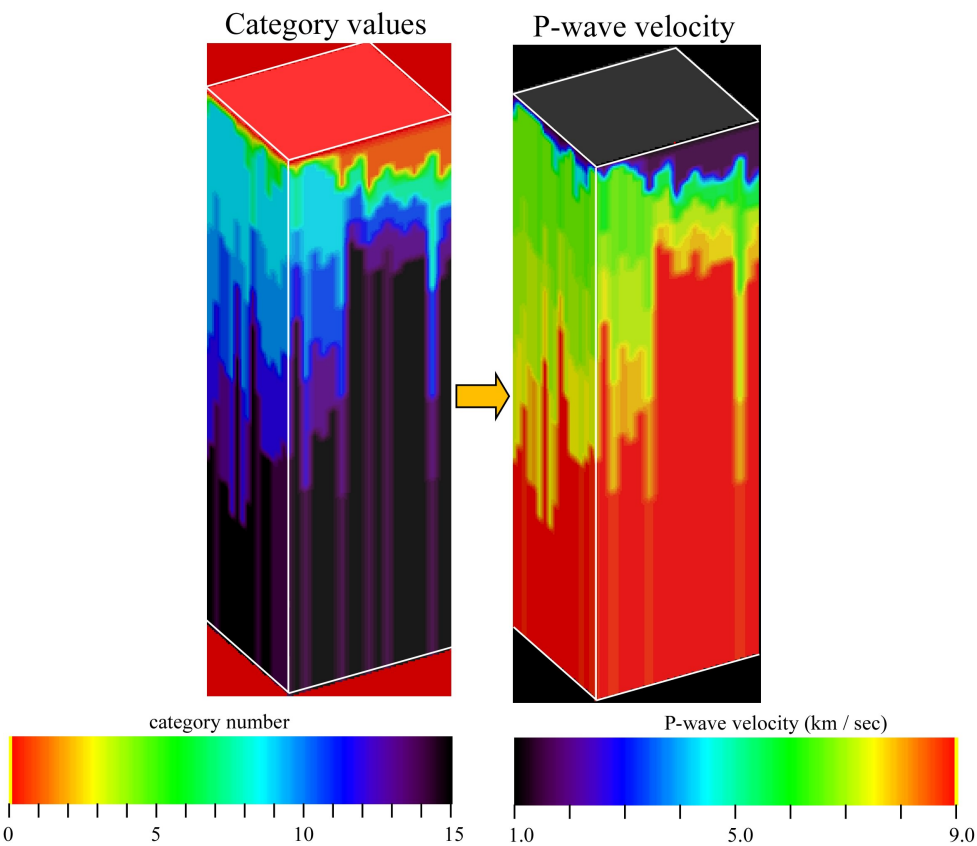

Figure 2.10. Map category to physical data values.

<u>Advanced data transformations</u>

Transform from one physical property to another when multiple data sets are inverted jointly. The MCMC method simultaneously uses many types of data to refine our understanding of complex subsurface systems. Suppose that the system under study consists of a plume of fluid injected into a subsurface layer. The actual connection of a plume model to an observation is made via a forward model: given a possible plume configuration, the forward model predicts the values that would be observed by actual measurement. Then, the predicted data are compared to the measured data, yielding an estimate of the probability that the proposed plume model is in fact the true model. By staging these comparisons in a series, we can identify probable plume models that are consistent with all available data. For example, consider an application where CO_2 is injected into a geologic reservoir and that cross-borehole electrical resistance and borehole tiltmeter surveys have been made to monitor the spatial extent of the CO_2 flood. To solve the forward problem for the electrical survey data, it is necessary to transform a given plume configuration into a 3-D model of electrical resistivities using an appropriate petrophysical model. To solve the forward problem for the tiltmeter data, we need to transform the plume model to a 3-D pressure field using a different petrophysical model. In a staged inversion, these data transformations could be generalized such that a single

software module can be used regardless of the type of data available and the petrophysical model needed.

Pre-processing of data differences. In some applications, the goal of the inversion is to characterize differences that develop due to some forcing condition such as injecting fluid into a layer, or due to different frequencies being used to probe the target. In these situations, the observations of interest are not the measured quantities themselves, but rather, time (or frequency) differences in the data. This type of perturbation analysis belongs to the class of time-dependent, 4-D problems (function of x, y, z, and t) rather than the simpler 3-D, time-independent problems. An example of a 4-D problem would be the use of INSAR (interferometric synthetic aperture radar) data used to monitor small changes in ground surface elevation due to tectonic or man-made events. We could consider the development a general module that any application could use to generate the data differences.

<u>Modeling of realistic spatial distributions of a physical parameter</u>

Natural geologic systems exhibit spatial variability of physical properties such as seismic velocity. This spatial variability is typically not random, i.e., it exhibits trends that are related to the various geologic processes that produced or distributed the geologic materials. When these trends can be quantified, they can be used as a constraint to stabilize the MCMC inversion. For example, in sedimentary basins, layers of gravel, sand, silt and clay tend to be deposited in a particular order, and tend to have particular correlation lengths. In many geologic settings, available data suggests that the values of a given property vary in a gradual or "smooth" way from one location to the next. We can make use of these trends to produce subsurface models that honor these observations. The models produced can also honor data collected along a profile, and spatial correlation data.

Honor vertical profile information: When data collected along a profile are available (e.g., borehole geophysical logs, core sample properties), we can use it to constrain the properties of the subsurface models proposed by the MCMC process. The goal of this approach is to produce models that exhibit the same properties along the line where the profile data was observed. The profile data can be considered as 100% accurate, and thus the proposed model has a probability $p = 1.0$ that it will honor the data exactly. The data can also be considered as somewhat uncertain and the models proposed have a probability $0.0 < p \leq 1.0$ that it will honor the data within some tolerance; in this case p is inversely related to the uncertainty.

Honor surface information (maps): This approach is similar to the one above; in this case, the goal is to produce models that exhibit the properties observed along the plane located at the ground surface. As before, the data can be considered as 100% accurate or have some measure of uncertainty.

Honor smooth trend - spatial correlations in a local neighborhood: (refer to **Figure 2.11**). Many subsurface environments tend to exhibit properties that vary smoothly between a given locality and its adjacent neighbors. When inverting for such models it is desirable to include a constraint that forces the search to only consider smoothly varying models. In these situations, one can introduce a dependency between the local value (e.g., voxel or profile) and that of its neighbors. When the pixels consist of continuous values, the simplest approach is neighborhood averaging. Each pixel is replaced by the average value of the pixels contained in some neighborhood about it. This approach tends to blur the image. An alternative approach is to use a median filter. In this case, a neighborhood around the pixel under consideration is used, but this time the pixel value is replaced by the median pixel value in the neighborhood. This approach has the advantage that the sharpness of image edges is preserved.

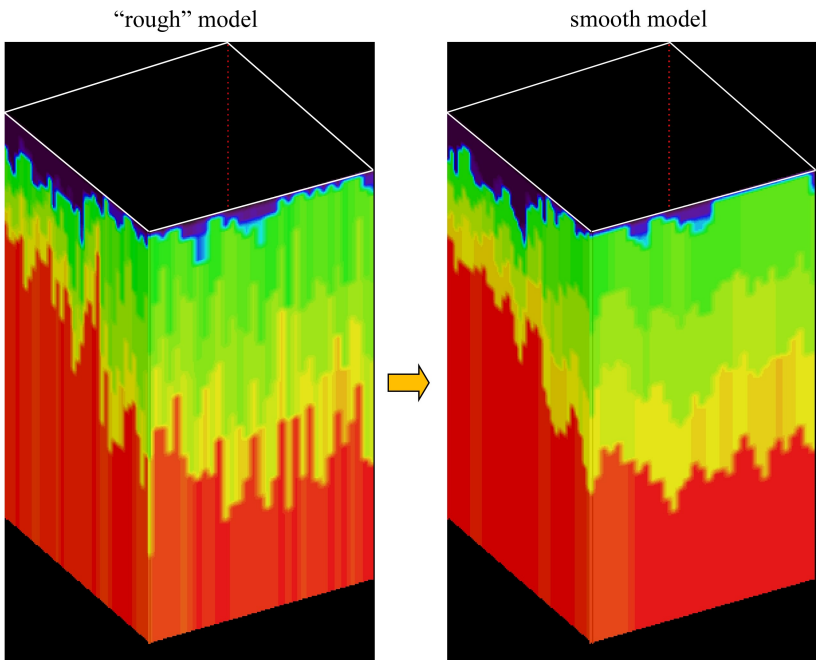

Figure 2.11. Transform a "rough" model to a model whose properties vary smoothly from one location to the next.

In cases where the image values consist of categories (e.g., sand, gravel and clay) we can consider the following approach. If a given voxel consists of sand, there is a probability p such that $0.0 < p \leq 1.0$ that the surrounding voxels also consist of sand. We can impose this type of smoothness constraint by forcing adjacent regions to match identically (or within some tolerance) a given voxel or profile. The values of p can be selected using prior data (data that can be used to compute or to sample the prior probability distribution) and expert judgment. We could develop a generalized module that incorporates one or more of these approaches.

Honor spatial correlation data – Fourier transform approach: This method can also produce realistic models of subsurface heterogeneity using prior knowledge. It is

conceptually similar to the previous approach but is implemented by characterizing correlation length scales in the frequency domain. This method models variability as a stochastic process that can be characterized by its spatial covariance structure (e.g. Hubbard et al., 1999). Spatial covariance is used to determine the degree to which two or more spatial random variables are related (e.g. the thickness of sand and clay layers) The spatial covariance structure is based on data such as core sample observations, outcrop maps and the distances separating the observations. It assumes that the subsurface model is composed of variations characterized by different length scales. The spatial covariance function is expressed as a spectral density function using Fourier transformation.

Honor spatial correlation data and profile information – TSIM: TSIM is a categorical geostatistical approach that can produce realistic models of subsurface heterogeneity using prior knowledge (Carle, 2003). This methodology combines disparate types of observational data such as geophysical borehole logs, geologic insight, and geostatistical trends to produce a consolidated body of knowledge indicating those layer configurations that are most consistent with the available data. It makes use of data such as the number and type of materials (e.g. gravel, sand, clay), the relative volumetric proportions for each of the materials, and horizontal and vertical correlation lengths. It uses the transition probability statistic to analyze spatial variability of geologic materials and to formulate co-kriging equations. This approach makes use of spatial cross correlations (how different geologic materials tend to locate in space relative to each other). It also honors profile information in a way that is more realistic than described previously under "Honor vertical profile information". TSIM assumes that the profile data has a measure of uncertainty; thus, the probability that a given model will honor the profile data is no longer 1.0; instead it is inversely proportional to the data uncertainty.

<u>Post-processing model transformations</u>

As the Markov chain generates samples, it is important to verify that these samples are statistically representative of the posterior distribution. The methodology used to perform this verification is referred to as "convergence analysis". Convergence analysis tools developed during a previous project assumed that the posterior samples consisted of 2-D or 3-D distribution of values within a mesh; all the mesh elements are assumed to be of the same size.

The same assumption is made by the tools previously developed for "posterior analysis"; i.e., the tools used to estimate the properties of the posterior distribution produced by the MCMC approach. These tools summarize the relevant information in the posterior samples so that it can be visualized and understood.

Transform a columnar model to a 3-D mesh model: (refer to **Figure 2.12**). An application of interest produces models where the properties of the domain are defined by a series of layers arranged within columns. Multiple columns specify the distribution of seismic property values in 3-D space. To use existing tools, it is necessary to transform from one

parameter space (column model) to another (3-D mesh model). We are developing the tools necessary to perform this transformation.

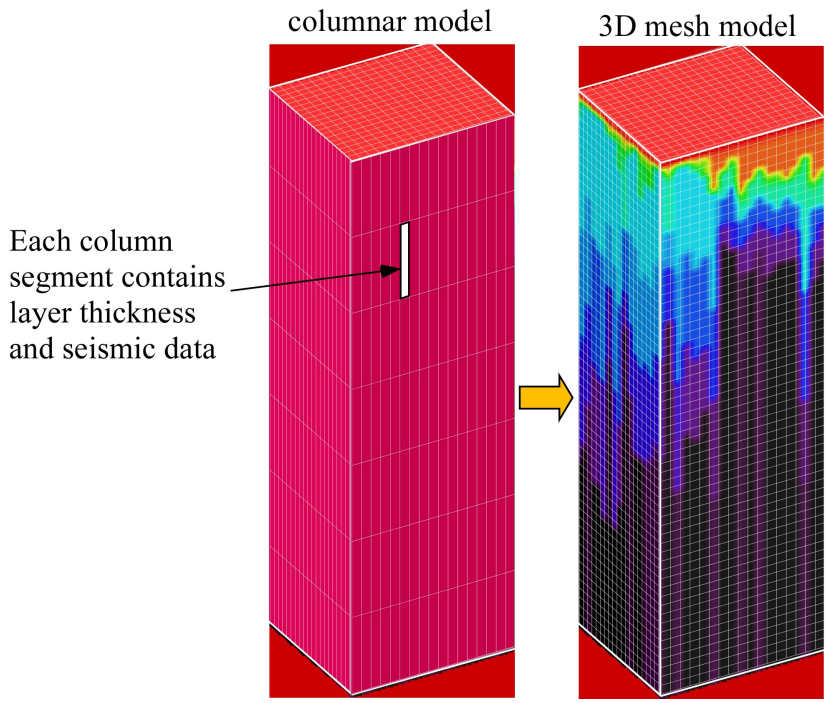

Figure 2.12. Transform a columnar model to a 3D mesh model.

Probability distribution transformations

There are several probability distribution used when generating various types of random variables used by the MCMC process. In general, these require transforming from one type of distribution to another such as transforming a uniform distribution to a normal (or vice versa), or a general distribution to a standard distribution (or vice versa). We have used the following transformations for our work:

a) *Transform a general normal distribution to a standard normal distribution.*
b) *Transform a standard normal distribution to a standard uniform distribution.*
c) *Transform a standard uniform distribution to a standard normal distribution.*
d) *Transform a standard normal distribution to a general normal distribution.*

Implemented transformations

Several codes have been constructed to implement some of the transformations listed above. Most of these programs have been coded in the popular Python programming language. Python is an object-oriented language used for both stand-alone programs and scripting applications (Lutz and Ascher, 1999). The few that are written in "C" are identified by a "*".

Transformation:	**Module name:**	**Class/method:**
3-D model to 3-D		

volume		
3-D model to 2-D map	starting_model_for_swaves.py forward_body_script.py	Surface_Waves Project
3-D model to 2-D cross-section	layers08.py	Layer_Plot.ConvertLayersToXYZ
3-D model to 1-D profile	models.py	Model.get_vertical_profile
3-D model to surface	forward_body_script.py	ForwardBody.Project
2-D cross-section to 1-D profile		
3-D model to voxel		
2-D grid to pixel	starting_model_for_swaves.py	Starting_Model.Surface_Waves
Integrate properties along a 2-D great circle path	BodyForward.c SurfForward.c	BodyForward SurfForward
Integrate properties along a general 2-D path		
Integrate properties along a general 3-D path	forward_body_script.py	Forward_Rcvr.Conversion
Map category to physical data		
Transform from one physical property to another when multiple data sets are inverted jointly		
Pre-processing of data differences		
Honor vertical profile information	forward_rcvr_script.py	Forward_Rcvr.Conversion
Honor surface information (maps)		
Honor smooth trend - spatial correlations in a local neighborhood	models.py	Model.NeighborhoodWeightedAverageFilter
Honor spatial correlation data – Fourier transform approach		
Honor spatial correlation data and profile information – TSIM		
Transform a columnar model to a 3-D mesh model	super_script_meshmodel_v2_file_list.py	chng_col2grid
Transform general normal distribution to a standard normal distribution	Sampler.c*	RDist()
Transform standard normal distribution to a standard uniform distribution	Sampler.c*	P()
Transform standard uniform distribution to a	Sampler.c*	UniformToNormal()

standard normal distribution		
Transform a standard normal distribution to a general normal distribution	Sampler.c*	RDist()

Table 1 lists the transformations that have been implemented as part of this project.

Applications

Here we present how we are currently employing some of these transformations in an application to construct regional models of the crust and upper mantle for the Yellow Sea – Korean Peninsula (YSKP) region, using three different data types (receiver functions, surface wave dispersion, body wave travel times). A detailed description of this work has been submitted for publication in Earth and Planetary Science Letters: "Reconciling data using Markov Chain Monte Carlo: An application to the Yellow Sea - Korean Peninsula region" (Pasyanos et al., 2004). Below we list the various steps in the processing sequence and the transformations used for the YSKP work.

Base Sampler
    Honor smooth trend

Stage 1 (Receiver functions)
    3-D model to 1-D profile

Stage 2 (Surface wave dispersion)
    3-D model to 2-D map
    Integrate along great circle path

Stage 3 (Body wave travel times)
    3-D model to 2-D map
    3-D model to surface
    Integrate along great circle path

Convergence analysis
    Transform a columnar model to a 3-D mesh model

Clustering analysis
    Transform a columnar model to a 3-D mesh model

While we have primarily focused our efforts on the YSKP application, we can now easily apply the transformations developed to other geophysical problems. We now present an example of the procedure for the construction of a seismic model of the Las Vegas Valley and southern Nevada using data sets available for this region. This model is of interest for predicting ground motions for the Test Site Readiness Program. The order of the stages listed is arbitrary.

Base Sampler
    Honor smooth trend

Stage 1 (Geotechnical shear wave profiles)
    3-D model to 1-D profile

Stage 2 (Seismic reflection profiles)
    3-D model to 2-D map
    3-D model to surface
    Integrate along great circle path

Stage 3 (Receiver functions)
    3-D model to 1-D profile

Stage 4 (Surface wave dispersion)
    3-D model to 2-D map
    Integrate along great circle path

Stage 5 (Body wave travel times)
    3-D model to 2-D map
    3-D model to surface
    Integrate along great circle path

Stage 6 (Gravity)
    3-D model to 2-D map

Convergence analysis
    Transform a columnar model to a 3-D mesh model

Clustering analysis
    Transform a columnar model to a 3-D mesh model

Section 3: Applications

Abstract

In an effort to build seismic models that are most consistent with multiple seismic data sets, we have applied a new method known as the Stochastic Engine (SE). The SE uses a Markov Chain Monte Carlo (MCMC) technique to sample models from a prior distribution and test them against multiple data types to generate a posterior distribution. While computationally expensive, this approach has several advantages over a single deterministic model: the use of prior model information; the reconciliation of different data types that constrain the model; the estimation of uncertainties of model parameters, properly migrating data uncertainties into model uncertainties; allowing non-Gaussian or multi-modal distributions; and the estimation of uncertainties on predicted observable signals. In this paper, we use this method to determine the crust and upper mantle structure of the Yellow Sea and Korean Peninsula (YSKP) region. We outline the technique and how it has been implemented, observe the behavior of the inversion, and demonstrate with examples the many advantages of this approach.

Introduction

A classic dilemma in earth sciences is the inability to reliably fit several data types simultaneously. Attempts to fit one type of data produce results that are often contrary to results obtained using other types of data. Trying to fit both data sets simultaneously are often difficult or impossible. Here we present the results of a stochastic method that should be applicable in these and other scenarios. This method, which the developers have referred to as the Stochastic Engine (SE), employs Bayesian inference though the use of a Markov Chain Monte Carlo (MCMC) algorithm. The algorithm is used to sample models from a prior distribution, test it against one or more data types, and generate a posterior distribution most consistent with the data and priors. The posterior has a distribution that is consistent with the uncertainties of the data. The technique has been applied to many applications ranging from statistical physics to the spread of infectious diseases. In the earth sciences, this method has recently been successfully applied to electrical resistivity changes (Ramirez et al., 2004). We have applied the same technique to seismic data to determine crust and upper mantle structure on a regional scale.

Methodology

The method draws samples from the prior distribution, evaluates their acceptability, and then forms sample averages to estimate the properties of the system under evaluation. MCMC draws these samples by running a cleverly constructed Markov chain for a long time (Gilks et al., 1996). A Markov chain is a sequence of random values whose probabilities at a time interval depends upon the value of the number at a previous time. A good example would be a random walk. The Markov chain that we construct here

follows a set of rules that preferentially moves to more likely states in the model space, but sometimes transitions to less likely states. The result is a chain that can sample the complete model space, efficiently inspect high-likelihood regions, but that is unlikely to get trapped in local extrema. MCMC methods invert data by probabilistically sampling the model space with the Markov chain and comparing observations predicted by the proposed model to observed data and preferentially accepting models that produce a good fit, generating a posterior distribution.

Both Bayes theorem and sampling methods like Monte Carlo have been extensively employed in the earth sciences. Tarantola (1987) employed a Bayesian framework for tomography, which incorporates a prior background model into the tomographic inversion. More recently, Taylor et al. (2003) have applied this method to Lg attenuation tomography in eastern Asia. While these types of methods succeed at including prior information, they assume the model follows Gaussian statistics. Furthermore, these techniques seek to find a single solution to the problem where the uncertainties are also normally distributed. In our formulation of the problem, no assumptions have been made about the prior distribution of models and we find, in fact, that in many instances Gaussian distributions are not applicable. Instead of a single model, what we seek are the distributions of models (which may or may not be Gaussian) that are most consistent with both our prior information and our observations. This approach treats the non-uniqueness of the inversion directly by allowing for alternative models that are consistent with the data and ranking them according to their consistency.

MCMC originated in statistical physics and more recently has starting finding applications in many different fields. Application in the earth sciences has been more limited. In at least one recent application, Shapiro and Ritzwoller (2002) employed MCMC, along with a linearized inversion and simulated annealing, to invert for shear velocity structure using surface waves. The methodology proposed here differs from previous earth science applications by using multiple stages, each having different data sets. There are several advantages to this approach. By ordering our stages properly, we can quickly reject models that cannot fit the observations that are easiest to calculate. The data sets that are more computationally intensive to forward model are relegated to later stages, increasing the efficiency of the problem. Again, what is also different here is that we do not seek a single "best" model or to simplify a distribution of models into a model with uncertainties. The posterior distribution of models is the answer to the problem that we seek.

Region

We have selected as our study area the region of Eastern Asia in the vicinity of the Yellow Sea and Korean Peninsula (YSKP). A map of the region is shown in **Figure 3.1**. While the circum-Pacific subduction zone is very seismically active, the YSKP area has a very low seismicity rate, with some areas of the region being essentially aseismic. Furthermore, while some countries in the region like Japan are well-instrumented with seismic sensors, there are areas in which we have limited access to seismic data. While

we have selected this region for its interest for explosion monitoring, this area is also tectonically interesting. Whereas cratons usually have fast velocity lithospheric roots that extend deep into the asthenosphere and act as keels, the Sino-Korean Paraplatform (the craton on which the Korean Peninsula resides) has been affected by the subduction of the Pacific and Philippine Sea Plates and has an unusually slow lithospheric upper mantle. We wish to examine the unique nature of the cratonic root under this region.

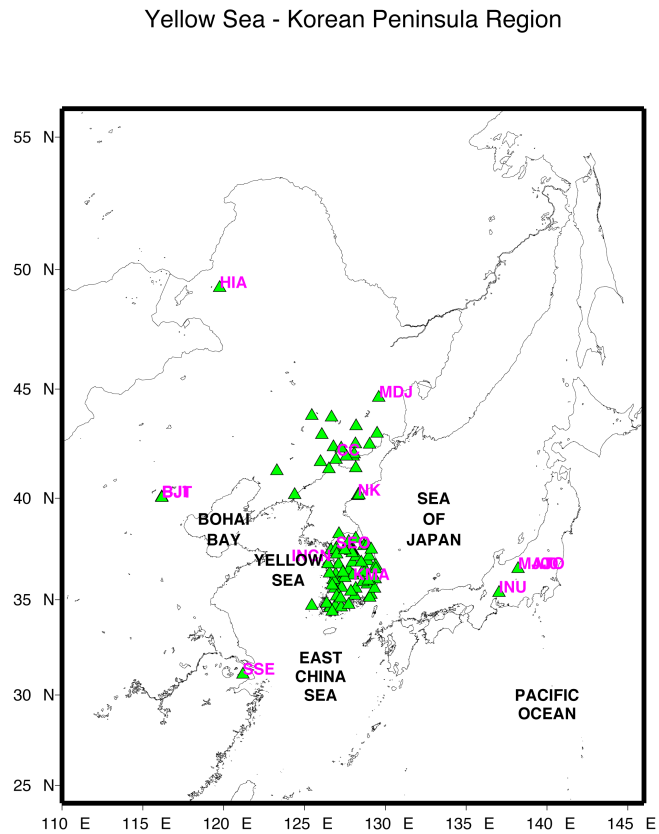

Figure 3.1. Tectonic map of Yellow Sea – Korean Peninsula region. The map also shows stations in the YSKP region.

<u>Base sampler</u>

The first step in implementing the SE is to develop a base sampler which samples the prior distribution. Before even doing this, however, the model must be parameterized. This can be done in any number of ways, but is somewhat critical to the success of the approach. For our problem, for example, we could set up a large regular grid in longitude (x), latitude (y), and depth (z), where each node or cell is specified by seismic parameters (i.e. Vp, Vs, density). A disadvantage of this approach, however, is that we would expect to need finer resolution in the shallower sedimentary layers than a layer buried deep in the upper mantle. While this could be rectified by a variable grid, there would likely be a

need to impose smoothness between adjacent nodes in order to make the problem tractable.

We have instead decided to take a different approach. While the model is specified by a regular grid in latitude and longitude, vertical parameterization is instead accomplished through a series of layers representative of geologic structure (i.e. water, upper sediments, lower sediments, upper crust, middle crust, lower crust, upper mantle) which are not only specified by their seismic parameters, but also by their thickness. This had many advantages including reducing the size of the problem, imposing realistic constraints, and making the resulting models more interpretable. Laterally, we have selected a 2 degree by 2 degree parameterization between 23º and 57º latitude and 109º and 147º longitude, for a total of 323 (17 x 19) cells. Combined with 6 variable layers (all but water) and 4 parameters for each layer (thickness, Vp, Vs, density), this results in a total of 7752 free parameters (323 x 6 x 4). While we would eventually like to go to higher resolution (i.e. 1º x 1º) and eventually include other parameters (i.e. Qp and Qs), this parameterization allowed us to test the methodology and run in a reasonable time period.

The starting model selected for the region was CRUST 2.0 (Bassin et al., 2000) which, conveniently enough, is parameterized along the same lines. In addition to specifying actual values (Vp, Vs, density, thickness) at each point, we also needed to provide the base sampler with specifications on the variations on these parameters. This is accomplished by providing probability distributions on each value along with minimum and maximum values. This was accomplished by looking at the natural variations in the CRUST 2.0 model for this region. For example, for the upper crust layer, the variation of P-wave velocities in the layer is 0.5 km/s with a minimum of 5.0 km/s and a maximum of 6.2 km/s. For S-wave velocities, we find a variation of 0.4 km/s with extrema of 2.5 km/s and 3.6 km/s, densities have values of 0.7 g/cm^3, 2.6 g/cm^3, and 2.8 g/cm^3, and thicknesses values of 4 km, with 1.7 km and 16 km the minimum and maximum. In the water layer, we do not allow the seismic parameters (P-velocity 1.50 km/s, S-velocity 0.00 km/s, density 1.02 g/cm3) or the thickness (determined from ETOPO5, NGDC, 1998) to vary.

From the starting model and these constraints, the base sampler can initially select models consistent with the a priori distribution. The seismic parameters are all selected independently which produces variations in Vp:Vs and Vp:density ratios. The ratios are checked, however, and unrealistic values of these ratios for the layer type are discarded.

Finally, in order to reduce dependence on the starting model and introduce some randomness to initial models for each chain, we have randomly selected 10% of the columns and swapped them with other columns. This randomization is done for each of the two chains in our search. Separate starting points allow us to move within model space along different routes. Compared to standard Monte Carlo sampling, this is often a quicker way of sampling the model space, particularly in complex problems. It is generally useful to choose starting points that are widely dispersed in the target distribution, both to ensure coverage of the target distribution, as well as to monitor

convergence. Prior to convergence, the run has not had enough time to properly sample the model space. The initial period, referred to as the "burn-in", is still unduly influenced by the starting model. When convergence is finally reached, each Markov chain will have had a sufficient number of steps to independently sample the model space. After convergence, the chains should have similar posterior distributions, which should be stationary over enough iterations, referred to as "long run" stationary. For example, the results from the next 1000 iterations should not be markedly different than the previous 1000 iterations (although the previous 10 and next 10 iterations might be). Convergence can be monitored by detecting when the Markov chains have forgotten their starting points and by comparing sequences drawn from different starting points and checking that they are indistinguishable (Gelman, 1996).

Markov Chain Monte Carlo

The Markov Chain Monte Carlo method provides a flexible framework that can be adapted to perform a variety of analyses and inference tasks. It uses a Markov chain state/transition structure to control the sampling process. MCMC techniques enable simulation from a targeted distribution by embedding it as a stationary distribution of an ergodic Markov chain and simulating the chain until it approaches equilibrium. For Bayesian analysis, we are able to adapt the approach to simulate and estimate posterior distributions that embody our available prior information and newly acquired observational data. MCMC algorithms can assume a variety of forms with the most useful to us being the Metropolis framework.

The SE approach is a derivative of the Metropolis algorithm (Metropolis et. al., 1953) as described by Mosegaard and Tarantola (1995). This particular MCMC algorithm has demonstrated significant potential in solving inverse problems involving complex physical systems and supports several key enhancements necessary to mitigate the combinatorial demands underlying the MCMC methodology. Within our stochastic framework, the solution to an inverse problem is an estimate of the posterior probability distribution defined on the corresponding space S of possible solutions (otherwise referred to as states). For any potential solution $s_0 \in S$, the SE will provide an estimate of the probability and confidence that state s_0 is indeed the true solution to the given system. This allows future analysis to focus upon the most likely explanations of system behavior – thereby improving both the efficiency and effectiveness of follow-on efforts. Moreover, results generated from the SE (i.e., the estimated posterior distribution, predictions, hypothesis testing, etc.) may be incrementally updated as more data becomes available.

The inverse problem under consideration may be described as follows. Let D and M denote the data space and model space respectively, and suppose that there exists a mapping G such that:

$$\underline{d} = G(\underline{m}) \qquad (1)$$

where $\underline{m} \in M$ is a parameter vector describing the state of the system of interest and $\underline{d} \in D$ is a vector of measurements taken on that system. The inverse problem occurs when a vector of data values is observed, say $\underline{d}_0$, and we want to determine the value of the parameter vector $\underline{m}_0$ that gave rise to $\underline{d}_0$. In geophysical applications, this problem is substantially under-constrained and ill-posed. In such cases, the search for a deterministic solution for $\underline{d}_0$ that is unique and possesses a high degree of confidence is virtually impossible, and hence a probabilistic solution is likely to be superior to any classical deterministic optimization approach.

The stochastic approach employs a variation of the Metropolis algorithm to generate a sequence of samples from M at rates proportional to their posterior probabilities. These samples enable the estimation of the posterior distribution $F_{M|D}$ using the sample frequencies corresponding to the sampled models $\underline{m} \in M$. By design, the models generated most frequently are most consistent with both our prior information on M and the observations being processed. Assuming mild regularity conditions, the estimated posterior distribution $\hat{F}_{M|D}$ converges strongly to the true underlying distribution $F_{M|D}$. This estimate is the basis for all subsequent analysis and inference including estimation, prediction, confidence assessment and risk/reward trade-off analysis.

Given that the information used to drive the simulation is taken from two distinct sources (prior knowledge and observational data) the sampling process can be viewed as consisting of two separate components. The first component generates samples according to an identified prior distribution $\mathsf{r}(\underline{m})$ defined over the model space M. These samples are called proposal states and constitute possible solutions to the inverse problem, (refer to the left hand side of **Figure 3.2**, base representation box). Specifically, this sampling process is manifested as a Markov chain **Q**, with one-step transition probabilities between states $\underline{m} \in M$ designed to produce a long-run stationary distribution equal to the prior $\mathsf{r}(\underline{m})$. In statistical terms, **Q** samples $\mathsf{r}(\underline{m})$.

The second component takes the form of a decision process that either accepts or rejects the proposal sample generated from the a priori Markov chain (labeled "stage 1" in **Figure 3.2**). For each visited state, forward simulators are used to predict values of measurable parameters such as travel times. These predictions are then compared to corresponding measurements to determine the likelihood $L(\underline{m})$ that the given state $\underline{m} \in M$ produced the observed data. An accept/reject decision based upon this likelihood is used to modify the prior sampling process. The result is a new Markov chain **R**, which samples the posterior distribution $\mathrm{P}(\underline{m})$.

Formally, Bayes rule relates the prior and posterior distributions as follows:

$$\mathrm{P}(\underline{m}) = \mathsf{r}(\underline{m})L(\underline{m}) \qquad (2)$$

The likelihood $L(m)$ is a measure of the degree of fit between the data predicted assuming the model *m* and the observed data. For this study, we assumed a likelihood function of the form:

$$L(\underline{m}) = k\,\mathbf{exp}\left[-\sum_i \left|\left(d_{pred,i} - d_{0,i}\right)^n / \mathsf{s}_i^n\right|\right] \qquad (3)$$

where $d_{pred,i}$ is the predicted data for a given model *m*, $d_{0,i}$ is the vector containing the observed measurements, s_i is the estimated data uncertainty, *k* is a normalization constant, and $n \geq 1$. For the results described below, we assumed that $n = 2$. Note that Eq. 3 assumes that the estimated data errors are uncorrelated.

The decision to accept or reject a proposed state is made on the basis of likelihood comparisons as indicated by the decision box of stage 1, located on the left hand side of **Figure 3.2**. Suppose that the current state of the Markov chain is $\underline{m}_i$ and that a randomized rule based upon the one-step transition probabilities propose a move to state $\underline{m}_j$. If these transitions were always accepted, then the simulation would be sampling from the prior distribution. Instead, suppose that the proposal transition is only accepted according to the following rules:

1) For both the current and proposal states $\underline{m}_i$ and $\underline{m}_j$, compute the respective likelihoods $L(\underline{m}_i)$ and $L(\underline{m}_j)$ that these models produced the observed data.

2) If $L(\underline{m}_j) \geq L(\underline{m}_i)$, then accept the proposed transition with probability 1 and move to state $\underline{m}_j$. (Note: The algorithm always accepts the transition when the new state provides a better explanation of the data than the current state.)

3) If $L(\underline{m}_j) < L(\underline{m}_i)$, then use a randomized decision rule and accept the proposed transition with probability $L(\underline{m}_j)/L(\underline{m}_i) < 1$ and move to state $\underline{m}_j$. Otherwise, transition back to state $\underline{m}_i$. (Note: By allowing the random walk to transition to a less likely state, the process can move out of a local extrema).

The samples generated through this three-step process will have a limiting distribution that is proportional to the desired posterior distribution $F_{M|D}$. That is, the search tends to hover in regions of space M containing states with greater prior propensities and higher likelihoods.

Since we have observational data of differing types (e.g., surface wave dispersion and body wave travel times), our approach takes advantage of the "cascade" rule (Mosegaard and Tarantola, 1995). In these cases, the errors in prediction are often independent and hence the total likelihood expression factors into distinct terms – one for each data type. For the above example, the total likelihood can be expressed as follows:

$$L_{total}(\underline{m}) = L_{surf}(\underline{m}) * L_{body}(\underline{m}) \qquad (4)$$

This probabilistic structure can be leveraged to streamline the transition process employed by the Stochastic Engine. Mosegaard and Tarantola (1995) indicate that performing a single Metropolis transition step (step 3 above) that uses the entire likelihood expression $L_{total}(\underline{m})$ is equivalent to performing a sequence of Metropolis steps – one for each term in the above expression. Once a model is drawn from the prior distribution, the forward model is solved initially for the first data type alone (step 1 as illustrated by stage 1 in **Figure 3.2**). At this juncture, the proposed model may be rejected or accepted (steps 2 and 3 above). If the decision is to reject the proposal, then the forward models in stage 2 (**Figure 3.2**) are not executed. The Markov chain returns to the last accepted state, a new state is proposed and the decision process begins anew with the first data type. If the decision at this stage is to accept the proposal, the next data set is considered in stage 2, its corresponding forward model is run and a decision to accept or reject is made based upon its likelihood. This continues through all of the different types of data until the proposal either is accepted at all stages or is rejected at one stage and starts over at the beginning of the sequence. Note that proposals accepted by both stages are ones that are most likely to be consistent with the prior and both sets of observational data. This approach can easily be extended beyond two stages to incorporate additional types of data.

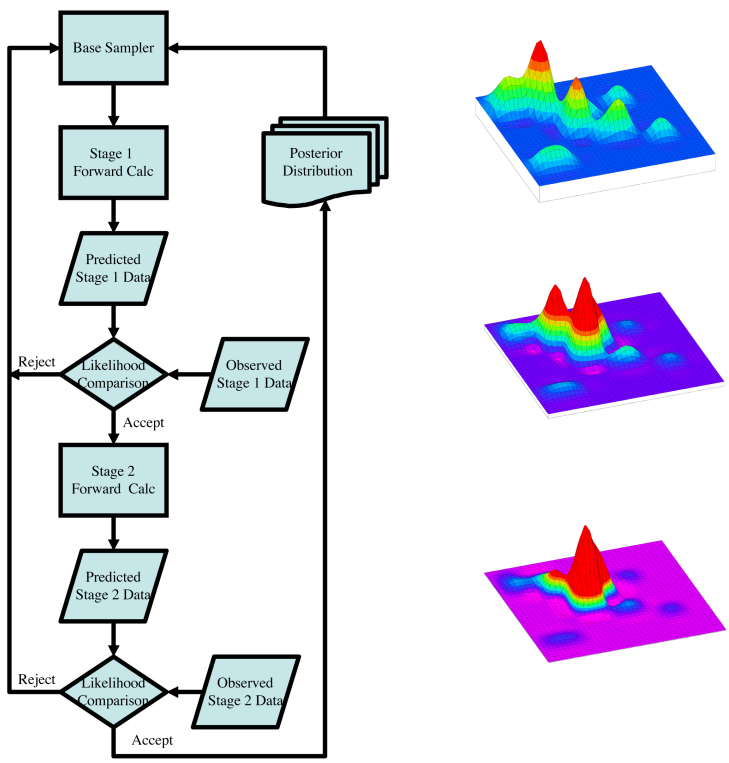

Figure 3.2. Flowchart of the stochastic inversion algorithm, along with a schematic of the likelihood function distribution of the model space at three different stages: prior

distribution, posterior distribution after data from a single stage, posterior distribution after data from two stages.

Data

We have used seismic data from a number of sources in the region. Stations in the Global Seismic Network (GSN), both IRIS/USGS and IRIS/IDA constituted the bulk of data for the region. Data from the China Digital Seismic Network (CDSN) provided coverage of eastern China. Seismic data from several stations were obtained from the International Monitoring System (IMS). Where possible, we have used data from regional networks, like those of the Korean Meteorological Administration (KMA) and Korean Institute of Geoscience and Mineral Resources (KIGAM) in South Korea. We also made use of two PASSCAL deployments in northeast China and in North Korea. Taken together, these stations provide good coverage of our study area (**Figure 3.1**).

The two observational data types that we currently use to construct the model are surface wave group velocities and body wave travel times. Surface wave dispersion measurements have been made for thousands of paths across the region using a multiple filter analysis technique and, as shown in **Figure 3.3**, provide excellent coverage (Pasyanos et al., 2003). The data is assembled by taking measurements at 5 second intervals between 10 and 100 seconds. The short period data is sensitive to the shallowest earth structure (i.e. sediments), intermediate period data to deeper crustal structure (i.e. average crustal velocity, crustal thickness), and long period data more exclusively to upper mantle structure. Combined they do an excellent job of predicting average S-wave velocity structure. This surface wave data, however, has the downside of large tradeoffs between velocities and the depth of discontinuities, allowing multiple models consistent with the same surface wave data. The resolution of the surface wave data (~1º) is greater than the current lateral parameterization. The uncertainties of the group velocity measurements are derived from the broadness of the energy arrival and generally range from 0.05 km/s to 0.10 km/s for closer events.

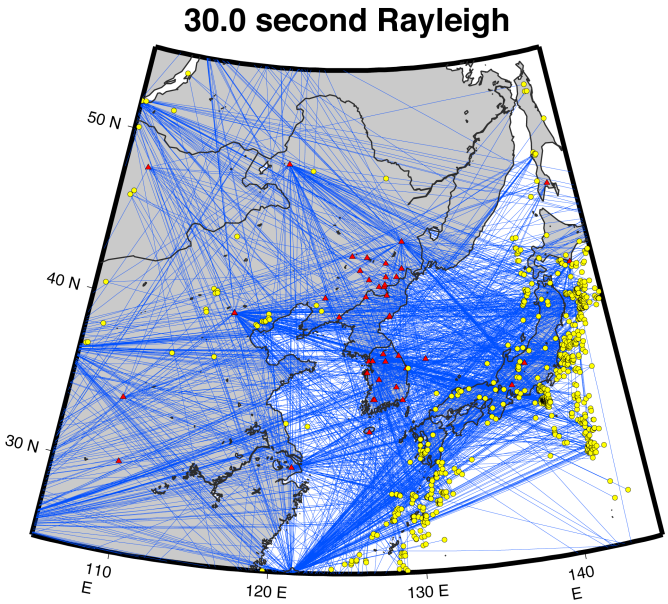

Figure 3.3. Path map of Rayleigh wave group velocity measurements at 30 sec. Path distributions at each period are slightly different.

The second data type that we use is body wave travel times, most notably Pn (P-velocity upper mantle head wave), Pg (crustal P-wave), Sn (S-velocity upper mantle head wave), and Lg (crustal guided S-wave). Paths for the various waves are shown in **Figure 3.4**. Coverage is generally not as good as the coverage of surface waves. This data, however, is complementary to the surface wave data both because some phases are more sensitive to P-wave velocities, but also because it samples the earth in a matter much different than the surface waves. For example, Pn is sensitive specifically to the P-wave velocity in the uppermost mantle, just below the Moho. By combining this with the surface wave data, we can reduce the non-uniqueness caused by the tradeoffs discussed above. The uncertainties of the travel times are derived from the phase and the estimated quality of the pick, and reflect the ambiguity of the timing of the arrival. First arriving P-phases generally have smaller uncertainties than phases picked in the coda of earlier-arriving phases.

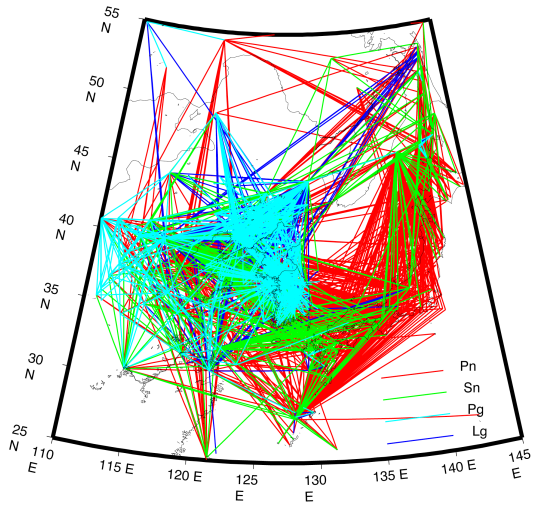

Figure 3.4. Path map of body wave travel times. Colors indicate Pg (cyan), Pn (red), Lg (blue), and Sn (green) phases.

While, at this point, we have only included Rayleigh wave group velocities and regional travel time measurements, it would be relatively easy to add more datasets in the inversion. Data sets that would be naturally suitable for extensions of this method include Love wave group velocities, receiver functions, and waveform modeling. The last would probably have the strongest constraints on the model along paths, but would also be the most computationally expensive to implement.

Results

We have made a run consisting of two chains each running 8000 iterations. The simulation was performed on a Compaq computer consisting of 512 processors with a processor speed of 667 MHz. For the run, we used 32 processors, with 16 processors dedicated to each chain. Although the codes have been parallelized, each iteration took an average of about 3.5 minutes, in total taking about 20 days. Separately, we have also run a two-chain "prior" distribution. In these runs, the models proposed by the base sampler are automatically accepted. Because no forward calculations are involved, these runs are extremely fast. They provide us with estimates of prior distribution properties, with which we can compare and contrast the posterior distribution's estimates. In both cases, we have thrown out models from the first 3/4 of the iterations in order to ensure that we are past burn in.

The first profile that we consider is at the southeastern corner of our model in the Pacific Ocean. **Figure 3.5** shows the profiles from the posterior distribution at that point. Colors indicate the profiles from the two different chains. The first thing to note is that the two chains seemed to be well-mixed, indicating that the convergence has been reached and that the chains have "forgotten" their starting point. The second thing to notes is that range of allowable models is very wide. This point, on the edge of our model, had little

or no data to truly test the proposed. In fact, if we look at the distribution of profiles in the prior, we find little or no difference. This effectively demonstrates, in fact, that in the absence of data, the model effectively reproduces the prior. Looking at histograms of parameters in the prior and posterior models (**Figure 3.6**), we find the same thing to be true. Also notice that the distribution of some parameters (like sediment thickness) is clearly non-Gaussian and any attempts to classify them as with mean and standard deviation as a Gaussian distribution would be inappropriate.

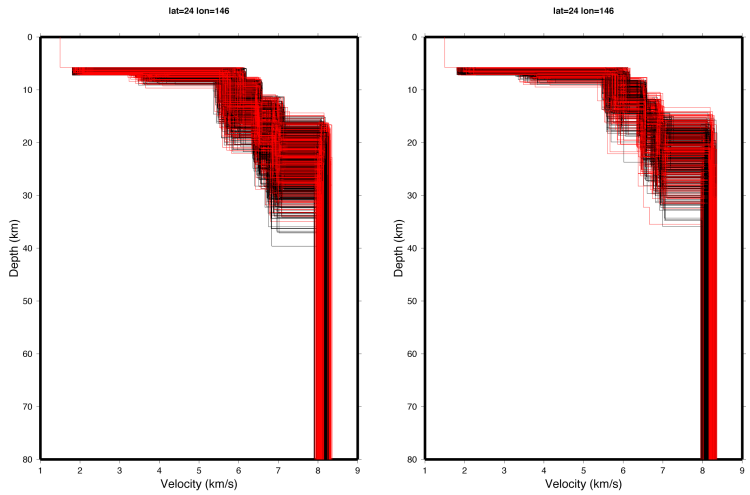

Figure 3.5. Prior (left) and posterior (right) model distributions for a point in the Pacific Ocean at the edge of our model. Red and black colors indicate results from the two chains.

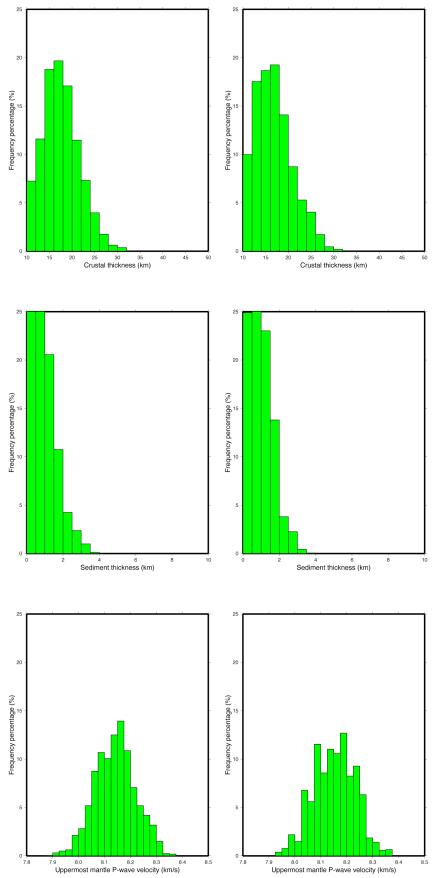

Figure 3.6. Histograms of model parameters for prior and posterior model for the point shown in the previous figure for parameters of crustal thickness, sediment thickness, and P-velocity of the uppermost mantle.

The second profile that we look at is one in the Yellow Sea at the center of our model. Here, we find the results to be quite different. **Figure 3.7** shows the profiles, once again demonstrating that the two chains are well-mixed. In this case, however, the range of allowable models has decreased significantly. The data which covers this region has preferentially sampled models which fit well. And the distribution of profiles is quite different than the distribution of the profiles in the prior. In the presence of data, then, the search preferentially samples regions of solution space that are most consistent with the data. Histograms of Moho depth, average crustal velocity, Pn velocity, and sediment thickness demonstrate the same (**Figure 3.8**). Not only have the peaks changed, but the distributions have gotten tighter, indicating lower uncertainties on these parameters. We find variance reductions of the parameters of 42%, 21%, and 56% for Pn velocity, crustal thickness, and sediment thickness. It is possible that these distributions might tighten even further as more iterations are made.

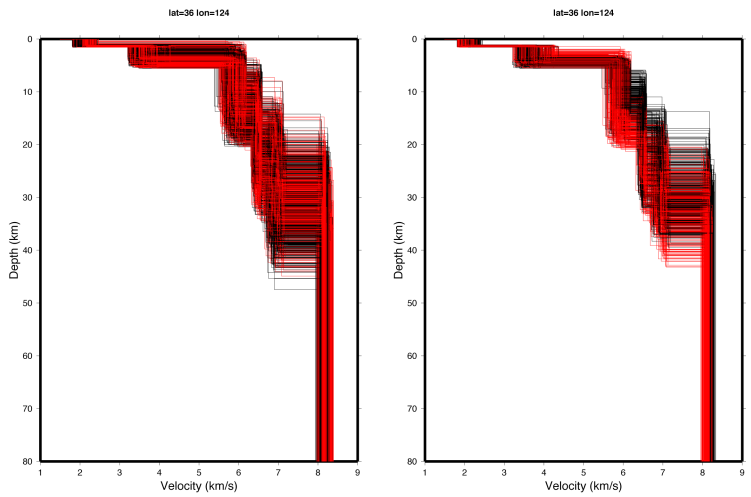

Figure 3.7. Prior (left) and posterior (right) model distributions for a point in the Yellow Sea at the center of our model.

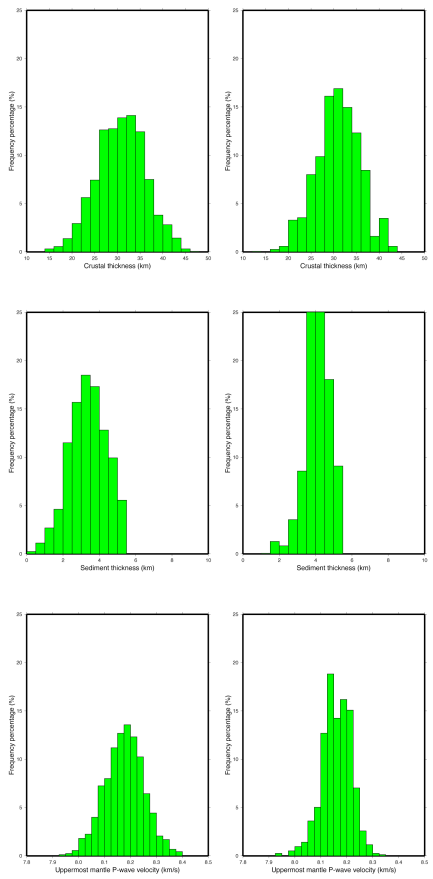

Figure 3.8. Histograms of model parameters for prior and posterior model for the point shown in the previous figure for parameters of crustal thickness, sediment thickness, and P-velocity of the uppermost mantle.

We have assembled the results from each point to create crustal thickness, sediment thickness, and average crustal velocity maps (**Figure 3.9**). In addition to the values, we have also plotted the uncertainties on each of these parameters, as determined by the standard deviation of the parameter for the posterior model. For crustal thickness, we find maps that are generally consistent with our understanding. We find differences from the prior model on the order of 5 km or so. Compared to the prior model, we are finding thicker crust in the Korean Peninsula and thinner crust in the Yellow Sea. Uncertainties are generally higher in the southwestern portion of our study area.

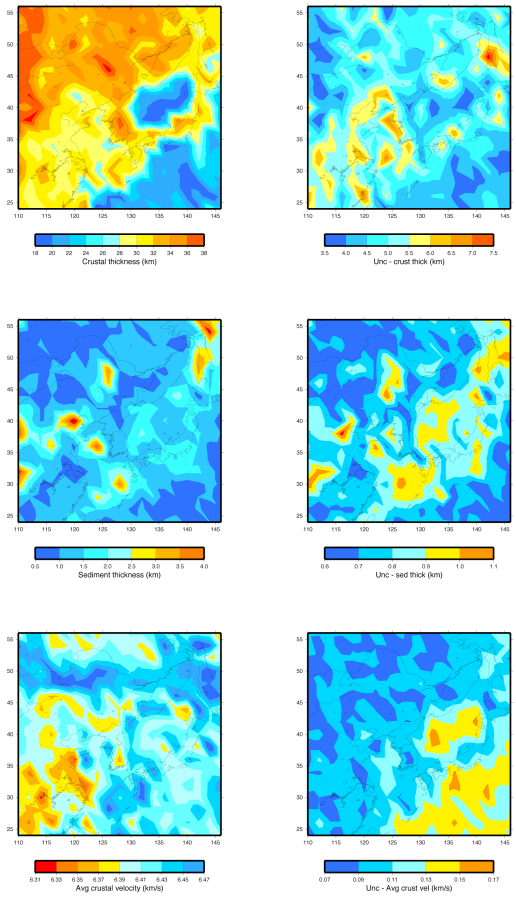

Figure 3.9. Maps of a) crustal thickness, b) upper mantle velocity and c) sediment thickness of the posterior model (left-hand columns), along with their associated uncertainties (right-hand columns).

Sediment thickness maps generally correspond closely to the prior model. There are differences, but they generally correspond to increases or decreases in thickness to existing sedimentary basins rather than the dissolution of existing basins or the creation of new basins. Differences on the order of a km or so are seen. Unsurprisingly, uncertainties are higher in basins and low in regions with little or no sedimentary cover.

Maps of average crustal thickness show generally low values in the Sino-Korean Paraplatform and higher in oceanic and other continental regions. Compared to the prior distribution, the velocities are slower in Chinese portions of the paraplatform, and faster in the Mongolia-Manchuria border area. There is a large difference in the uncertainties between oceanic and continental crust, with values being significantly higher in the former. This is likely due to the insensitivity of the surface wave data to changes in the thin oceanic crust.

Posterior Analysis

One significant advantage of stochastic models is their ability to properly estimate the uncertainties on observables. This is often hard to derive from single models because the uncertainties on any given model parameter (i.e. the thickness of the crust under point A or the upper mantle velocity at point B) is correlated with other model uncertainty and, often, these covariances are not taken into account. For example, while there might be a significant trade-off between certain parameters such as crustal thickness and crustal velocities, making the uncertainty on the individual parameters high, the overall uncertainty on the travel time through the uncertain media might actually be significantly less than that.

Stochastic models can handle this simply by estimating the observable through models in the posterior distribution, and noting the distribution of the observable. Here we show one example for predicting travel times. For every individual model in the posterior distribution, we have calculated the travel times of phases Pn, Pg, Sn, and Lg (**Figure 3.10**). The result is a distribution of predicted travel times for each phase, generally Gaussian in this case, that are the proper model uncertainties. For phase Pn, for example, we would estimate the velocity along this path as 7.68 ± 0.05 km/s.

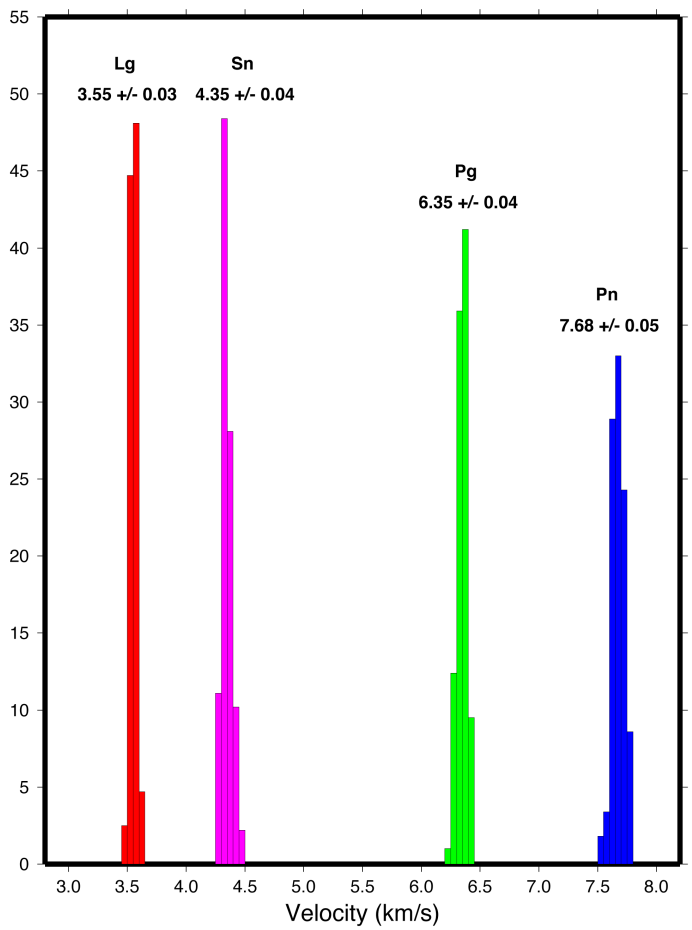

Figure 3.10. Predicted travel times of phases Pn, Pg, Sn, and Lg for a path. The travel times are distributions determined by calculating the travel time through all models in the posterior distribution.

We can take this one step further and use the predicted travel times to locate an event. Although the arrival times are the same for each model, there is a distribution of locations caused by variations in travel times. The resulting distribution of locations represents the location uncertainties caused by model errors. Picking errors, not shown here, would be considered separately.

We have taken 12 random models from the posterior distribution and used them to calculate a map of 30 second Rayleigh wave group velocity. The resulting group velocity map is shown in **Figure 3.11** and is compared to a tomographic inversion map at the same period. When comparing the two figures, keep in mind that while the group velocities predicted from models in the posterior distribution are fitting surface waves at all periods (as well as body waves), the tomographic inversion is only fitting data at this one particular period. Particularly given this fact, the fit is quite good.

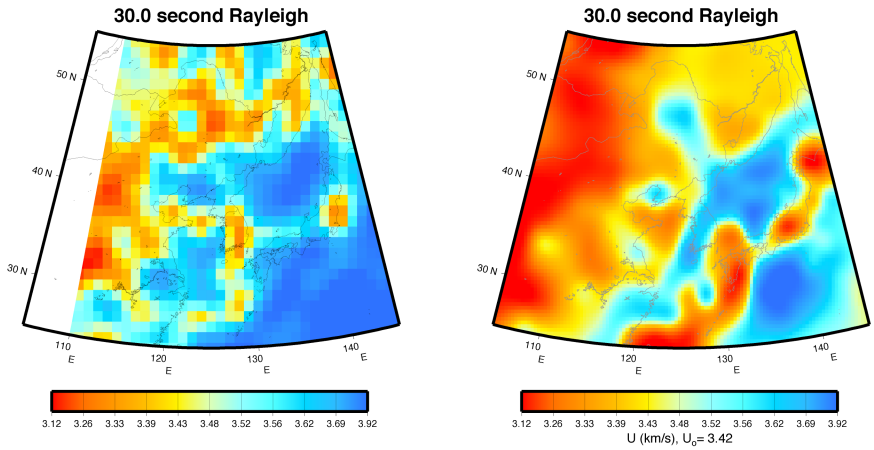

Figure 3.11. A comparison of the group velocities predicted by a posterior model to group velocities determined from the tomographic inversion of surface wave measurements at this period only.

Here, we provide an example of hypothesis testing that can be readily performed using our stochastic model. In nuclear explosion monitoring, the presence of a large P-wave and the absence of an S-wave are indicative of the event being an explosion. Propagation through oceanic crust, however, can block the Lg phase (the crustal guided S-wave) and make earthquakes look explosion-like. In other words, we wish to ask whether the lack of an Lg wave is due to source or structural effects.

In this case (**Figure 3.12)**, we have a bimodal distribution with peaks at 20 km and 35 km, each mode having about 5 km standard deviation uncertainties. The 35 km peak encompasses about 70% of the models, while the 20 km peak (which would block the Lg phase) encompasses the rest. In other words, if no Lg is observed along this path, there is a 30% chance that structural effects are the cause of the blockage, and that we have no information on the source.

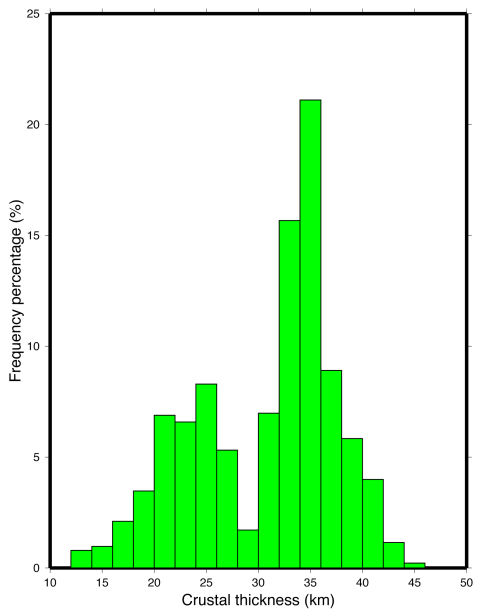

Figure 3.12. Histogram showing a bi-modal distribution of crustal thickness, with peaks at 20 km and 35 km.

Conclusions

Stochastic methods are an innovative technique for producing next-generation data-driven models. They have a number of advantages compared to traditional models, such as the ability to reconcile different types of geophysical data. An important component is the ability to predict new observables with proper uncertainties. We have started to apply this technique with success to the YSKP region. In the future, we aim to improve the model by incorporating more data sets (i.e. receiver functions, waveforms, etc.) and increasing the horizontal resolution.

Section 4: Future Applications

We have started applying or plan on applying the results of this LDRD to several other projects. First, several improvements to the model have been made since the conclusion of this project. The Ground-Based Nuclear Explosion Monitoring (GNEM) program has funded improvements to the regional model for the Yellow Sea – Korean Peninsula region. We have added more data types (such as gravity and Love waves) and more data for existing data types (such as more travel times), allowing us to increase the resolution of the model. Furthermore, there are plans to submit proposals responding to the next (Fiscal Year 2005) Solicitation for Proposals for "Nuclear Explosion Monitoring Research and Engineering" Broad Agency Announcement (BAA) from the National Nuclear Security Administration (NNSA), the Air Force Research Laboratory (AFRL) and the Army Space and Missile Defense Command (SMDC). These proposals would be working with other institutions (universities and private companies) to develop models for other regions of interest.

The second application is that the regional Korean model is being used as a base model for the NAI LDRD "A New Capability for Regional High-Frequency Seismic Wave Simulation In Realistic Three-Dimensional Earth Models To Improve Nuclear Explosion Monitoring" [05-ERD-019]. The motivation of this project is to improve current methods of seismic nuclear explosion monitoring, by substituting model-based calibrations for empirical measurements where such measurements are limited or do not exist. An aggressive approach to extending these strategies to broad regions would supplement or replace empirical calibrations with calibrations computed from geophysical (Earth) models. A model-based signal processing paradigm will require development of scientific and engineering enablers: 3-dimensional stochastic geophysical models, codes for calculating accurate, high-frequency (0.2 – 10 Hz) synthetic seismograms at ranges to 2000 kilometers and algorithms for exploiting realistic synthetics for event detection, location and identification. The benefits of model-based signal processing are anticipated to be an extension of the current parameter-based monitoring practice to broad regions without calibration data (aseismic regions), and very-low-magnitude event detection, location and identification capabilities in select portions of aseismic regions and in seismically-active areas with gaps in coverage. Long and short scale-length models specify the three-dimensional variation of the five elastodynamic parameters required to compute the response of the Earth to seismic excitation. These parameters are the seismic compressional (P-wave) and shear (S-wave) velocities (V_P & V_S), density (ρ) and attenuation quality factors (Q_P & Q_S, controlling attenuation of P- and S-wave, respectively). These parameters depend mostly on depth, due to chemical layering and central compression of the Earth. However, significant variation (10% or more near the Earth's surface) of these parameters results from thermal and compositional variations related to plate tectonics and geologic structure. We can use observations of seismic waves, such as travel times, surface wave dispersion, and other geophysical observations to estimate long scale-length models of the Earth using tomography and inversion techniques. Recently, we have pioneered the use of the Stochastic Engine for the estimation of regional-scale (~1000 km) models of

seismic velocity structure of the Earth on scale lengths as short as 100-200 km. This method uses Markov Chain Monte Carlo (MCMC) sampling to estimate the 3D structure of the Earth from multiple data sets and results in many model realizations that best fit the data as well as realistic uncertainties.

We are also interested in applying this technique and methodology to build regional models in southern Nevada. Models developed for this region would have applications for test-site readiness, namely how might recently-constructed buildings in Las Vegas fare with a resumption of nuclear testing at the nearby Nevada Test Site (NTS). By more accurately determining earth structure, we can better predict ground motions from scenario events. Models in this region could be built fairly easily since many of the same datasets and forward codes used in the YSKP region would also be applicable here. A stochastic model could be developed from six data sets (several which already exist): 1) geotechnical shear wave profiles, 2) seismic reflection profiles, 3) receiver functions, 4) surface wave dispersion, 5) body wave travel times, and 6) gravity.

Acknowledgements

This work was enabled by LLNL Laboratory Directed Research & Development (LDRD) research on the Stochastic Engine Initiative. We thank Roger Aines, John Nitao, and Bill Hanley for their contributions to the project. This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract W-7405-ENG-48. This is LLNL contribution UCRL-TR-209573.

References

Aines, Roger D., Robin L. Newmark, John J. Nitao, William G. Hanley, Steven Carle, Sailes Sengupta, David B. Harris, Ron Glaser, Richard A. Levine, Zhang Yong, and Kathleen M. Dyer, The Stochastic Engine Strategic Initiative.

Bassin, C. G. Laske, and G. Masters (2000), The current limits of resolution for surface wave tomography in North America, EOS Trans. AGU, 81, F897.

Carle, S. (2003), Integration of soft data into categorical geostatistical simulation, UCRL-JC-152653, Lawrence Livermore National Laboratory, Livermore, CA.

Crotwell, H.P., T.J. Owens, and J. Ritsema (1999), The TauP ToolKit: Flexible seismic travel-time and raypath utilities, *Seism. Res. Lett.,* 70, 154-160.

Davis. J.P. and I.H. Henson (1993), User's Guide to Xgbm: An X-Windows System to compute Gaussian beam synthetic seismograms. Teledyne Geotech Alexandria Laboratories, Alexandria, VA, 1.1 edition.

Gelman, A. (1996), Inference and monitoring convergence, in Markov Chain Monte Carlo in Practice, Chapman & Hall, New York, 131-143.

Gilks, W.R., S. Richardson, D.J. Spiegelhalter (1996), Introducing Markov chain Monte Carlo, in Markov Chain Monte Carlo in Practice, Chapman & Hall, New York, 1-19.

Hole, J.A. and B.C. Zelt (1995), 3-D Finite difference reflection travel times, *Geophys. J. Int.*, 121, 427-434.

Hubbard, S., Y. Rubin, and E. Majer (1999), Spatial correlation structure estimation using geophysical and hydrogeological data, *Water Resources Research*, vol. 35, pp. 1809-1825.

Keller, Randy (2003), GEON (GEOscience Network)-A First Step in Creating Cyberinfrastructure for the Geosciences, Electronic Seismologist, July/August 2003.

Kennett, B.L.N. (1985), *Seismic wave propagation in stratified media*, 342 pp., Cambridge University Press, Cambridge, England.

Komatitsch, D. and J. Tromp (1999). Introduction to the spectral element method for three-dimensional seismic wave propagation, *Geophys. J. Int.,* 139, 806-822.

Komatitsch, D., J. Ritsema and J. Tromp (2002). The Spectral-Element Method, Beowolf computing and global seismology, *Science*, 298, 1737-1742.

Larsen, S.C. and C.A. Schultz (1995), ELAS3D: 2D/3D Elastic Finite-Difference wave propagation code, Lawrence Livermore National Laboratory, 19pp.

Lutz, M. and D. Ascher (1999), *Learning Python*, O'Reilly and Assoc., Sebastopol, 366pp.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, (1953), Equation of state calculations by fast computing machines, J. Chem. Phys., 1, 1087-1092.

Mosegaard, K., and A. Tarantola (1995), Monte Carlo sampling of solutions to inverse problems, J. Geophys. Res., 100, 12431-12447.

National Geophysical Data Center (NGDC) (1998), ETOPO-5 Bathymetry/ Topography data, Data Announcement 88-MGG-02; National Oceanic and Atmospheric Administration, U.S. Department of Commerce.

Pasyanos, M.E., W.R. Walter, and M.P. Flanagan (2003), A Lithospheric Study of Eastern Asia using Surface Wave Dispersion, EOS Trans., AGU, 84, F1071.

Pasyanos, M.E. , G.A. Franz, and A.L. Ramirez (2004), Reconciling data using Markov Chain Monte Carlo: An application to the Yellow Sea - Korean Peninsula region, submitted to *Earth Planet. Sci. Lett*.

Ramirez, A. L., J.J. Nitao, W.G. Hanley, R.D. Aines, R.E. Glaser, S.K. Sengupta, K.M. Dyer, T.L. Hickling, W.D. Daily (2004), Stochastic Inversion of Electrical Resistivity Changes Using a Markov Chain, Monte Carlo Approach, in press, *J. Geophys. Res.*, UCRL-JRNL-155048 Rev. 2, Lawrence Livermore National Lab., Livermore CA.

Shapiro, N.M., and M.H. Ritzwoller (2002), Monte-Carlo inversion for a global shear-velocity model of the crust and upper mantle, Geophys. J. Int., 151, 88-105.

Tarantola, A. (1987), Inverse Problem Theory, Elsevier Science, Amsterdam, The Netherlands.

Taylor, S.R., X. Yang, and W.S. Phillips (2003), Bayesian Lg attenuation tomography applied to Eastern Asia, Bull. Seism. Soc. Amer., 93, 795-803.

Vidale, J., (1988) Finite-difference Calculation of Travel Times, *Bull. Seismol. Soc. Am.*, 78, 2062-2076.