JOURNAL OF GEOPHYSICAL RESEARCH: SOLID EARTH, VOL. 118, 1650–1676, doi:10.1002/jgrb.50123, 2013

3-D multi-observable probabilistic inversion for the compositional and thermal structure of the lithosphere and upper mantle. II: General methodology and resolution analysis

J. C. Afonso,[1] J. Fullea,[2,3] Y. Yang,[1] J. A. D. Connolly,[4] and A. G. Jones[2]

Received 14 June 2012; revised 4 February 2013; accepted 6 February 2013; published 8 April 2013.

[1] Here we present a 3-D multi-observable probabilistic inversion method, particularly designed for high-resolution (regional) thermal and compositional mapping of the lithosphere and sub-lithospheric upper mantle that circumvents the problems associated with traditional inversion methods. The key aspects of the method are as follows: (a) it exploits the increasing amount and quality of geophysical datasets; (b) it combines multiple geophysical observables (Rayleigh and Love dispersion curves, body-wave tomography, magnetotelluric, geothermal, petrological, gravity, elevation, and geoid) with different sensitivities to deep/shallow, thermal/compositional anomalies into a single thermodynamic-geophysical framework; (c) it uses a general probabilistic (Bayesian) formulation to appraise the data; (d) no initial model is needed; (e) compositional a priori information relies on robust statistical analyses of a large database of natural mantle samples; and (f) it provides a natural platform to estimate realistic uncertainties. In addition, the modular nature of the method/algorithm allows for incorporating or isolating specific forward operators according to available data. The strengths and limitations of the method are thoroughly explored with synthetic models. It is shown that the a posteriori probability density function (i.e., solution to the inverse problem) satisfactorily captures spatial variations in bulk composition and temperature with high resolution, as well as sharp discontinuities in these fields. Our results indicate that only temperature anomalies of $\Delta T \gtrsim 150$°C and large compositional anomalies of ΔMg# > 3 (or bulk $\Delta Al_2O_3 > 1.5$) can be expected to be resolved simultaneously when combining high-quality geophysical data. This resolving power is sufficient to explore some long-standing problems regarding the nature and evolution of the lithosphere (e.g., vertical stratification of cratonic mantle, compositional versus temperature signatures in seismic velocities, etc) and offers new opportunities for joint studies of the structure of the upper mantle with unprecedented resolution.

Citation: Afonso, J. C., J. Fullea, Y. Yang, J. A. D. Connolly, and A. G. Jones (2013), 3-D multi-observable probabilistic inversion for the compositional and thermal structure of the lithosphere and upper mantle. II: general methodology and resolution analysis, *J. Geophys. Res. Solid Earth*, *118*, 1650–1676, doi:10.1002/jgrb.50123.

1. Introduction

[2] There are two main sources of information available to constrain the present-day physical state and chemical composition of the Earth's lithosphere and sub-lithospheric upper mantle: the interpretation of geophysical observables (e.g., gravity anomalies, traveltime data, surface heat flow, etc.) and studies on exhumed mantle samples (e.g., xenoliths, tectonically exposed mantle bodies). Experimental petrology/mineralogy [cf. *Ringwood*, 1975; *Herzberg*, 1999; *Walter*, 2004] and numerical simulations [e.g., *Tackley and Xie*, 2002; *Brandenburg et al.*, 2008; *Nakagawa et al.*, 2010; *Tirone et al.*, 2012] offer an important complement, but they cannot constrain the present-day compositional and temperature structure of the upper mantle per se. Both geophysical observables and exhumed mantle samples have their own set of, virtually independent, advantages and limitations when used to make inferences about the upper mantle. Geophysical observables offer a larger and more continuous spatial coverage, but their conversion into estimates of composition and temperature is populated with difficulties [e.g., *Jones-Evans-and-Eaton*,

[1]GEMOC-CCFS, Department of Earth and Planetary Sciences, Macquarie University, Sydney, Australia.
[2]Dublin Institute for Advanced Studies, Dublin, Ireland.
[3]Institute of Geosciences (IGEO) CSIC-UCM José Antonio Novais, Madrid, Spain.
[4]Institute of Geochemistry and Petrology, ETH Zurich, Switzerland.

Corresponding author: J. C. Afonso, GEMOC-CCFS, Department of Earth and Planetary Sciences, Macquarie University, Sydney, Australia. (juan.afonso@mq.edu.au)

2169-9313/13/10.1002/jgrb.50123

2009; *Afonso et al.*, 2010; *Cammarano et al.*, 2011; *Becker*, 2012; *Afonso and Schutt*, 2012, Afonso et al., Paper I, this issue]. Exhumed mantle samples, on the other hand, represent direct evidence of the compositional and thermal structure of the upper mantle at the time of exhumation. However, their spatial and temporal coverage are highly discontinuous, their original compositions may have been modified by or during the exhumation process, and there may be an inherent bias in that large mantle samples (> 40–60 cm in diameter) may be too heavy for transportation to the surface by magmatic events [*O'Reilly and Griffin*, 2010]. Ideally, these sources should be used together to provide a more complete picture of the thermochemical state of the Earth's upper mantle than that resulting from the evaluation of them individually.

[3] The main question is then how to combine both sources of information into an internally consistent, general, and objective inverse framework. By internal consistency we mean that all physical and chemical parameters are formally linked together by a unique and sound physicochemical model. Likewise, generality and objectivity require the inverse framework to be applicable to a large number of scenarios (e.g., different tectonic settings, different terrestrial bodies) and to a multitude of data with the minimum (necessary) input of subjective information. Probabilistic formulations (statistical inference) of such inverse problems are particularly well suited for the task at hand, simply because the actual solution to the problem is probabilistic in nature. Since (i) the available information from geophysical measurements and/or exhumed mantle samples is always incomplete and subject to uncertainties, (ii) the physical theories we use to make predictions may be imperfect, and (iii) the sensitivity of our measurements to variations in some of the parameters of interest are nonlinear and relatively weak, any inference made on the physical state of the Earth's mantle is necessarily probabilistic. In other words, the best one can, and should, achieve is to obtain a probability density over the parameters of interest, rather than a single, necessarily non-unique solution [cf. *Tarantola and Valette*, 1982; *Mosegaard and Tarantola*, 2002; *Congdon*, 2006; *Biegler et al.*, 2011].

[4] Within the context of statistical inference, the most general solution to the inverse problem is represented by a joint probability density function (PDF) in the parameter and data space, $\sigma(\mathbf{d}, \mathbf{m})$, known as the a posteriori PDF [cf. *Mosegaard and Tarantola*, 2002; *Tarantola*, 2005; *Congdon*, 2006]. This PDF summarizes all the information we have on both the model parameters **m** and the observable data **d**, and represents our best "state of knowledge." Under most practical circumstances, we can integrate out the data vector component from $\sigma(\mathbf{d}, \mathbf{m})$ to obtain the marginal PDF in the model space as [see *Mosegaard and Tarantola*, 2002; *Tarantola*, 2005; Afonso et al., this issue]

$$\sigma(\mathbf{m}) \propto \rho(\mathbf{m})\, L(\mathbf{m}) \tag{1}$$

with

$$L(\mathbf{m}) \propto \int \rho(\mathbf{d})\, \theta(\mathbf{d} \mid \mathbf{m})\, d\mathbf{d} \tag{2}$$

where $\rho(\mathbf{m})$ is the a priori PDF describing all the information in the parameter space that is independent from the actual measurements, $L(\mathbf{m})$ is the so-called *likelihood function*, which describes the probability of obtaining the observed data **d** given **m** (i.e., a measure of how good the model **m** is in explaining the data), $\rho(\mathbf{d})$ the a priori PDF describing data uncertainties, and $\theta(\mathbf{d} \mid \mathbf{m})$ the conditional probability describing the uncertainties associated with predictions from our theoretical models. In the general case, where there are no formal mathematical expressions for these PDFs and the number of parameters is large, a sampling scheme based on a Monte Carlo analysis is necessary. The computational cost associated with such probabilistic analysis, however, can be notoriously large. Traditional linearized inversion methods would be preferable given their computational efficiency; however, they are not well suited to deal with (i) the non-linearity of the system, (ii) the much larger sensitivity of geophysical observables to temperature than to compositional anomalies, (iii) the non-uniqueness of the compositional and thermal fields (more than one valid solution), (iv) the poorly known correlations between modeled physical parameters and geophysical observables, and (v) the well-known trade-off between temperature and composition in wave speeds [e.g., *Khan et al.*, 2007; Afonso et al., this issue]. These are all intrinsic issues related to the current problem of inverting directly for temperature and composition, and thus, any method attempting to provide reliable estimates of the thermochemical structure of the lithosphere and sub-lithospheric upper mantle must address them.

[5] Multi-observable probabilistic inversions based on thermodynamic grounds offers a robust means to invert for the thermochemical structure of the Earth's mantle and has the potential to minimize, and in some circumstances overcome, the above-mentioned problems. However, such multi-observable probabilistic schemes are largely an understudied field. The closest method to the one presented in this paper is that of *Khan et al.* [2007, 2011a, 2011b]. The Bayesian method used by these authors is truly nonlinear and thermodynamically self-consistent, and it represents one of the most advanced and well-suited approaches available. Our present implementation of the Bayesian method differs from that of *Khan et al.* [2007, 2011a, 2011b] in (i) the treatment of a priori information on the compositional space [Afonso et al., paper I, this issue], (ii) the integration of more (3-D) geophysical observables, and (iii) the way in which we account for theoretical (modeling) uncertainties. These differences are mostly related to, and justified by, the different scales of interest. While *Khan et al.* [2011a, 2011b] deals with large-scale whole-mantle structures, our interest is on smaller scale ($\lesssim$ 100–200 km) structures, suitable for regional ($\lesssim 1.0 \times 10^6$ km^2) studies of the lithospheric and sub-lithospheric upper mantle.

[6] In this work, we include the simultaneous, internally consistent, inversion of gravity and geoid anomalies, surface heat flow, electrical conductivity, absolute elevation, Rayleigh and Love dispersion data, body-wave data, and petrological data within a probabilistic (Bayesian) context. We have purposely chosen these observables because they are differently sensitive to shallow/deep, thermal/compositional structure and anomalies, which allows superior control of the lateral and vertical variations of the bulk properties. Our implementation is specifically designed to circumvent the aforementioned problems affecting traditional linearized inversions. It also offers critical insights into the incompatibilities between predictions from

stand-alone methods (e.g., seismic versus MT studies) and it can easily incorporate a priori geochemical/petrological information, when such information is available and have reliable error estimates. Importantly, given the modular nature of the method, it can be applied to either all of the abovementioned observables (simultaneously) or to a subset of them, depending on data availability.

[7] This paper focuses on the general inversion methodology and its computational implementation. The advantages and limitations of the method are thoroughly discussed and illustrated with synthetic examples. The information that controls the estimations of $L(\mathbf{m})$ and $\rho(\mathbf{m})$ in equation (1) are discussed in detail in an accompanying paper [Afonso et al., this issue]. Readers interested in the entire methodology will benefit from reading part I first. We emphasize here that the new probabilistic method described in these two papers is not new in the sense of the Bayesian formulation but rather in the practical implementation/adaptation to solve a specific inversion problem (strictly, make inferences about the problem) that seeks to provide reliable estimates of the temperature and bulk composition of the Earth's upper mantle. The application of the method to a real case scenario will be presented in a forthcoming publication (Afonso et al., Part III).

2. General Methodology

[8] In the most general case, there will be no explicit mathematical expressions to calculate $\rho(\mathbf{m})$ and/or $\sigma(\mathbf{m})$, and exhaustive Monte Carlo sampling will be needed [*Mosegaard and Tarantola*, 1995]. This is particularly relevant for the 3-D problem discussed in this paper, since using complicated priors and Monte Carlo sampling to invert for major-element composition and temperature in true 3-D with high resolution is practically impossible. This is simply a consequence of both the large number of parameters needed to define a full 3-D model (see below) and the much larger number of samples for which the 3-D forward problem needs to be solved in order to obtain reliable estimates of the posterior $\sigma(\mathbf{m})$.

[9] In order to make the general problem tractable, we choose to divide the inversion into *two main stages* (Figure 1). In the first stage, we subdivide the 3-D model into its constituent 1-D rectangular columns (Figure 2) and invert a *subset* of the data vector associated with each column individually. As a result of this column-by-column inversion, we obtain first-order estimates of marginal PDFs describing 1-D thermochemical structures that are compatible with 1-D data. We then use these PDFs as a priori information in a second "refinement stage" in which the full 3-D problem is addressed within a formal Bayesian formalism (explained below). This two-stage approach shares some similarities with the so-called "empirical" and "hierarchical" Bayes methods [cf. *Carlin and Louis*, 2000]. The main characteristic that distinguishes empirical from full Bayes methods is the way in which the priors are assigned or defined. In the full Bayes approach, only priors that are defined before the inclusion of any observation are considered true priors (i.e., information independent of data). The empirical approach, on the other hand, relaxes this requirement and allows some or all available data to inform the priors defining the parameters of interest. Without entering the controversial debate of empirical versus full Bayes approaches [e.g., *Scales and Snider*, 1997; *Scales and Tenorio*, 2001; *Kitanidis*, 2011, and references therein], we note that different forms of the empirical

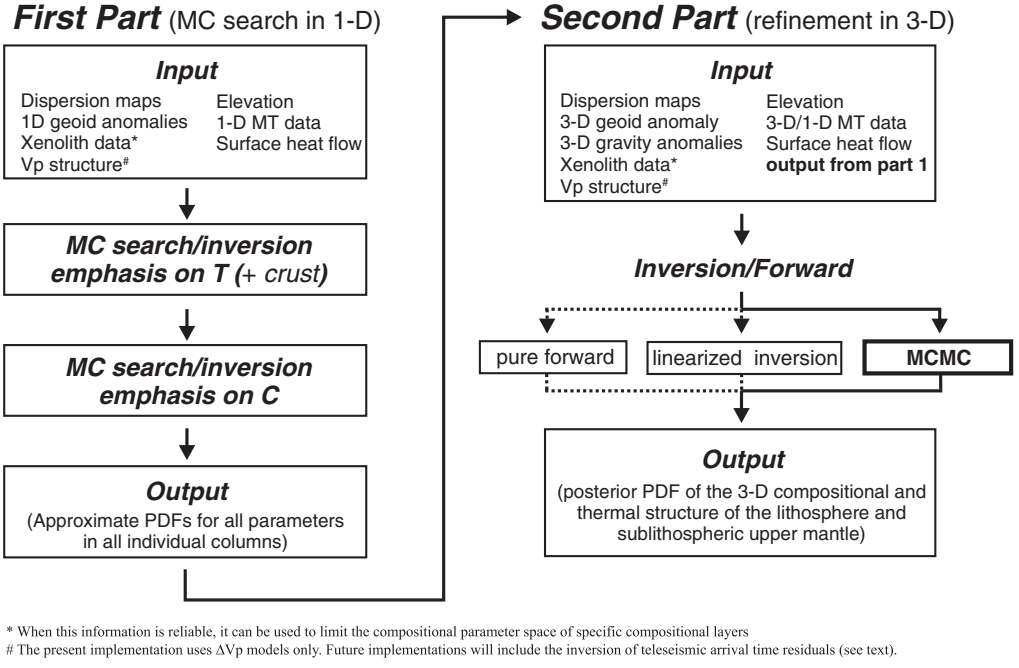

Figure 1. Flow chart illustrating the two-part inversion approach presented in this study. In this paper, we only describe the Markov Chain Monte Carlo inversion method (solid lines) for the second part. Other potential approaches, such as pure forward or linearized inversions (dashed lines), are simpler and more restrictive and therefore will not be discussed here.

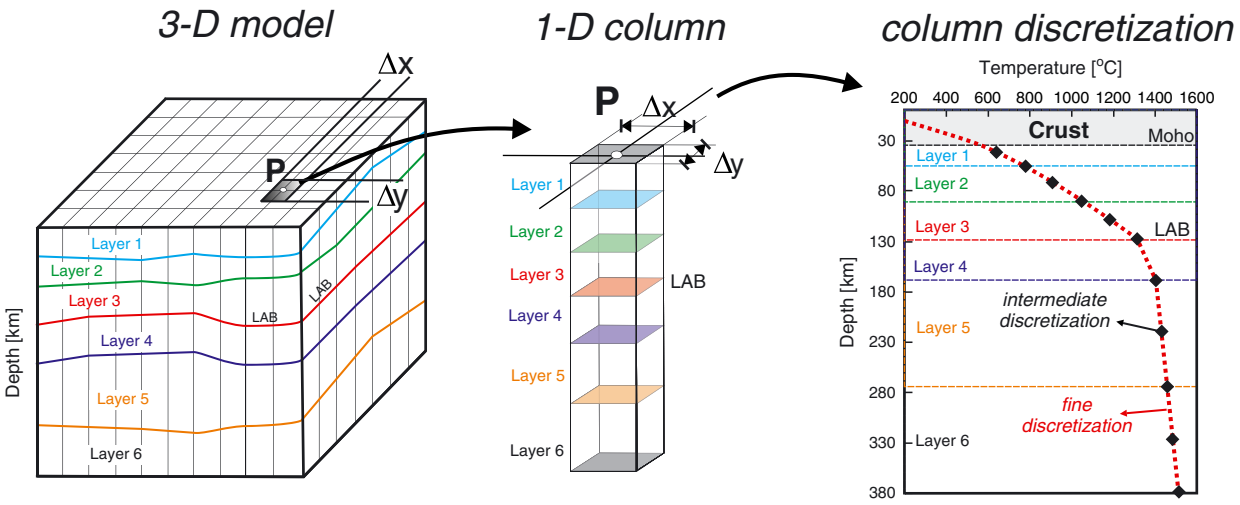

Figure 2. Discretization scales used in the inversion. A 3-D model can be thought of as made up of a collection of 1-D columns. Each column has three discretization scales. The finest discretization scale (*computation nodes*) is used in the numerical solution of the forward problem, both in 1-D and 3-D. The vertical grid step is typically set to ≤ 2 km. The intermediate discretization scale (*thermodynamic nodes*) is used in the solution of the Gibbs free energy minimization problem. Typically, 10–20 nodes are used in the intermediate discretization. The third discretization scale (*compositional layers*) refers to the actual number of independent compositional layers allowed in the model. The number of compositional layers in most practical applications varies between 6 and 8. See text for details.

Bayes approach have been proven to be some of the most practical and effective tools for the solution of geophysical inverse problems [e.g., *Gouveia and Scales*, 1998; *Oh and Kwon*, 2001; *Malinverno and Briggs*, 2004; *Sambridge et al.*, 2006; *Kitanidis*, 2011]. Although we do not use a formal empirical Bayes formalism, we implicitly borrow concepts from the empirical paradigm by allowing some of the data to define, in an approximate manner, the priors used in the final 3-D stage.

[10] The described division of the inversion scheme into a preliminary 1-D stage and a final 3-D stage is sensible for the following two reasons. First, the observables used in the first 1-D stage (dispersion maps, elevation, MT data, 1-D geoid anomaly, surface heat flow) are largely controlled by the 1-D structure of the lithosphere and upper mantle. Second, the initial 3-D models constructed by assembling the best-fitting 1-D columns, and used as starting models in the 3-D refinement stage, typically represent already acceptable models (i.e., they explain most of the data well). We discuss these two main stages further in the following sections.

3. 1-D Column-by-Column Inversion

[11] As noted above and in Paper I, the response of our system (i.e., predicted observables) to perturbations in the parameter space (e.g., Al_2O_3 content, depth to the lithosphere-asthenosphere boundary, etc.) is nonlinear, and the number of free parameters and their initial variation ranges are large. We therefore are forced to rely on an extensive sampling of the parameter space through stochastic nonlinear techniques. To this end, here we adopt a modified version of the direct search algorithm known as the Neighborhood Algorithm (NA) [*Sambridge*, 1999a]. The objective of the search is to find models in the entire multi-dimensional model space that are compatible, to a satisfactory degree, with observed data. The original version of the NA does this by randomly sampling the entire parameter space, guided by the fitting properties (good and bad) of all previous samples [*Sambridge*, 1999a]. This self-guided non-uniform sampling approach concentrates the sampling around the regions (neighborhoods) where the models best fit the data and avoids unnecessary sampling in areas where the fit is not optimal. In this way, the entire model space is well sampled, computational efficiency is maximized, and the problem of entrapment in local minima is reduced. In an optional second stage (the appraisal stage), the entire ensemble of "acceptable" models obtained during the search stage is re-sampled and analyzed using a Bayesian (probabilistic) inference framework [*Sambridge*, 1999b]. We do not use this capability here, since our final Bayesian analysis is performed during the 3-D stage.

[12] The original NA requires four main control parameters to be defined prior to execution: the number of initial samples to be generated (N_i), the maximum number of iterations (*maxit*), the total number of random models generated in each iteration of the search (N_s), and the number of best-fitting neighbor cells to be re-sampled per iteration (N_r). For instance, if $N_i = 100$, *maxit* = 20, $N_s = 100$, and $N_r = 10$, then the algorithm will create 100 samples randomly distributed in the n-dimensional parameter space, and in each successive iteration, only the 10 best-fitting cells will be re-sampled by generating 10 new models within each of these cells (N_s / N_r). The total number of models generated during the complete search is 2100 ($N_i + N_s \times maxit$). The values needed for these parameters in order to perform a "good" or "successful" search are strongly dependent on the actual problem at hand, and they must be chosen based on trial

experiments. In general, large values for N_s and N_r makes the algorithm more explorative, thus reducing the problem of converging to local minima and guaranteeing good sampling of the entire parameter space. This is particularly relevant to our problem, since the search process, which includes solving the forward problem, must be repeated for each column making up our model (typically between 100 and 400 columns) before we can construct the 3-D model for the next inversion stage. Therefore, the tuning of these four parameters to warrant both computational efficiency and acceptable solutions is critical. In order to maximize the efficiency of the search, and customize the NA to our problem, we have modified the original NA in two ways. Each of these modifications will be introduced below as the relevant problems are being discussed.

3.1. The Forward Problem in 1-D: Parameters and Discretization Scales

[13] At this point, we need to define (i) the variables or parameters to be used during the search of the parameter space and (ii) the numerical discretization used to solve the forward problem of each realization or sample. The parameters that we choose to define a particular 1-D realization are the thermal thickness of the lithosphere (i.e., depth to the lithosphere-asthenosphere boundary or LAB) and the major-oxide compositions of each of the *n*-compositional layers defined for the simulation (Figure 2). We use the depth to the LAB as an independent parameter because it defines the first-order thermal structure of the column (section 3.2.1). Each column typically would have six to eight compositional layers of variable thickness; the number of compositional layers is kept constant in all the columns during the entire inversion procedure. Each of these layers is defined by its five major oxides within the system CaO-FeO-MgO-Al_2O_3-SiO_2 (strictly, since CaO+FeO+MgO+Al_2O_3 +SiO_2 = 100, only four oxides are needed). We choose this system instead of the more complete system Na2O-CFMAS because the added computational cost does not translate in any significant improvement of the results (Appendix B in Paper I). We have modified the original NA to sample the compositional space following the method discussed in Paper I (Afonso et al., this issue). Briefly, this method uses two "key" oxides (Al_2O_3 and FeO) as independent parameters (Figures 3a–3c and 4a–4c in Paper I) and additional, dependent, Gaussian distributions (Figures 3d, 3d, 4d, and 4e in Paper I) to sample the remaining CaO and MgO according to their occurrence probabilities. This approach is able to generate independent random samples that cover the entire compositional spectrum observed in real mantle samples without oversampling low-probability regions.

[14] In the present implementation, the solution of the 1-D forward problem depends on three different discretization scales (Figure 2). The finest discretization scale is used in the numerical solution of the heat-transfer, potential field, isostasy, Love and Rayleigh dispersion curves, and Maxwell's equations (Appendix A). We will refer to the nodes making up this mesh as the *computation nodes*. In this case, the grid step is typically set to $\leq$ 2 km. The intermediate discretization scale is used to calculate equilibrium assemblages and associated thermophysical properties (e.g., seismic velocities, compressibility, bulk density, etc.) by free energy minimization at specific P-T-C conditions. We use components of the software Perple_X [*Connolly*, 2009] to solve the Gibbs free energy minimization problem together with the database and thermodynamic formalism of *Stixrude and Lithgow-Bertelloni* [2011]. The number of nodes used in this discretization varies between 10 and 20, and we will refer to them as the *thermodynamic nodes*. All properties calculated at the thermodynamic nodes are then interpolated to the computation nodes during the iterative solution of the forward problem (see below). The third, and coarser, discretization scale refers to the actual number of independent *compositional layers* used in each column (Figure 2). Typically, each compositional layer will include at least two thermodynamic nodes. When the location of a thermodynamic node coincides exactly with a limit of a compositional layer, the composition of the node is assumed to be that of the compositional layer immediately above it. Since the bulk composition in each layer is constant, compositional anomalies of smaller scale than that defined by the number of compositional layers assumed in the inversion will not be resolvable. In this context, six to eight layers usually suffice given the resolution of the observables to vertical gradients in composition.

[15] In order to compute the stable assemblages (by free energy minimization) and their associated properties in the thermodynamic nodes, one needs to know the temperature, pressure, and composition of each node. This is straightforward in the case of temperature and composition (Appendix A), since they are the independent variables. Pressure, on the other hand, depends on the integrated bulk density of the column, which in turn depends on pressure, temperature, and the compositional structure of the column. Therefore, an iterative scheme must be used to compute the lithostatic pressure at the nodes [e.g., *Fullea et al.*, 2009]. This is particularly important when calculating absolute values of physical properties, such as density or elastic moduli, since the accumulated error in pressure when using non-iterative schemes can reach values > 500 MPa at depths > 220 km (i.e., > 7% error). This is equivalent to an error of $\sim$ 15 km in node depth. Likewise, neglecting changes in the acceleration of gravity with depth can introduce systematic errors in the calculation of pressure, albeit significantly smaller ($\sim$ 100 MPa at 380 km depth, equivalent to 0.8% error). We therefore use a linear correction factor in our computations to account for this effect. The density-pressure iterations are stopped when the difference in calculated pressure between two successive iterations becomes less than 70 MPa (i.e., assumed error in pressure). This is equivalent to an uncertainty of only $\sim$ 2 km in the node depth. In practice, one or more "master pressure profiles" can be constructed a priori and used as the initial guesses in the iteration scheme. We find that by correctly tuning these profiles, the number of necessary iterations during the 1-D forward problem can be reduced to 0 at shallow levels and to 1 or 2 at the deepest levels (z > 300 km depth).

3.2. Temperature Versus Compositional Parameters

[16] It is important to know the sensitivities of the observables to the different parameters in the model in order to devise an efficient implementation of the 1-D search. As discussed in Paper I, the sensitivity of geophysical observables to changes in bulk composition is much weaker than to variations in temperature, or equivalently, to variations in

the LAB depth. Moreover, different bulk compositions can result in similar physical properties of the aggregate, thus offering equally good fits to geophysical observables.

[17] The problem of having a system in which the individual contributions of the model parameters to the global likelihood are significantly different (i.e., data with different sensitivities) is illustrated in Figure 3. In this simplified example with only two parameters, observations (data) are strongly sensitive to changes in parameter m_1 but only slightly sensitive to parameter m_0. As a result, if the initial random sampling (e.g., N_i in the NA algorithm) does not produce models close to the true values of the two parameters, a situation can arise in which the proximity to the true value of m_1 in the initial sampling (case shown in Figures 3a–3c) bias the evolution of the search algorithm to wrong values of m_0. This occurs because the combined likelihood $L(m_0, m_1)$ is dominated by the contribution from m_1 (Figure 3d). Consequently, those models that are close to the true value of m_0, but relatively far from the true value of m_1, are left unsampled in subsequent iterations (Figures 3b and 3c). Although Markov Chain Monte Carlo (MCMC) methods are arguably better suited to handle this problem, a similar situation could arise if a MCMC simulation without optimal adaptation is not let to run long enough to guarantee good sampling (Figure 3e). This is particularly relevant when the prior is vague and the proposal distribution inaccurate [cf. *Gilks et al.*, 1996; *Gregory*, 2005].

[18] In principle, the above-mentioned difficulty could be overcome or minimized by (i) increasing the values of N_i, N_s and N_r in the NA (defined at the beginning of section 3), (ii) applying weighting factors to $L(m_0, m_1)$ in order to increase the sensitivity to m_0, or (iii) by subdividing the search in two parts, each designed to maximize the differential sensitivities. While (i) would be possible in low-dimensional cases, it rapidly becomes impractical as the dimensionality of the parameter space increases [e.g., *Sambridge*, 1998]. Similarly, weighting $L(\mathbf{m})$ adds an artificial and arbitrary effect to the system that, in most cases, is difficult to justify objectively. Arguably, (iii) represents a better strategy as long as both the forward problem and the misfit function can be specifically tailored to take advantage of the different sensitivities to m_0 and m_1 (somewhat similar to the "cascade" approach of *Mosegaard and Tarantola* [1995]). In this way, the first search focuses on constraining acceptable models using only parameter m_1 (i.e., m_0 taken constant) and a simplified forward problem. Given the strong sensitivity of the data to m_1 (Figure 3d), we only need to generate a modest number of models in the initial sample (inasmuch as they are uniformly distributed) to greatly reduce the initial range of potential values for m_1, as constrained by the data. In the second part, the search algorithm is focused on sampling parameter m_0 more extensively by allowing only small variations in m_1 (already constrained during the first stage; Figure 3f).

[19] A similar, but more problematic, issue is having more than one possible bulk composition fitting the data equally well. In Paper I, we showed that samples with almost identical physical properties tend to form a trend in the compositional space, particularly at pressures pertaining to lithospheric depths. In the same paper, we discussed a probabilistic approach that succeeded in reducing the ensemble of acceptable samples consistent with the data. However, one critical aspect upon which the method relies is that the search performed in the first stage (which defines the priors for the second stage) must sample well the entire compositional space, including the trends of acceptable models (*Afonso et al*., Paper I, this issue). In the context of direct search methods such as the NA, if the original sample is too small to properly explore the region defined by the trend of acceptable samples, successive iterations could concentrate the search only in some parts of the acceptable region that nevertheless do not contain the true value (Figures 3g–3i). Short MCMC simulations with vague priors and/or proposal distributions could potentially be affected by the same problem.

[20] Although the above illustrative examples are deliberately simplified, they do portray real complications identified during extensive trial-and-error experiments. Based on this experience, and considering the differential sensitivities of observables to temperature and composition, we subdivide the 1-D column-by-column inversion stage into two parts. The emphasis of the first part is on temperature structures (the main parameter affecting observables), while the second is designed to explore more thoroughly the compositional space.

3.2.1. 1-D Search With Emphasis on Temperature Structure

[21] The search in this first stage involves generating random samples (via NA) of LAB depths within the range $50 < \text{LAB} < 380$ km. This range is large enough to include most of the variability commonly thought to exist in continental settings [e.g., *Pasyanos*, 2010; *Fishwick*, 2010; *Hasterok and Chapman*, 2011, and references therein]. We also generate samples of the *average* composition of the lithosphere (i.e., the model contain only one compositional layer; the composition of the sub-lithospheric mantle is assumed constant). This is necessary to avoid the potential, albeit unlikely, problem that arises when the actual (real) average lithospheric composition is significantly different from the assumed one. In such a case, the search/inversion could compensate the large compositional contrast with relatively large and erroneous variations in the LAB or in the internal thermal structure of the lithosphere.

[22] Once a particular LAB depth and average lithospheric composition have been chosen, we solve the 1-D heat transfer equation under the assumptions of steady state in the lithosphere and adiabatic gradient in the sub-lithospheric upper mantle (Appendix A for numerical details). Although these are standard assumptions in thermal modeling of the lithosphere [cf. *Jaupart and Mareschal*, 2011], in general one could expect deviations from these assumptions both within the lithosphere (e.g., transient effects, local heating) and in the sub-lithospheric mantle (e.g., hot plume material). In order to account for this possibility, and to keep the search as general as possible, we introduce additional variables to the parameter space. These new variables are discrete temperature deviations ΔT_n from the conductive geotherm within the lithosphere, and discrete deviations from the nominal adiabatic profile $\Delta \mathrm{d}T/\mathrm{d}z_n$ in the sub-lithospheric mantle. Since we only allow the algorithm to propose finite increments of $\Delta T_n = 100$°C and set the minimum possible departure from the conductive geotherms to 200°C, we force the search to be sensitive to

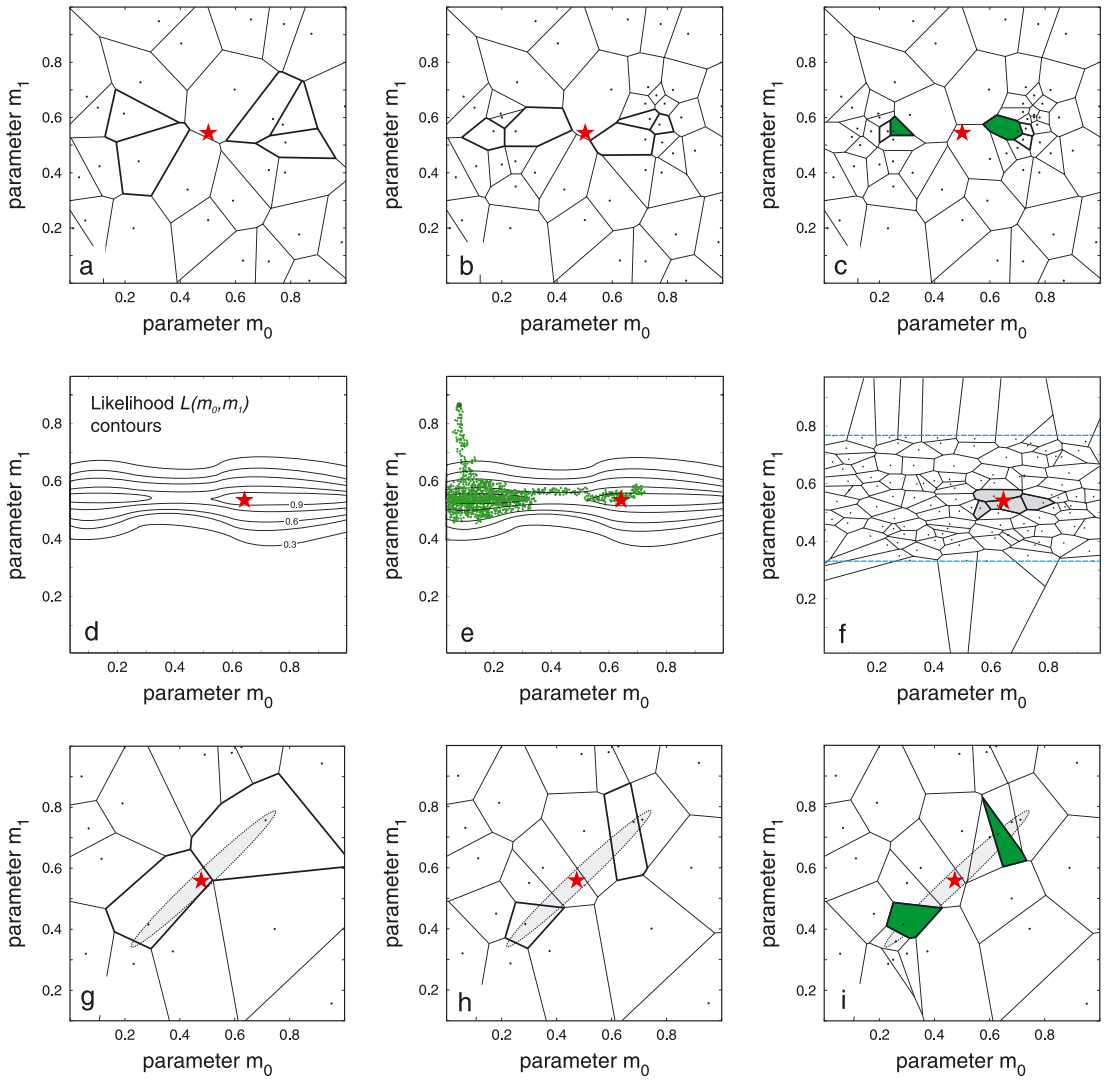

Figure 3. (a–c) A direct search of a two-dimensional parameter space is performed with the Neighborhood Algorithm [*Sambridge*, 1999a]. The initial sample (a) is made up of 30 random samples (N_i = 30). Only the best five cells (thick borders) are chosen for resampling (N_r = 5), and three new random samples are created in each cell (N_s = 15). Given the strong sensitivity of the data to parameter m_1, the sampling has been concentrated in two regions that define m_1 well but do not include the true solution (red star). (d) The total likelihood of the system $L(m_0, m_1)$ is dominated by the contribution from parameter m_1. (e) A MCMC simulation with vague priors and an inadequate proposal distribution converges rapidly to a good value of m_1 but slowly toward good values of m_0. If the chain is not let to run until global convergence, the final result can be misleading. (f) Similar to Figures 3a–3c, but now the search has been divided into two parts. The first part significantly narrowed the original variation range of m_1, and therefore, the sampling in the second part is focused on parameter m_0. In this case, 50 samples have been enough to locate the true solution (red star), as indicated by the five best-fitting cells (in gray). (g–i) The gray ellipse denotes the region of models with identical fitting properties. Fifteen samples are generated in the first iteration (g) and the two best cells (thick borders) are chosen for resampling (h). Since the original sampling did not sample well the gray region, subsequent sampling is concentrated in an area of acceptable fitting that nonetheless does not contain the true value.

only relatively large temperature anomalies. Similarly, for the adiabatic gradient we use discrete values $\Delta dT/dz_n$ = 0.10°C km^{-1}. By doing this, we also minimize the ambiguities that could arise at this stage from the trade-off between temperature and composition in seismic velocities. Under most circumstances, compositional variations in peridotites can result in *maximum Vs* variations of ∼ 1.5–1.7% [*Afonso and Schutt*, 2012; Afonso et al., Paper I, this issue], which is equivalent to maximum changes in temperature of ∼ 200°C [e.g., *Afonso et al.*, 2010]. Hence, by setting the minimum possible departure from the conductive geotherms to 200°C, we effectively remove local compositional effects from *this part* of the inversion (relaxed later). We emphasize that this is also entirely consistent with, and justified by, the intrinsic

theoretical uncertainties associated with the calculation of conductive geotherms (operator $\mathbf{g}_T$ in Paper I), which can be of the order of $\pm$ 80°C [*Joniken and Kukkonen*, 1999; *Vilá et al.*, 2010].

[23] In summary, this first part only aims to identify first-order temperature structures and large thermal anomalies from the conductive and adiabatic regimes. Given the strong sensitivities of the observables to variations in temperature, this first part of the search converges rapidly. As we show below, temperature anomalies are successfully identified and the initial range of 50 < LAB < 380 km is typically reduced to $\lesssim$ 40 km with the evaluation of only 3500–6000 random models per column.

3.2.2. 1-D Search With Emphasis on Composition

[24] This part is the most computationally demanding and time-consuming part of the method. Here we generate random samples of (i) LAB depths (within the new reduced range, Section 3.2.1), (ii) deviations ΔT_n for the lithospheric thermodynamic nodes in increments of 50°C, (iii) deviations $\Delta \mathrm{d}T/\mathrm{d}z_n$ in increments of 0.05°C km^{-1}, and (iv) five-oxide compositions for all compositional layers defined in the model. Any large temperature anomaly from conductive or adiabatic profiles detected in the first part is transferred to the second part.

[25] While the range of LAB depths typically varies only between $\pm$20–25 km in this part, the compositional space to sample is extremely large. Each thermodynamic node (intermediate discretization) has four quasi-independent variables (SiO_2 is not independent) when the system CFMAS is used to describe bulk compositions. Note that although the number of independent oxides (parameters) in the sampling method described in Paper I is only two (Al_2O_3 and FeO), the NA needs to consider a four-dimensional space (i.e., the regressions in Figures 3 and 4 in Paper I have a finite "thickness," represented by Gaussian distributions describing CaO and MgO as functions of Al_2O_3). Therefore, in this stage, the total number of parameters is $\sim$ 31–36 per column (4 oxides $\times$ layers + 1 parameter for LAB depth + ΔT_n + $\Delta \mathrm{d}T/\mathrm{d}z_n$).

[26] The sampling of the compositional space needs to be exhaustive in order to locate a representative ensemble of acceptable models. Importantly, not only the number of samples generated in each iteration N_s and the number of re-sampled best-fitting cells N_r need to be large, but more critically, the initial sample N_i must be large (e.g., Figures 3g–3i). In practice, we found that the combination N_i = 120,000–140,000, N_s = 5000–7000, N_r = 200–300, and *maxit* = 40–60 achieves satisfactory results. Therefore, the final number of models (for which a forward problem is solved) generated with the NA is of the order of 320,000–560,000 per column. With such large number of forward evaluations, it is critical that the forward problem (heat-transfer, potential field, isostasy, Love and Rayleigh waves, and Maxwell's equations) is thoroughly optimized. In our current 1-D implementation, the slowest part of the forward problem is the solution of the free energy minimization problem. We adopt the general strategy of solving the minimization problem ex tempore for any composition drawn during the inversion process in each thermodynamic node. By doing this, we do not need to limit the compositional space to a few pre-selected compositions, but rather the data "decides" which compositions need to be sampled. Each time one minimization problem is solved, we save the relevant information (i.e., seismic velocities, densities, etc) in pre-allocated CPU RAM memory vectors, for later use. In this way, if a similar sample in the parameter space (similar T, P and composition, see below) is drawn by the NA at later iterations, the minimization problem does not need to be solved again, instead we retrieve the information from the previously calculated sample. Importantly, as the simulations saturate the pre-allocated memory vectors, the algorithm automatically selects the least used values saved in the memory vector, and updates the register with the most recently computed value. Since by design, both MCMC simulations and the NA concentrate the sampling around specific regions of good fit as they proceed, the newly generated samples become gradually more similar to the previous ones, and thus the minimization problem is solved fewer times. This results in a rapidly increasing performance as the search/inversion proceeds. In the case of the NA, the actual gain in performance depends upon a slight modification in the way the algorithm generate new samples (Appendix B).

[27] It remains to be defined what we mean by "similar samples." We consider two samples to be similar (strictly, identical physical properties) if they are located within the same T-P-C cell in our discretized T-P-C space. The discretization parameters for this T-P-C space are chosen before starting the search. Unless otherwise indicated, we will use ΔT = 3°C, ΔP = 5 MPa, ΔFeO = 0.2 wt%, ΔAl_2O_3 = 0.2 wt%, ΔMgO = 0.5 wt%, and ΔCaO = 0.2 wt%.

3.2.3. A 1-D Synthetic Example

[28] A typical set of results from the full 1-D search stage described above is shown in Figures 4–6 for a single column of a 3-D synthetic model. The "observed" data used in the 1-D search/inversion included surface heat flow, geoid height, Love and Rayleigh dispersion curves (6 < period < 170 s), 1-D *Vp* structure, and absolute elevation. Each of these synthetic observables was obtained by adding Gaussian random errors to the predictions from the original 3-D model. The latter consisted of six compositional layers; three in the lithosphere and three in the sub-lithospheric upper mantle. The bulk compositions assigned to each of the compositional layer are listed in Table 1, together with their respective Mg# values. The misfit function used with the NA was defined as

$$\mathtt{misfit}(\mathbf{m}) = \frac{1}{2}\left[(g(\mathbf{m}) - \mathbf{d}_{\mathrm{obs}})^T C_D^{-1} (g(\mathbf{m}) - \mathbf{d}_{\mathrm{obs}})\right] \qquad (3)$$

The computation of the covariance matrix C_D used in this example is described in detail in Section 4.

[29] The ensemble depicted in Figures 4–6 has been selected from the entire populations of generated samples according to their fitting properties. Only those samples that fit the data within $\pm$ the variance of the data uncertainty (i.e., diagonals of C_D) are plotted and used to estimate PDFs with a Gaussian kernel density estimator [e.g., *Venables and Ripley*, 1999].

[30] There are a number of important observations to note in Figures 4–6. First, the retrieved PDF for LAB depths is an excellent representation of the true value, reflecting the fact that the observables are mostly sensitive to temperature structure. These results also reassure the benefits

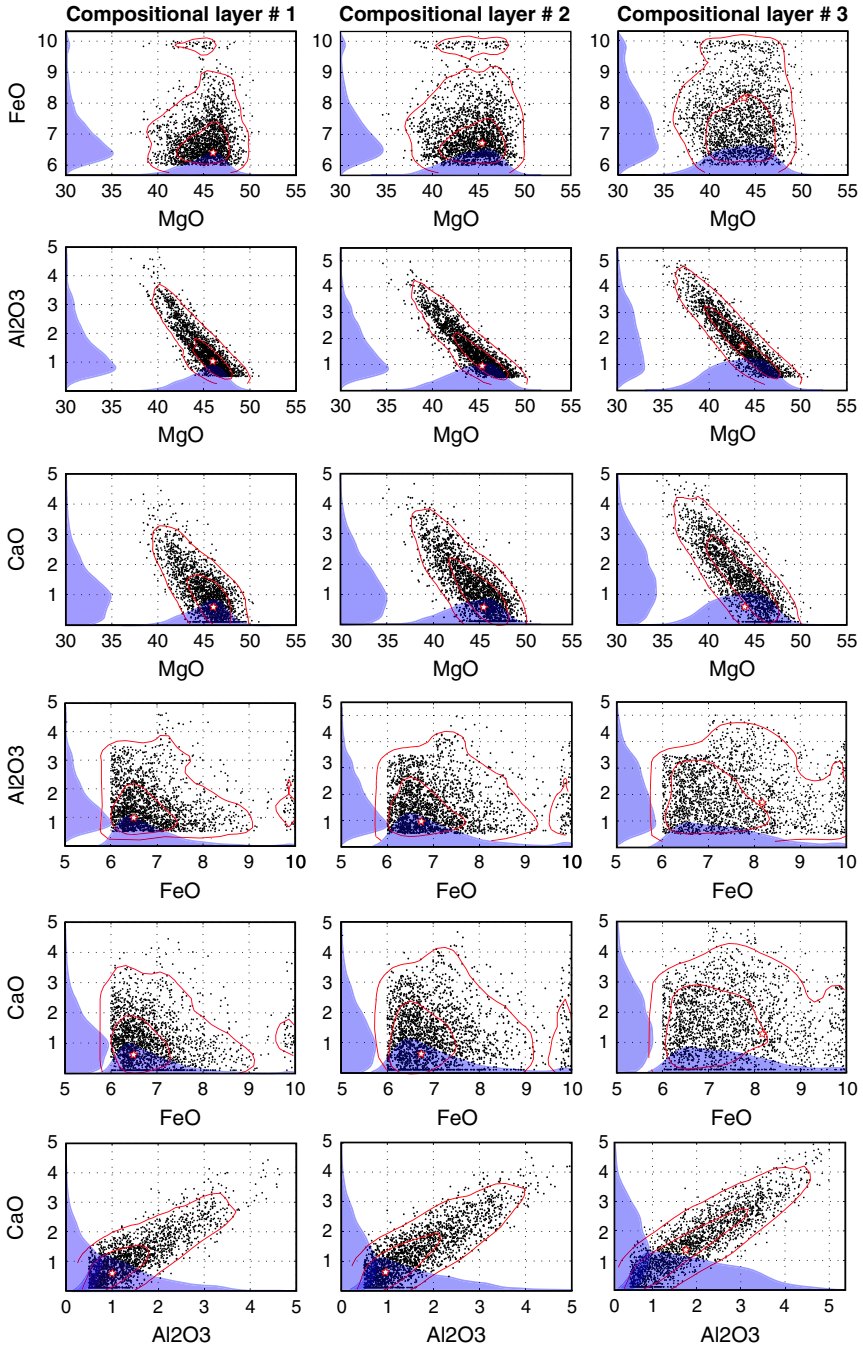

Figure 4. Ensemble of acceptable models from the 1-D column-by-column inversion described in the text. The synthetic data used in the inversion were geoid height, absolute elevation, Rayleigh dispersion curves, 1-D *Vp* anomalies, and surface heat flow. The PDFs to be used as data-generating priors in the 3-D refinement stage are also shown together with their 50% and 95% regions. A Gaussian kernel density estimator *Venables and Ripley* [1999] was used to compute the PDFs associated with the data points. The sharp cutoffs in the samples (e.g., Al_2O_3 and FeO) are due to the limits imposed during the compositional sampling, as described in Paper I (Afonso et al., this issue).

of dividing the 1-D inversion into temperature-oriented and a composition-oriented parts. Second, the PDFs for compositional layers 1, 2, and 6 can also be considered good representations of the true compositions; the PDFs exhibit relatively narrow peaks of high probability close to the true value. This suggests that the concentration of the sampling by the NA has been successful in producing natural PDFs for the second 3-D stage of the inversion. Although in most cases the true composition is located within the 50% contour interval of the posterior PDF, some results for layers 3, 4, and 5 are less clear. Here the resulting PDFs tend to be broader (i.e., the NA could not concentrate the sampling) and do not provide significant compositional information at this stage. Note that the poor localization by

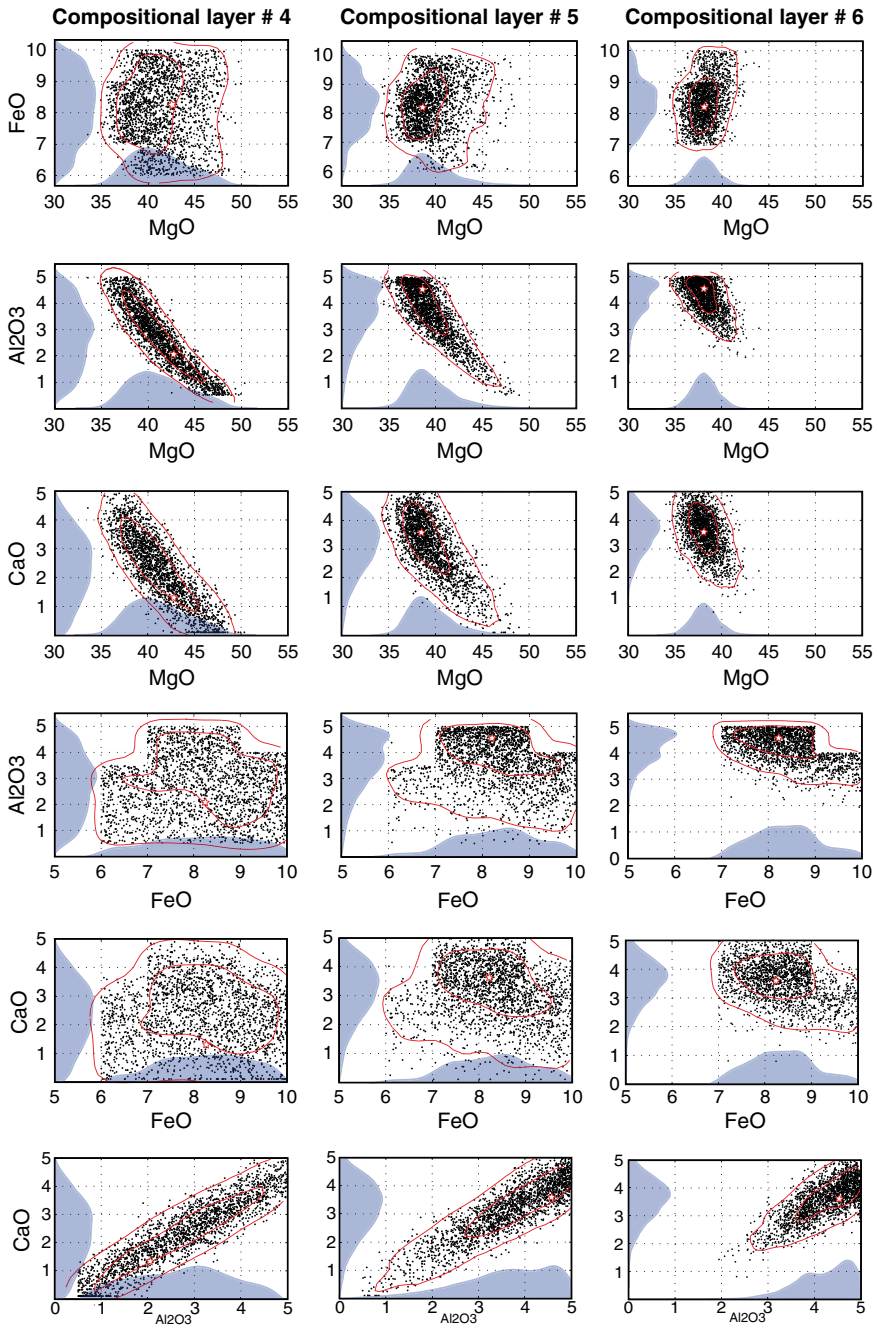

Figure 5. (continued).

the NA is not a problem of the search algorithm, but rather the unavoidable consequence of having acceptable samples spread over a large region of compositional space. In Paper I, we use synthetic rock samples to show that compositions close to the mean of the entire compositional space are difficult to isolate by Monte Carlo searches guided by residuals based only on the fitting properties of the sample. In these cases, the complex modal versus chemical effects that control the properties of the assemblages at lithospheric conditions result in ensembles of acceptable models large enough to cover most of the original variability observed in exhumed mantle samples (e.g., Figures 10 and 11 in Paper I). The situation is aggravated when vague priors and realistic observables + uncertainties are explicitly considered in the modeling and data vectors (as in the present example). Therefore, only those compositional anomalies that are significantly distant from the background composition and produce relatively large signatures in the observables are expected to be located (probabilistically) by the search. While in principle the refinement 3-D stage of the inversion could narrow the posterior PDFs of these layers, we do not expect major improvements. This is due to the simple fact that the observed data (plus all the uncertainties affecting the inversion) are not sensitive enough to significantly increase the value of $L(\mathbf{m})$ over the prior PDFs during the 3-D stage (section 5). In simple terms, we cannot extract more information from the observed data than it actually contains.

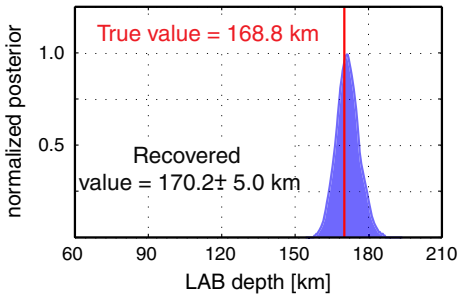

Figure 6. Posterior distribution obtained during the 1-D inversion/search for the LAB depth. This PDF is used as data-generating prior in the second 3-D refinement stage.

[31] Another relevant observation from Figures 4–5 is that the distribution in the deepest layer is clearly more localized than in the shallower layers. This observation is consistent with, and can be explained by, the pressure-dependent modal-compositional effect described in Paper I. Briefly, the combined effects of larger gt/opx ratios and the decrease in the capacity of pyroxenes to accommodate Al_2O_3 as pressure increases, result in a greater sensitivity of the secondary parameters *Vp*, *Vs*, and bulk density to changes in primary parameters such as bulk composition. This in turn translates into somewhat larger sensitivities of observables (particularly seismological observables) to changes in composition (especially Al_2O_3 content) at depths $\gtrsim$ 300 km, consistent with the results from the analysis in Paper I.

[32] An important result from this example is that even at this early stage of the inversion, we have been able to identify (on a statistical basis) vertical gradients in composition independent of the thermal structure. Furthermore, the recovered mean Mg# for each layer are close to the true values (Table 1). Although these results are encouraging, by no means should they be taken as evidence that the present method can distinguish between temperature and compositional effects at all scales. In fact, we emphasize that the results in Figures 4–5 clearly suggest that differences in bulk Mg# $\lesssim$ 3 are not expected to be resolvable with confidence by either the present or any other method based on geophysical observables. That the average Mg# of the acceptable ensembles are close to the true value simply reflects the tendency of the acceptable samples to cluster around the true solution; something described in detail in Paper I. However, the associated dispersion observed in the PDFs of Figures 4–5 (see also the standard deviations in Table 1) do not allow a statistically robust distinction between the Mg# of the three intermediate layers (3, 4, and 5). Simply put, with the current quality and quantity of observed geophysical data, we can only aim to simultaneously identify and/or distinguish, in a probabilistic manner, extreme compositional anomalies (e.g., highly depleted or highly fertile mantle domains) and relatively large temperature anomalies.

4. Theoretical and Observational Uncertainties

[33] Before moving on to the 3-D refinement stage, we need to discuss how to deal with both observational and theoretical uncertainties within the present inversion scheme. For most practical purposes, these uncertainties can be described with a Gaussian model, in which case we can write

$$\rho(\mathbf{d}) \propto \exp\left[-\frac{1}{2}(\mathbf{d}-\mathbf{d}_{\rm obs})^T C_d^{-1}(\mathbf{d}-\mathbf{d}_{\rm obs})\right] \quad (4)$$

$$\theta(\mathbf{d}|\mathbf{m}) \propto \exp\left\{-\frac{1}{2}[\mathbf{d}-g(\mathbf{m})]^T C_t^{-1}[\mathbf{d}-g(\mathbf{m})]\right\} \quad (5)$$

as PDFs for observational and theoretical data, respectively. The covariance matrices C_d and C_t incorporate the uncertainties (and their inter-correlations) in both types of data. When these expressions are used for $\rho(\mathbf{d})$ and $\theta(\mathbf{d}|\mathbf{m})$, it can be shown [*Tarantola and Valette*, 1982] that the likelihood function in equations 1 and 2 takes the form

$$L(\mathbf{m}) \propto \exp\left\{-\frac{1}{2}[g(\mathbf{m})-\mathbf{d}_{\rm obs}]^T C_D^{-1}[g(\mathbf{m})-\mathbf{d}_{\rm obs}]\right\} \quad (6)$$

where $C_D = C_d + C_t$. Sometimes, double exponential or Laplacian PDFs are preferred for modeling observational data since they are less affected by potential outliers [e.g., *Claerbout and Muir*, 1973; *Tarantola*, 2005]. This is particularly useful in cases where the uncertainties are uncorrelated. If correlations exit, the use of Laplacian PDFs requires much more cumbersome functions [e.g., *Eltoft et al.*, 2006; *van Gerven et al.*, 2009], and therefore they are much less popular in practical geophysical applications.

[34] There are no generally accepted rules for the computation of the covariance matrix in equation (6) for problems involving multiple datasets. In particular, C_d is commonly represented as a diagonal matrix (no correlations), while C_t is typically neglected. In our case, however, uncertainties associated with some of the observables are correlated (e.g., gravity and geoid anomalies, Rayleigh and Love dispersion curves), and therefore a full-rank C_D would be needed. One of the main difficulties is the large dimension of the data vector $\mathbf{d}_{\rm obs}$ (typically, the number of elements in $\mathbf{d}_{\rm obs} > 8208$), which makes the associated full-rank matrix $C_D = C_t + C_d$ simply too big to be practical in MCMC simulations. Fortunately, C_D is typically sparse, and for most practical purposes, it can be subdivided into smaller submatrices (blocks) for an easier computation of equations 6 and 3 (see below).

[35] The diagonal elements of C_d can readily be obtained from independent analyses of the individual data (see Paper I for a full discussion on observational uncertainties). For dispersion curves and/or gravity and geoid anomalies, some

Table 1. Bulk Compositions Used in the Synthetic Model of Section 3.2.3. The Original and Recovered Mg# of Each Layer Is Also Included

Layer #	SiO_2	Al_2O_3	FeO	MgO	CaO	Mg#	Mg_R#[a]
1	46.08	1.00	6.45	45.88	0.59	92.7	92 ± 0.9
2	46.34	0.98	6.75	45.34	0.59	92.3	91.7 ± 1.1
3	44.88	1.76	8.21	43.87	1.29	90.5	91.1 ± 1.2
4	45.84	2.06	8.24	42.55	1.31	90.2	90.0 ± 1.4
5	45.45	4.55	8.18	38.18	3.64	89.3	89.4 ± 1.1
6	45.45	4.55	8.18	38.18	3.64	89.3	89.0 ± 0.8

[a] Mean ± standard deviation of the recovered values.

of the non-diagonal elements can also be estimated during the processing of the raw data [e.g., *Bensen et al.*, 2007; *Lin et al.*, 2009]. However, the potential correlations between all the different observables considered here is difficult (if not impossible) to estimate with confidence in real case scenarios.

[36] The evaluation of C_t is arguably more important. It provides information not only on the reliability of our forward problem but also can inform us about unforeseen sources of error in our numerical schemes. Furthermore, uncertainties in the numerical solution of the forward problem are typically of the same order as, if not larger than, those affecting observations. A proper estimation and incorporation of theoretical uncertainties is therefore essential. Here, we follow a procedure similar to that in *Gouveia and Scales* [1998] to compute estimates of C_t. Since, in principle, C_t is independent of the vector of model parameters **m**, a single representative covariance matrix C_t could be computed prior to the inversion and then used throughout the subsequent analysis.

[37] With the above considerations, and noting that the approach to compute of C_t and C_d is independent of the general inversion formalism (i.e., the Bayesian analysis is valid with or without an explicit definition of these matrices), in what follows we describe a practical, yet representative, approach for the estimation of C_t and C_d and their incorporation into the inversion scheme (see also *Gouveia and Scales* [1998] and *Sambridge* [1999b]).

4.1. A Practical Approach for Computing C_D

[38] Besides the intrinsic uncertainties associated with the models of anelastic attenuation (Appendix A4), the main source of theoretical uncertainty in our method is related to the three discretization scales used to numerically solve the forward problem (Figure 2 and section 3.1). The purpose of estimating the covariance matrix C_t is therefore to incorporate representative calculation errors (and their correlations) associated with the artificial discretization of a real thermochemical continuum. We estimate the elements of C_t as follows:

[39] 1. Create a reference 3-D thermochemical model in which all three (vertical) discretization scales have a node spacing of = 1 km (ultrahigh resolution). The horizontal node spacing in each direction should be at least one half the spacing in the model used during the inversion.

[40] 2. Solve the 3-D forward problem of this ultrahigh resolution model and compute all the observables.

[41] 3. Rediscretize the 3-D model using the discretization scales described in the text. The new vertical node spacing is 2 km for the computation nodes, ~20–30 km for the thermodynamic nodes and ~40–50 km for the compositional layers. The node spacing of the last two scales are average values because they vary depending on the actual LAB depth (see text).

[42] 4. Compute the observables associated with this rediscretized 3-D model and calculate the difference ΔO_i between its observables and those computed in item 2.

[43] 5. Repeat 1–4 as many times as necessary, always starting with a different high-resolution model.

[44] 6. Use the results of all simulations to compute the elements of the diagonal (variances) of C_t as $\sum(\Delta O_i)^2/N$, where N is the number of repetitions in 5.

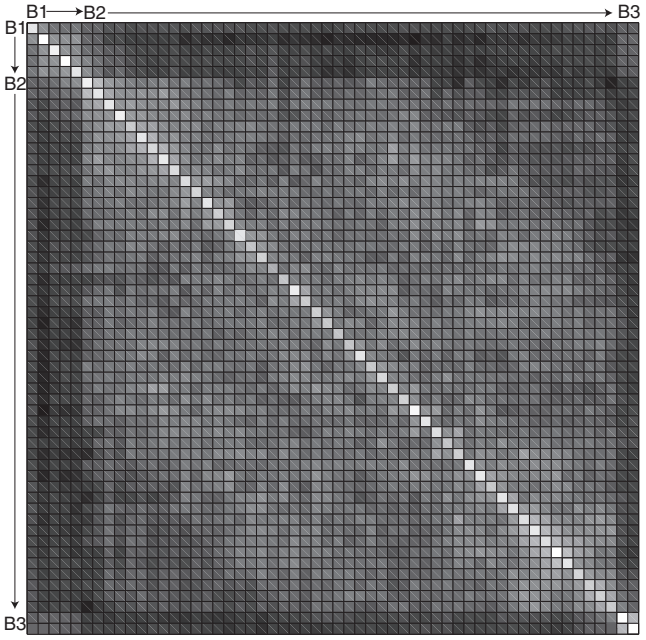

Figure 7. Subset (one column only) from the total covariance matrix used in this work to describe theoretical uncertainties. Gray scale indicates strong (white) and weak (black) correlations. Block B1 includes surface heat flow, Bouguer and free-air anomalies, geoid height, and elevation; Block B2 includes all the seismic observables (e.g., dispersion curves for individual periods, *Vp* for individual nodes, etc.); Block B3 includes the two MT observables, namely, apparent resistivity and phase. When Bouguer and free-air anomalies, geoid height, and elevation are removed from this matrix (i.e., those exhibiting correlations within and between columns), the resultant reduced matrix is a real example of C_{D2} (see text).

[45] In practice, since the solution of the high-resolution model is computationally demanding, we only used eight iterations in this paper. Given the independence of C_t on the vector of model parameters **m**, further iterations can be performed at any time and their results used to update the global C_t.

[46] The non-diagonals elements need to capture the extent to which random fluctuations in the calculated data vector $g(\mathbf{m})$ are correlated. We obtain estimates of these correlations by modifying one column at a time and computing the effect on the observables of all other columns in the 3-D model (note that for this to be useful, one need to use the same discretization as in subsequent real inversions). Given the nature of the chosen observables and the dimensions of each column (i.e., horizontal resolution), it turns out that the only observables that exhibit non-negligible correlations, both within columns and between columns, are the geoid height, gravity anomalies, and elevation. All other observables display significant correlations only within individual columns (Figure 7). We therefore assemble C_t in the following order: gravity anomalies for all columns, followed by geoid heights for all columns, followed by elevation for all columns, followed by all other observables ordered by column (e.g., surface heat flow of column 1 + seismic observables of column 1, followed by surface

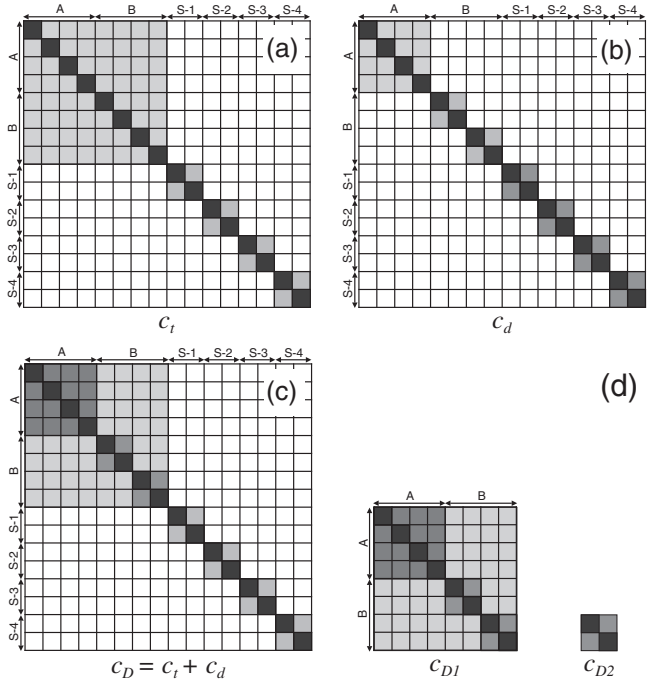

Figure 8. (a) Hypothetical matrix C_t for a 3-D model composed of only four rectangular columns and a data vector of length = 16 (i.e., observables A and B have one datum per column; observable S has two data per column). Theoretical uncertainties associated with observables A and B are correlated within and between columns, and therefore, they are assembled first for all columns (i.e., A1 A2 A3 A4 B1 B2 B3 B4). Theoretical uncertainties associated with observable S are correlated only within columns, and therefore, they are assembled per column after A and B (i.e., S11 S12 S21 S22 S31 S32 S41 S42). (b) Same as Figure 8a but for matrix C_d. Note that in this example, observational uncertainties for observables A and B are uncorrelated. (c) Total matrix C_D is simply the sum of C_t and C_d (for Gaussian uncertainties). If both theoretical and observational uncertainties for observable S are only correlated within columns, C_D can be subdivided into two master submatrices C_{D1} and C_{D2} for a simpler computation of the likelihood function (Figure 7 for an augmented example of C_{D2}).

heat flow of column 2 + seismic observables of column 2, etc.). This ordering results in a sparse matrix C_t (Figure 8a) with a structure that significantly simplifies the computation of likelihood functions (see below). Finally, we scale all non-diagonal elements to be consistent with the diagonal elements computed previously (steps (i)–(vi) above).

[47] The final covariance matrix C_D in equation (6) is obtained by simply adding C_t to the covariance matrix C_d. Ideally, C_d would be obtained from a previous analysis of the individual observables, and in general, it will not exhibit the same structure as C_t. For instance, while uncertainties in measured geoid and gravity anomalies are expected to be correlated, observational uncertainties in elevation may be independent of those associated with gravity measurements (Figure 8b). For the synthetic cases used in this paper, we use the synthetic data as observed data and compute C_d as follows:

[48] 1. Solve the 3-D forward problem of the synthetic model and compute all the observables.

[49] 2. Add moderate Gaussian noise to the model parameters and compute the new set of observables.

[50] 3. Generate as many realizations N_d as necessary by repeating 1–2. This generates a matrix with dimensions $N_d \times \mathbf{d}_{\text{obs}}$. Each row vector represents a new realization or measurement of the full data vector.

[51] 4. Compute the unbiased covariance matrix $C_d = \frac{1}{N_d-1}\sum_{i=1}^{N_d}(\mathbf{d}_{\text{obs}} - \mathbf{d}_{\text{mean}})(\mathbf{d}_{\text{obs}} - \mathbf{d}_{\text{mean}})^T$, where $\mathbf{d}_{\text{mean}}$ are the elements of the mean vector from the sample N_d.

[52] For this work, we have only generated 50 realizations ($N_d = 50$). Finally, we ignore the correlations between elevation and gravity/geoid, since they are an artifact of the actual numerical method that is not representative of real observational uncertainties (except in cases such as satellite-derived bathymetry/geoid). Furthermore, to be consistent with the computation of C_t, we only allow correlations within columns for the rest of the observables (Figure 8b). This is a suitable assumption, since the computed correlations between columns are small. To minimize the artificial effect of the arbitrary initial random noise assigned to the model parameters and obtain a C_d representative of real observational noise, one should normalize the elements of C_d by forcing the diagonal values to coincide with realistic uncertainty values (σ^2) for each variable. Representative uncertainties for each observable are discussed in detail in Paper I.

[53] The final structure of C_D allows us to subdivide the computation of $L(\mathbf{m})$ in two steps that involve only two matrices of modest size (Figures 8c and 8d):

$$L_1(\mathbf{m}) \propto \exp\left[-\frac{1}{2}(g_1(\mathbf{m}) - \mathbf{d}_1)^T C_{D1}^{-1}(g_1(\mathbf{m}) - \mathbf{d}_1)\right] \quad (7)$$

$$L_2(\mathbf{m}) \propto \exp\sum_{1}^{N}\left[-\frac{1}{2}(g_i(\mathbf{m}) - \mathbf{d}_i)^T C_{D2}^{-1}(g_i(\mathbf{m}) - \mathbf{d}_i)\right] \quad (8)$$

where $\mathbf{d}_1$ is a subset of the total data vector containing the data that shows correlations within and between columns, $\mathbf{d}_i$ are subsets of the total data vector containing the data that shows correlations within columns only, $g_1(\mathbf{m})$ and $g_i(\mathbf{m})$ are the appropriate forward operators for $\mathbf{d}_1$ and $\mathbf{d}_i$, respectively, N the number of columns for which there is a subset $g_i(\mathbf{m})$, C_{D1} the (sub)covariance matrix describing the uncertainties associated with the subset $\mathbf{d}_1$, and C_{D2} the (sub)covariance matrix describing the uncertainties associated with the subsets $\mathbf{d}_i$. The total likelihood is simply $L(\mathbf{m}) = L1(\mathbf{m})\, L2(\mathbf{m})$.

5. 3-D Refinement Stage

[54] The 1-D nonlinear inversion process described in section 3 greatly reduces the large initial range of model parameters. In particular, the original range of LAB depths is usually reduced as much as 95% and its resulting posterior is thus very narrow (Figure 6). For the compositional space, on the other hand, the posterior PDFs are wider, but in most cases, they exhibit defined regions of high probability (Figures 4–5). A preliminary pseudo 3-D model can

be constructed simply by assembling the best-fitting model columns obtained in the 1-D inversion part. This first pseudo 3-D model is used as the starting model for the refinement 3-D part (e.g., starting point in a MCMC simulation). During this second refinement part, the physical or forward problem (e.g., potential field equations) of each new drawn model is solved in full 3-D geometry, except for Maxwell's equations, which are approximated as 1-D due to the time taken by the 3-D forward solver (see section C for details).

[55] It transpires that, if the 1-D inversion was exhaustive, the initial pseudo 3-D model already fits most of the 3-D data well. In principle, this allows the modeler to choose among a number of approaches for the refinement stage; foremost amongst these are: (i) a purely forward approach, (ii) a linearized inversion method, or (iii) a Bayesian analysis. In this paper, we will deal only with the Bayesian approach, which we consider the most general case. Although an exhaustive comparison of the virtues and limitation of these methods is beyond the scope of this paper, we note that a linearized inversion represents a particularly attractive option, since it would significantly reduce the computation time. However, as with any other linear inversion method that includes the modeling of multiple observables, the success of the linearized approach relies upon an optimal regularization of the model, which is a difficult problem itself [cf. *Zhdanov*, 2002]. Moreover, preliminary tests with synthetic models indicate that if the compositions of particular columns or layers in the initial pseudo 3-D model are relatively far from their true value (something that can be considered the rule rather than the exception), a linearized technique is not able to retrieve satisfactory estimates of the true composition. Despite these preliminary results, and considering the significant gains in computational efficiency, a thorough future assessment of the practicality and reliability of linearized methods for the current problem is warranted.

5.1. Sampling the A Priori Information

[56] The a priori information used in the 3-D refinement stage is obtained during the 1-D column-by-column inversion described in section 3. It consist of (i) first-order PDFs describing the acceptable distributions of major oxides for all the compositional layers making up the 3-D model, (ii) a PDF describing the LAB depth, and (iii) discrete ΔT and $\Delta \mathrm{d}T/\mathrm{d}z_n$ (departures from a conductive geotherm and nominal adiabatic profile, respectively) values associated with the thermodynamic nodes of the intermediate discretization. Since there are no analytical expressions for these PDFs, we use a Metropolis-Hastings algorithm to draw samples from these a priori probability densities [cf. *Mosegaard and Tarantola*, 2002; *Tarantola*, 2005].

[57] As described in *Mosegaard and Tarantola* [1995], it is important that the steps in the sampling of the prior do not lead to radical changes in the predicted observables. Otherwise, the probability of being accepted as samples of the a posteriori density (proportional to $L(\mathbf{m})$) becomes extremely low and the performance of the algorithm is seriously deteriorated. In our case, a simple way of sampling the prior without compromising the acceptance rate of the models proceeds as follows:

[58] 1. Randomly choose one column of the 3-D model; guarantee that the probability of being chosen is the same for all columns.

[59] 2. Use a Metropolis-Hastings algorithm to obtain a new sample of the column (i.e., different LAB depth and layers' compositions) according to its prior PDFs.

[60] The samples or models so generated correspond exactly to the a priori information we want to input into our problem. At the same time, this random walk does not perturb previous good samples to a large extent, thus favoring a greater acceptance rate by the Metropolis rule (see below).

5.2. Sampling the A Posteriori Probability Density

[61] The first sample in our MCMC simulation is the initial pseudo 3-D model. Once a new sample $\mathbf{m}_{\mathrm{new}}$ of the prior has been chosen as described above, the 3-D forward problem is solved and its likelihood function $L(\mathbf{m}_{\mathrm{new}})$ is obtained. We then decide whether $\mathbf{m}_{\mathrm{new}}$ is accepted as part of the chain or rejected according to the *Metropolis rule* [cf. *Mosegaard and Tarantola*, 1995; *Tarantola*, 2005]:

[62] 1. if $L(\mathbf{m}_{\mathrm{new}}) \geq L(\mathbf{m}_{\mathrm{prev}})$, then accept $\mathbf{m}_{\mathrm{new}}$

[63] 2. if $L(\mathbf{m}_{\mathrm{new}}) < L(\mathbf{m}_{\mathrm{prev}})$, then randomly decide to accept or reject $\mathbf{m}_{\mathrm{new}}$ with probability $\mathrm{P} = L(\mathbf{m}_{\mathrm{new}})/L(\mathbf{m}_{\mathrm{prev}})$ of accepting $\mathbf{m}_{\mathrm{new}}$

where $\mathbf{m}_{\mathrm{prev}}$ is the last accepted model (pseudo 3-D model if we just started the simulation). The final ensemble of accepted samples so obtained constitutes an estimate of the posterior $\sigma(\mathbf{m})$. It is customary in MCMC simulations to define a "burn-in" period in which the accepted samples are not included in chain [cf. *Gilks et al.*, 1996]. This has the purpose of minimizing or removing the "memory effect" of the original or starting sample, which in principle, could be far from the high a posteriori probability region. Since in our case the starting sample typically is already a good sample, defining a burn-in period is not strictly necessary. Nevertheless, to allow for the possibility of a poor starting sample, we usually adopt a burn-in period of 4500 samples.

5.3. A 3-D Synthetic Example

5.3.1. Synthetic Model and "Observed" Data

[64] In this section, we consider a basic example to illustrate the benefits and limitations of the probabilistic inversion method discussed above. The synthetic model is a 3-D Cartesian region that we subdivide into 144 (12 × 12) columns to be used in the 1-D stage. The total number of columns in a given model depends on both the total dimension of the region under study (typically $\lesssim 1.0 \times 10^6$ km^2) and the minimum surface area assigned for each column. The surface area of individual columns is, in turn, dictated by the assumptions used in the solution of the 1-D forward problem (Paper I and Appendix A), and for the present implementation, it should be $\sim$ 5000–10,000 km^2.

[65] The 3-D model includes four distinct compositional domains embedded at different depths within a slightly depleted "Tecton-type" background (Figure 9 and Table 2). The LAB and therefore the long-wavelength thermal structure of the model have an irregular geometry with localized lateral gradients (Figure 10a). The crustal structure is dominated by variations of shorter wavelength than those of the LAB (Figure 10b). Bulk density, *Vp*, and *Vs* within the crust are assumed to vary linearly with depth as $\rho(z) = 2670 + 9{*}z$

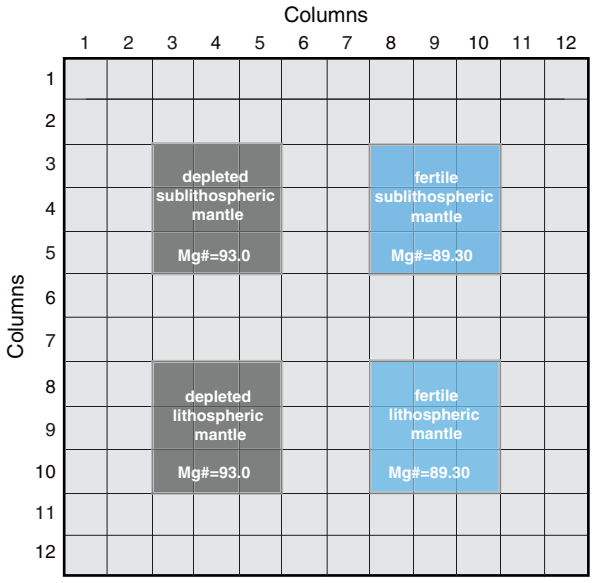

Figure 9. Columns making up our 3-D synthetic model. The surface projections of the four anomalous compositional domains together with their bulk Mg# are indicated. The bulk Mg# of the background is 89.9.

[kg m^{-3}], $Vs(z) = 3.45 + 4.25 \times 10^{-3} * z$ [km s^{-1}], $Vp(z) = 6.0 + 1.7 \times 10^{-2} * z$ [km s^{-1}], where z is in km.

[66] We generate the "observed" data for our test by first solving the 3-D forward problem associated with the 3-D model described above and then contaminating its predictions by adding Gaussian random noise. We do this on a finer grid (120 × 120) in order to simulate real data more closely and to be able to compute "observational" variances for the inversion (see below). Figure 11 shows three of these synthetic observables. The first-order general patterns of the geoid height, Bouguer anomaly and elevation are easily understood in terms of the assumed thermochemical structure of the model. For instance, the two major features in the elevation, namely, the high and low domains in the left side of the model, are the response to the highly depleted (buoyant) lithospheric domain in the lower left corner (Figures 9 and 12) and the denser (colder) lithospheric root in the upper left corner (Figure 10), respectively. The same two structures are also identifiable in the geoid height (Figure 11). As expected, the pattern of Bouguer anomalies shows a closer correlation with the crustal structure than with the long-wavelength structure of the thermal lithosphere. The only clear perturbation to this pattern is the large low anomaly region associated with the topographic high. Note that there is no crustal compensation in this case; the topography is compensated by the low density depleted body in the lithospheric mantle. Yet, since the density anomaly is relatively shallow and large, it influences considerably the pattern of gravity anomalies.

[67] Before starting the 1-D part of the inversion, we need to assign a unique observed value for each column plus uncertainties. For gravity anomalies, geoid height, surface heat flow, and topography, we first select the area associated with a particular column and subdivide it into 100 subregions (i.e., 10 by 10). We then take an individual value in each subregion from the observed data (Figure 11) and compute the average and variance of all 100 values. The averages so calculated represent the input data for each column, while the variances are added to the diagonals of C_d (section 4). This is possible because these variances are related exclusively to the averaging process (i.e., averaging the natural variability of the data over the area of a column), and therefore they represent an independent source of uncertainty.

[68] In all cases, synthetic seismic observables (and the methods to compute them) include corrections for high-temperature anelastic behavior as described in section A4. For the dispersion curves, we use the prediction from the true model and assume the following observational uncertainties (i.e., diagonals of C_d): 15 m s^{-1} for 6 s < periods < 30 s, 20 m s^{-1} for 30 s < periods < 60 s, and 25 m s^{-1} for periods > 60 s. For the *Vp* structure, we first subtract the *AK135* model [*Kennett et al.*, 1995] from each computation node to obtain the respective *Vp* anomalies (which are the only reliable information from body-wave tomography studies; see Paper I). We then add Gaussian noise with a standard deviation of 0.5% to these anomalies and used them as data in the inversion. This simulates maximum anomaly uncertainties of ∼ 1.2%. Larger uncertainties would render the use of *Vp* anomalies unviable for the present purposes/method. We emphasize, however, that the present choice of working with synthetic *Vp* anomalies is only for illustrative purposes. A more objective approach of including body-wave data into the inversion is discussed in section 7.

5.3.2. Results From the Inversion

[69] The final amount of data generated by the inversion is extremely large, and typical visualization techniques such as marginalization of the posterior is impractical unless we are interested in a small number of parameters. Since the solution to the problem is not a single model, but a probability density over the model parameters (in practice, an extremely large collection of acceptable models), the question of how to visualize this posterior and evaluate the resolving power of the method for 3-D models has no trivial answer. For instance, how do we visualize the marginal posterior for the LAB depth? Ideally, we would need to construct marginals of the LAB depth for different points on the surface of the model (e.g., one for each column), since each 2-D surface region has associated a probability density describing the depth to the LAB under that region. For illustration purposes, here we choose to describe the solution of our inversion by plotting the means of randomly chosen subsets from the entire collection defining $\sigma(\mathbf{m})$ (cluster analysis is another option). For each subset, we randomly choose 500 models from the entire posterior ensemble and compute the mean vector and covariance matrix of the subset. Figure 13 shows eight examples of such mean maps for the LAB, together with the original model. Since the method

Table 2. Bulk Compositions Used in the 3-D Synthetic Model of Section 5

Domain	SiO_2	Al_2O_3	FeO	MgO	CaO	Mg#
Depleted domains	44.70	0.50	6.50	48.20	0.10	93.0
Fertile domains	45.66	4.50	8.10	37.83	3.91	89.3
Background	45.60	3.50	8.00	39.80	3.10	89.9

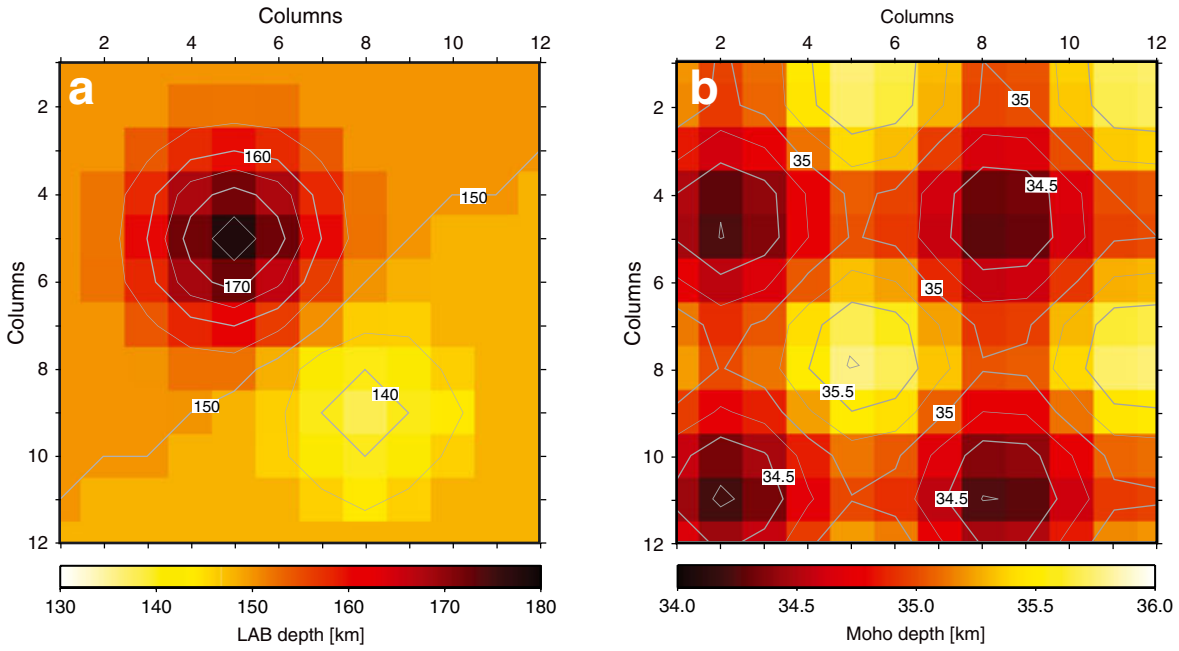

Figure 10. Contour maps of the (a) LAB and (b) Moho used in the 3-D synthetic model of Section 5. Note the different wavelengths of these two structures.

is particularly designed to generate model features with a frequency proportional to their a posteriori PDF, those features that appear more frequently and consistently in the subsets (and also in the total ensemble of acceptable models) are considered well resolved. Remember that our inversion scheme starts by assuming very vague priors for the thermal and compositional structure of the model, and therefore the features that appear with high frequency in the final collection are favored mainly by the likelihood function (i.e., data). From the maps in Figure 13, we can conclude that the inversion has been able to retrieve satisfactory solutions for both the two original anomalies (persistent in all maps) and the mean LAB depth.

[70] Since the LAB posterior has a simple form and is well behaved (no multiple peaks, etc.), plotting the mean of all possible solutions together with some estimation of the variance is informative. Constructing 2-D transects across the model showing the probabilities of having the LAB at specific depths (Figure 14) is therefore an appropriate option. The same is not true for the visualization of the compositional parameters.

[71] Figures 15a–15h and 16a–16h show the results for eight mean subsets for bulk Mg# and Al_2O_3 content, together with the initial true model. As in Figure 13, each subset has been constructed by randomly choosing 500 models from the entire posterior ensemble and computing the mean of the subset. The most salient features in Figures 15 and 16 are the two highly depleted domains. Importantly, both the lithospheric and sub-lithospheric depleted domains are well resolved with relatively high spatial resolution. The resolving power for the fertile domains, on the other hand, is less obvious, particularly in the case of Mg# (Figure 15). Most of the subsets in Figure 15 correctly depict the presence of an anomalously fertile sub-lithospheric domain at the northeastern lower corner of the model, although the true shape is not well recovered. Similarly, a fertile shallow region at the southeastern corner is a consistent, yet dimmer, feature across all subsets.

[72] A better picture of the anomalously fertile domains is obtained when looking at the bulk Al_2O_3 contents (Figure 16), a well-known indicator of depleted versus fertile peridotites. Although the depleted domains are also identifiable when looking at CaO and FeO, the same is not true for the fertile domains. Only Al_2O_3 offers a relatively clear picture of the fertile domains. This is no surprise, as bulk Al_2O_3 has the strongest effect on the physical properties of peridotitc assembleges (see also Paper I). That the highly depleted domains have been recovered with higher probability/resolution than the fertile domains simply reflects the fact that the contrast between the physical properties of the background and depleted assemblages is stronger than that between the background and the fertile domains. In particular, the irregular and patchy pattern recovered for the fertile domains is consistent with our previous observation (section 3.2.3) that differences in bulk Mg# < 3 are not expected to be resolvable with confidence by methods based on geophysical observables. It seems therefore that bulk Al_2O_3 rather than Mg# could be a better parameter to discriminate between compositional end members in thermodynamically constrained studies.

[73] Perhaps more important is the fact that the posterior probabilities of both horizontal and vertical gradients in composition are good representations of the true model (also in the 1-D example of section 3.2.3). In other words, the important step forward is not the recovery of accurate absolute amplitudes, but the robust discrimination/identification of both lateral and vertical thermochemical changes within the lithosphere and sub-lithospheric upper mantle. To this respect, what makes our method particularly well suited for discriminating thermal and compositional anomalies is the internally consistent combination of observables that are differentially sensitive to bulk density and seismic

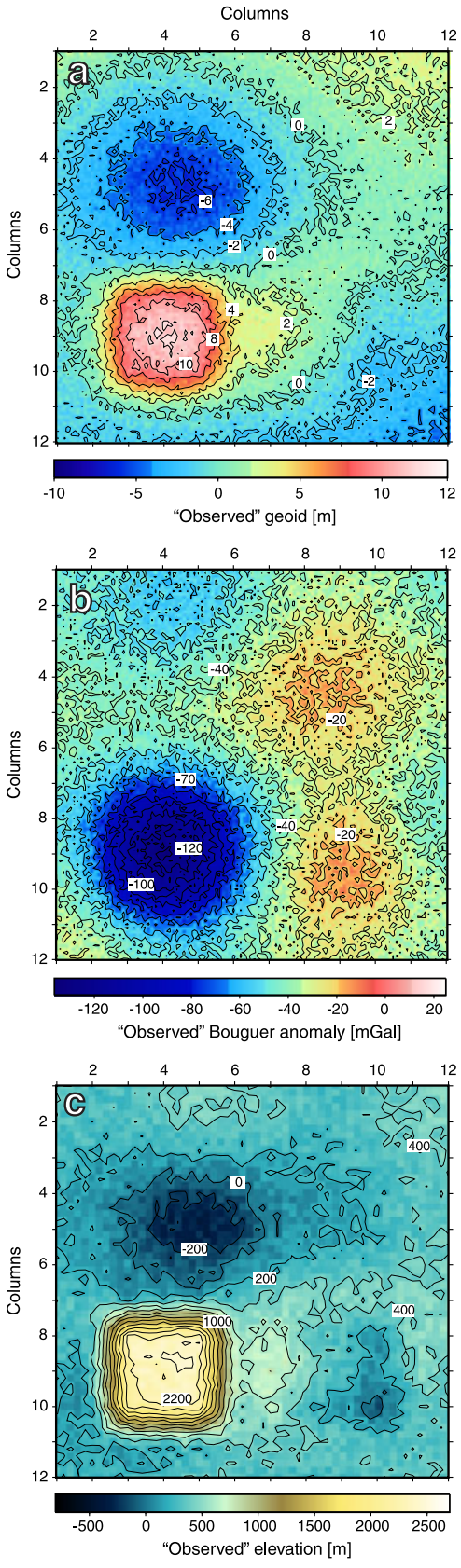

Figure 11. (a) "Observed" geoid, (b) Bouguer anomaly, and (c) topography datasets used in the 3-D inversion, generated by adding noise to the synthetic data predicted by the true model described in Section 5.3.1.

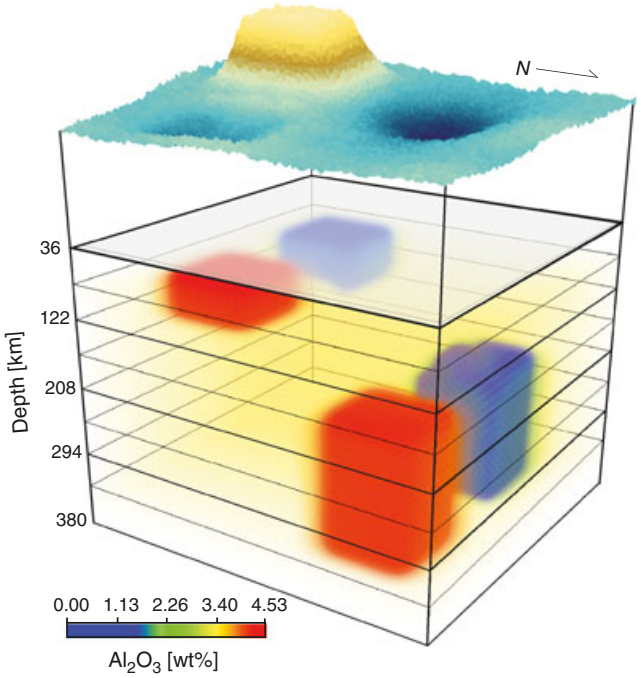

Figure 12. 3-D spatial distribution of the anomalous compositional domains in our synthetic model. The color scale denotes bulk Al_2O_3 content. Predicted topography plus noise (Figure 11) is also shown.

velocities. For instance, useful indicators such as ρ/Vs [*Forte and Perry*, 2000; *Deschamps et al.*, 2002; *Afonso et al.*, 2010] or topography/geoid ratios are intrinsically built in our method, and therefore, there is no need to assume ad hoc scaling factors or depth-dependent behaviors for them.

6. Further Comments on the Methodology

6.1. A Note on Computational Requirements

[74] As with any other high-dimensional problem solved with Monte Carlo techniques, the computational cost associated with the current method is high, and parallelization of the problem becomes essential. In principle, there are many different ways to parallelize our algorithms. In practice, however, only trial and error dictates which of them are best suited for the present problem. Despite the particular parallel implementation adopted, the number of columns in a typical 3-D model ranges between 144 and 400, and approximately, 180,000–300,000 models need to be evaluated in each column. This bring a total of forward evaluations to 25–120 million in the 1-D stage only; another 550,000–850,000 models need to be solved during the 3-D stage. The need for large clusters (ideal number of CPUs = number of columns) and efficient parallel implementations is thus evident. Without attempting to perform an exhaustive analysis of all possible techniques, we will briefly describe a particular parallelization scheme that has proven adequate.

[75] Two parallel versions exist for the NA algorithm [*Rickwood and Sambridge*, 2006]. The first one, referred to as the canonical version, distributes the forward problem to different processors but collects and processes all their information in the master node only. This implementation

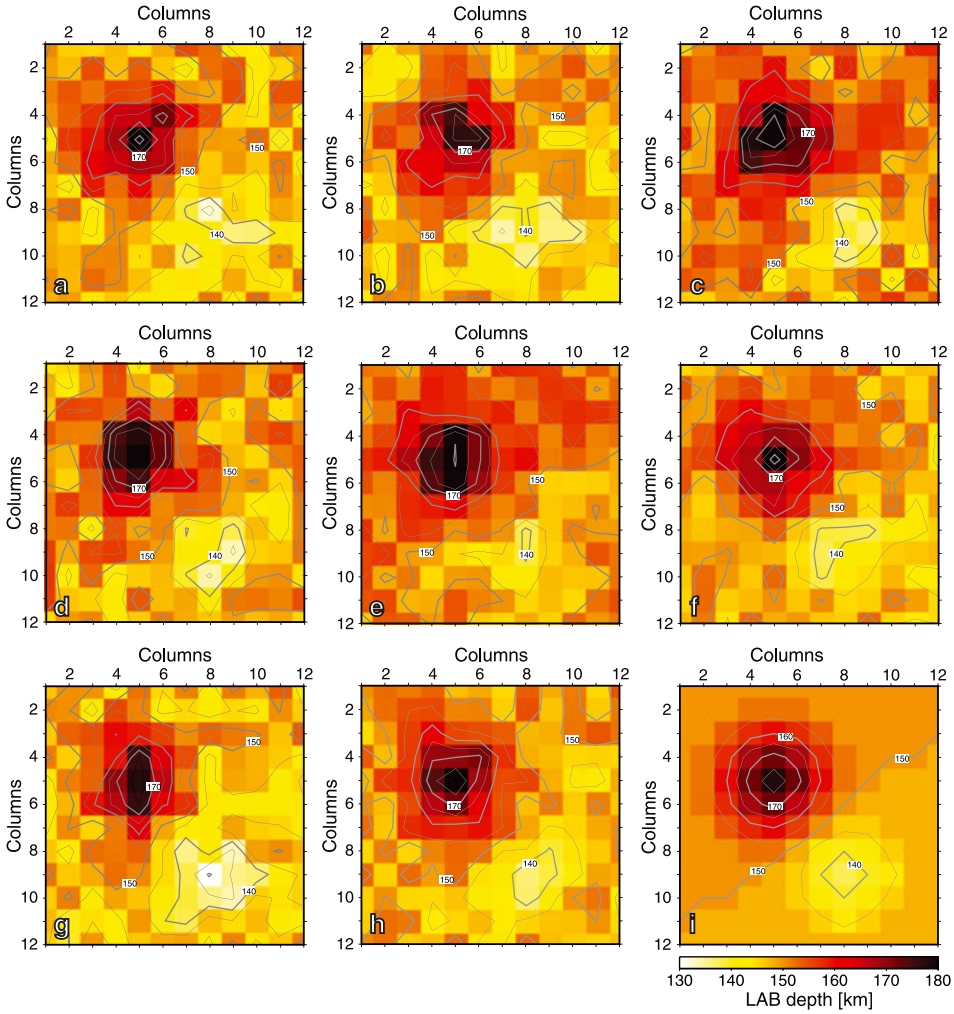

Figure 13. Mean LAB models of eight subsets of 500 random samples taken from the entire collection of acceptable models defining the posterior $\sigma(\mathbf{m})$. The true model is shown for comparison at the bottom right (i). Note that LAB anomalies with amplitudes close to the original values are persistent across all mean maps.

creates a bottle-neck at the master node level when the number of forward problems is as large as in our problem, since it is the master node that evaluates all the information from the computing nodes and generates the new samples for the next iteration. The second implementation attempts to overcome this limitation by truly parallelizing both the forward problem and the generation of samples. Unfortunately, when applied to our specific problem, neither of these implementations performed well as stand-alone. We therefore adopt the following strategy. Since the columns are considered independent in our current implementation of the 1-D inversion problem, we solve each 1-D problem (one column) as a serial problem in a single CPU. Each 1-D problem typically takes between 7 and 12 hs, depending on the total number of CPUs and memory available. Therefore, it is imperative to have access to at least $N_{cols}/2$ CPUs, where N_{cols} is the total number of columns in the 3-D model. The collection of all the information from each computing node is done by the master node at the end of the simulation. If more computing nodes were available, a better strategy could be to assign n nodes to each column and distribute the forward problem to these n nodes. This would require a total number of nodes of the order of number of columns $\times$ n, with $n > 2$. We have not tested this approach yet.

[76] For the 3-D stage of the inversion, we use a simpler and more popular parallelization strategy based on running different MCMC simulations in different CPUs [*Gelman and Rubin*, 1992; *Gelman*, 1996; *Wilkinson*, 2005]. This approach is particularly attractive when the burn-in period is short in comparison with the total number of samples, as it occurs in our problem. In this way, we can vary the properties of each chain and see whether they all provide consistent results. In particular, starting from different 3-D models taken from the PDFs obtained during the 1-D inversion stage can be useful, since it provides a measure of how dependent the solution is to the initial guess. Also, and perhaps more important, comparing different chains can inform us about the global convergence of the MCMC simulation [*Gelman*, 1996; *Rosenthal*, 2000; *Wilkinson*, 2005]. Given that this stage takes between 100 and 150 hs, future work

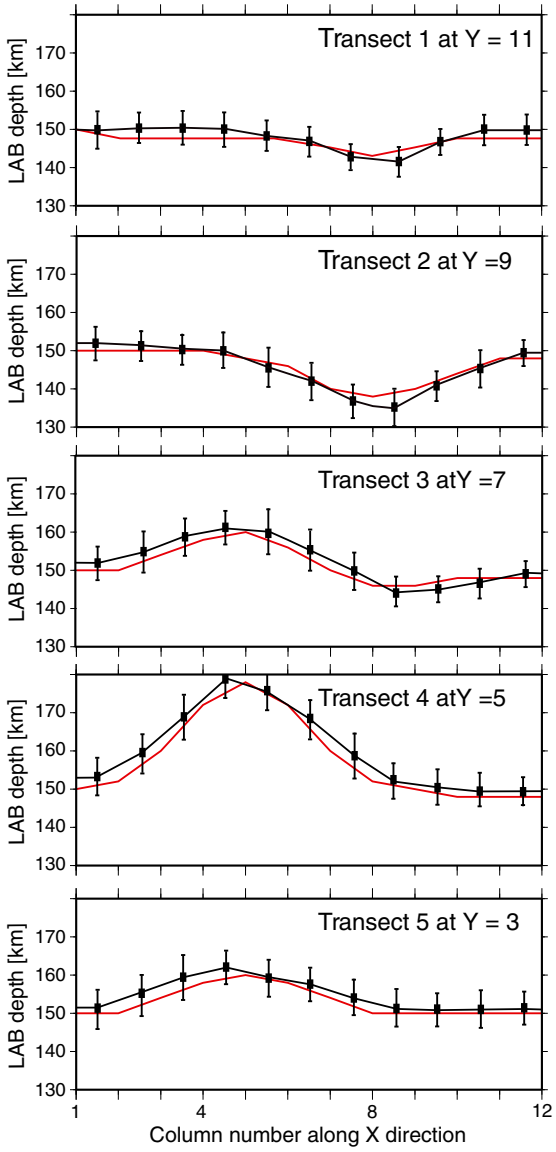

Figure 14. Mean values (black solid lines) for LAB depths and associated dispersion from the posterior ensemble of acceptable models. Bars denote one standard deviation from the mean value. Red solid lines denote the true values from Figure 10a.

on the parallelization of the actual 3-D forward operators and the independent chains will be highly beneficial (see Section 7.2).

6.2. Including Crustal Structure in the Inversion

[77] So far we have assumed that the density, resistivity and velocity structures of the crust is known from previous independent studies (reflection-refraction profiling, MT studies, ambient-noise tomography, etc.). While this may be true in some well-studied regions, it is likely that we will not have a thorough coverage of the crustal structure in all regions of interest. Therefore, since some of the geophysical observables used in our method are significantly sensitive to crustal structure (particularly Moho depth and average density), its signature has to be included in the inversion process (i.e., the crustal structure becomes an unknown).

[78] When this is the case, we propose to discretize the crust as either three or four consecutive layers and define the following additional variables in the 1-D temperature-oriented search (section 3.2.1): bulk density ρ, Vs, Vp (actually, Vp/Vs ratio), electrical conductivity, and layer thickness for each crustal layer. Note that the thermodynamic equilibrium problem is not solved within the crust, and therefore, we do not explicitly solve for its compositional structure. This is justified by the lack of reliable quantitative information on reaction kinetics and metastability at temperatures $< 500°C$. However, this is a restriction that can be easily relaxed if reliable data is available.

[79] The reason for inverting for the crustal structure during the 1-D temperature-oriented part is the strong sensitivity of the observables (particularly surface waves and isostasy) to both temperature and crustal structure. Therefore, similarly as for temperature, a large reduction of the initial parameter space for crustal parameters is achieved with a modest number of forward evaluations (25,000–60,000 for the combined temperature-crust case), and the resulting PDFs are typically narrow and well behaved. We therefore consider the best-fitting crustal model as known data during the second 1-D search with emphasis on composition (i.e., crustal structure is not allowed to vary). For the 3-D refinement stage, however, the full PDFs of crustal parameters are used to generate samples of crustal models.

[80] We acknowledge that the above discretization and approach to include the crustal structure in the inversion are not the only possible options, but they represent a sensible compromise between performance and robustness. The addition of a receiver function operator to the forward problem (work in progress, see next section) when inverting for crustal structure is of particular relevance. This could provide valuable additional constraints to determine the Moho structure and the location of the interfaces between the internal layers, provided reliable receiver function data exist in the region of interest.

7. Future Work

7.1. Additional Operators

[81] One of the main advantages of probabilistic methods such as the one described in this paper is the possibility of working with different forward operators in an internally consistent way. Likewise, the modular character of our implementation allows for including or excluding operators from the inversion according to available data. Since the main purpose of inverting multiple geophysical observables is to minimize the range of plausible or acceptable models consistent with data, it is a requisite that each forward operator contributes a different (an ideally independent) sensitivity to the model parameters.

[82] Two operators that in principle could provide valuable additional information, and that we have not formally included in this work, are the seismic-wave propagation and receiver function (RF) operators. The inversion of P or S teleseismic arrival time residuals for 3-D velocity variations beneath an array can be performed with a forward calculator such as the Fast Marching Method (FMM) to solve the forward problem of traveltime prediction [*Rawlinson and Sambridge*, 2004; *Rawlinson et al.*, 2006].

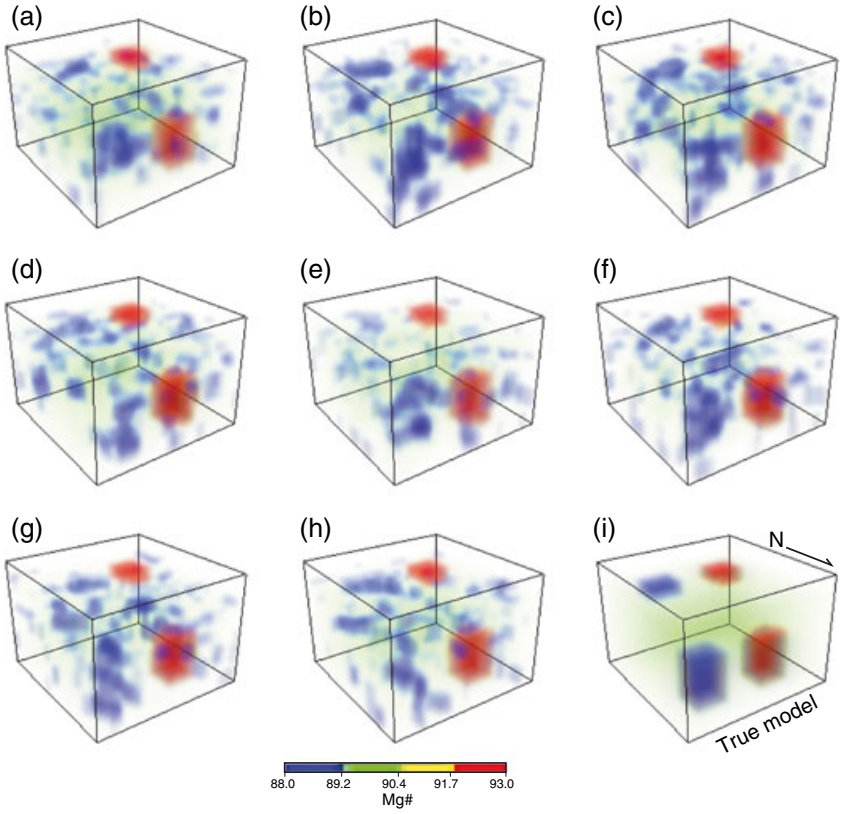

Figure 15. Mean-models of eight subsets of 500 random samples taken from the complete collection of acceptable models defining the posterior PDF. The color scale denotes the bulk Mg# of the model. The true model (i) is shown for comparison.

The FMM uses upwind entropy satisfying finite-difference operators combined with a computational front which tracks the first-arriving wavefront to robustly and efficiently compute traveltimes on a grid. Ray paths can be extracted a posteriori by tracking the traveltime gradient from receiver to source [*Rawlinson et al.*, 2006]. Each time the misfit function is computed for a particular model, traveltime residuals can be rapidly computed by a simple integration of slowness along each path. Although this assumes that ray path geometry is largely independent of velocity structure, this is an acceptable approximation for teleseismic paths, particularly given that any dependence is second order. Moreover, by updating the paths occasionally as the inversion proceeds, the issue of path dependence is explicitly considered. Although a thorough assessment of the viability/usefulness of this approach will be reported elsewhere (Afonso et al., Paper III), preliminary tests confirm its feasibility within the present Bayesian framework.

[83] The use of a RF operator is attractive due to the increasing volume of receiver functions recorded by permanent stations and portable networks, the availability of a number of efficient techniques for forward modeling [e.g., *Frederiksen and Bostock*, 2000; *Herrmann*, 2002; *Sambridge*, 1999a], and the unique sensitivity of RF to the Moho structure [cf. *Zhu and Kanamori*, 2000; *Kind et al.*, 2012]. These methods use information from a variety of secondary wavefields (or phases) scattered off from heterogeneities beneath the seismic stations to locate first-order discontinuities of the impedance structure. In regions where data are available, the implementation of a RF operator into the first stage of the inversion could provide critical information, particularly when combined with ambient-noise data. We are currently testing these capabilities.

7.2. Single Stage Inversion

[84] The apparent complexity of the inversion scheme described in this paper is justified by the actual complexity of the problem. However, some simplifications and optimizations seem possible. Foremost among these is the unification of the two parts (1-D and 3-D) into a single full 3-D stage driven by a self-adaptive MCMC. To this respect, recent progress on the development of delayed rejection methods, parallel tempering, and adaptive Metropolis algorithms for high dimensions are particularly promising [e.g., *Haario et al.*, 2005; 2006; *Hasenbusch and Schaefer*, 2010; *Roberts and Rosenthal*, 2006; *Craiu et al.*, 2009]. If these strategies can be combined into a single optimized MCMC [e.g., *Haario et al.*, 2006], they may offer a more effective way of implementing multi-observable probabilistic inversion schemes. Alternatively, such adaptive MCMC scheme could be used for both stages of the inversion (1-D and 3-D), instead of a first search with the NA and then a Metropolis-Hastings algorithm. In either case, however, the viability of such approaches relies heavily on the non-trivial problem of implementing an efficient parallelization of the individual chains [e.g., *Wilkinson*, 2005; *Brockwell*, 2006].

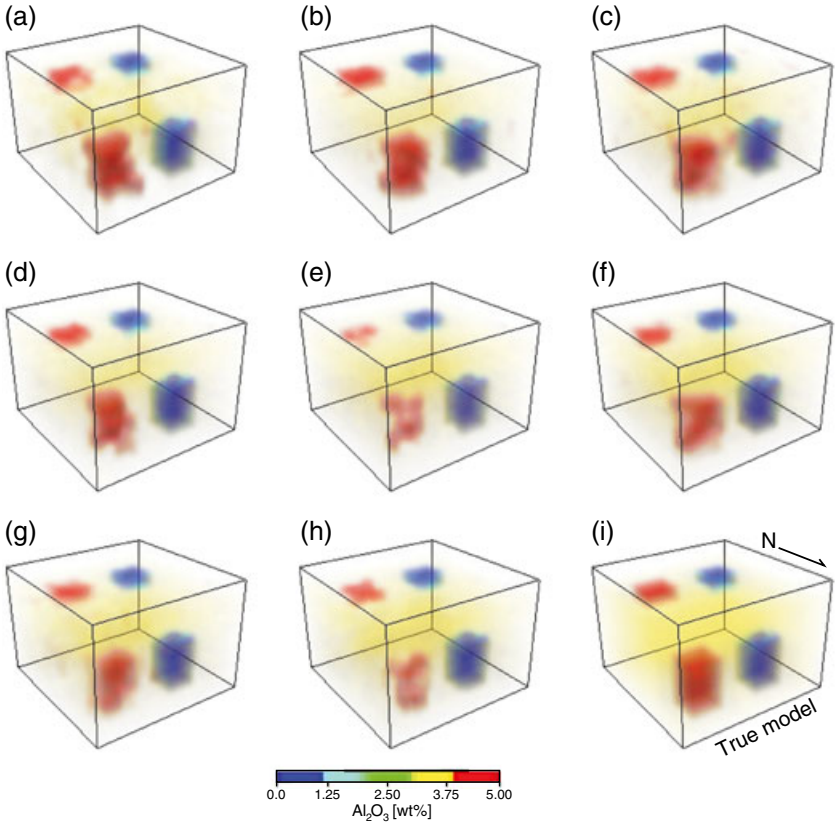

Figure 16. Same as in Figure 15, but now the color scale denotes the bulk Al_2O_3 content in the model. The true model (i) is shown for comparison.

8. Summary and Concluding Remarks

[85] There have been four major advances over the past 10 years that warrants the development of robust inverse methodologies capable of inverting directly for the thermochemical structure of the Earth's upper mantle: (i) improvements in instrumentation and data processing of geophysical data [e.g., *Shapiro et al.*, 2005; *Yang et al.*, 2008], concomitant with the installation of dense geophysical arrays in many regions of the world [e.g., EarthScope/US Array, PICASSO Array, Wombat Array, NECESS Array], (ii) recent progress in the thermodynamic and thermophysical modeling of the Earth's upper mantle [e.g., *Connolly*, 2009; *Stixrude and Lithgow-Bertelloni*, 2011; *Holland and Powell*, 2011], (iii) rapid growth in the computational power available to the scientific community, and (iv) comprehensive documentation of exhumed upper mantle samples and associated petrological/geochemical processes (see Afonso et al., Paper I, for references). The main purpose of this paper is to present a 3-D multi-observable probabilistic inversion method to study the thermal and compositional structure of the lithosphere and sub-lithospheric upper mantle, specifically designed to incorporate all of these advances in an internally consistent manner. The method is based on a Bayesian formalism (statistical inference), in which the solution to the inverse problem is fully characterized by the a posteriori probability density function, rather than single models.

[86] The present method offers several important advantages over more traditional inversion schemes when dealing with temperature and compositional structures simultaneously. First, the a priori compositional information is not based on restrictive ad hoc ranges or assumed age-composition relationships but on a statistical analysis of a large database of exhumed upper mantle samples from many different tectonic settings. This database can be easily augmented as more samples become available, or as theoretical petrological models evolve. Second, the geophysical observables used in the inversion have different sensitivities to temperature and bulk composition, as well as to their spatial distributions. Due to the internally consistent treatment of the observables, the use of important indicators for temperature versus compositional anomalies such as the ρ/Vs and/or topography/geoid ratios is intrinsically included in the method. Therefore, no ad hoc assumptions regarding the depth distribution of ρ/Vs are needed. This makes the method well suited to handle the simultaneous determination (probabilistically) of temperature and compositional anomalies. Third, both observational and theoretical uncertainties (and their inter-correlations) can be taken into account in a natural and general manner through probability density functions. In this way, the a posteriori probability density function contains all necessary information to assign realistic uncertainties on the final results.

[87] Although the bulk Mg# (MgO/[MgO+FeO]) of peridotites is typically considered the standard indicator of their composition and physical properties, we have shown that bulk Al_2O_3 content may be a better compositional proxy in the context of direct inversions for temperature and

composition. This is due to the strong effect that Al_2O_3 has on the bulk density and seismic velocities of peridotites, as well as its natural correlation with other important oxides (e.g., MgO, CaO) and with the average degree of depletion of the rock. Bulk "Al_2O_3 anomalies" in the upper mantle may therefore be one of the most informative indicators for scientific and large-scale prospecting studies.

[88] We have demonstrated that, even in the ideal case of incorporating multiple high-quality datasets, deep temperature anomalies $\lesssim$ 150°C and compositional anomalies ΔMg# < 3 are not simultaneously resolvable by either ours or any other method based on geophysical observables only. This is an unavoidable consequence of the uncertainties associated with both data and the current theoretical relationships between data and model parameters. Additional complications related to the non-uniqueness in the compositional space and the intrinsic behavior of peridotites' physical properties at high P-T conditions also contributes to this problem. Nevertheless, we emphasize that this resolving power in absolute magnitudes is sufficient for most practical and scientific purposes. It is not the ability to obtain accurate magnitude maps of temperature and compositional anomalies (which, in any case, is currently impossible) but rather the ability to simultaneously locate large anomalies that will allow us to examine long-standing problems about the nature and evolution of the lithosphere–sub-lithospheric upper mantle system (e.g., vertical stratification of cratonic mantle, compositional versus temperature signatures in seismic velocities, crust-mantle interactions, etc.). In this context, our results indicate that the present method is able to accomplish this (probabilistically) at horizontal and vertical scales of $\lesssim$ 100 and 40 km, respectively, provided the input data has equal or better resolution. Such high-resolution geophysical datasets are being attained in an increasing number of regions (e.g., USA, China, Australia, Japan, Spain, see above). The application of multi-observable probabilistic methods to these datasets opens new pathways to characterize the thermochemical structure of the lithosphere and upper mantle beneath these areas with unprecedented resolution.

Appendix A: 1-D Forward Problem

A1. Temperature

[89] The temperature structure is ultimately obtained by the inversion procedure (section 3.2.1). However, we need to define an initial temperature profile that is subsequently modified during the inversion by proposing finite changes ΔT_n and $\Delta \mathrm{d}T/\mathrm{d}z_n$. To obtain the initial thermal field in the model, we subdivide the numerical domain into three subdomains: (i) a conduction-dominated region in steady state (i.e., lithosphere), (ii) a convection-dominated region (i.e., convective sub-lithospheric upper mantle), and (iii) an intermediate zone where both conduction and convection are significant [cf. *Afonso et al.*, 2008; *Fullea et al.*, 2009]. We identify the bottom of the conduction-dominated domain as the LAB, and assign to it a value of 1315°C. This value is always associated with a specific thermodynamic node (defining the LAB), and thus, its temperature is not allowed to change during the inversion (only its position).

[90] Within the first domain, we solve the steady state heat conduction equation:

$$\nabla \cdot [k(z,T,P)\nabla T(z)] = -H(z) \quad \text{(A1)}$$

where T is temperature, P pressure, k thermal conductivity, H volumetric radiogenic heat production, and z the 1-D Cartesian coordinate (i.e., depth). We discretize equation (A1) with the finite-difference method; the nodal spacing is that of the computation nodes described in the text ($\leq$ 2 km). The boundary conditions are the temperature at the surface and at the bottom of the lithosphere (1315°C). The transition between the lithospheric mantle and the adiabatic sub-lithospheric mantle is modeled with a "thermal buffer" of constant thickness (35 km), but variable bottom temperature. This buffer mimics the thermal effect of a rheologically active layer present at the bottom of the upper thermal boundary layer in convecting mantle-like fluids [e.g., *Solomatov and Moresi*, 2000; *Zaranek and Parmentier*, 2004]. The temperature at the top of the buffer coincides with the LAB temperature; its bottom temperature, T_{bb}, is entirely controlled by the inversion (i.e., observables). Finally, the temperature in the sub-lithospheric mantle is given by $T_{\mathrm{sublith}} = T_{bb} + z * (\mathrm{d}T/\mathrm{d}z)_{\mathrm{adiabatic}}$, where z is depth and $(\mathrm{d}T/\mathrm{d}z)_{\mathrm{adiabatic}}$ is the adiabatic gradient. The latter is computed as $(T_{Bot} - T_{bb})/(z_{Bot} - z_{bb})$, where z_{Bot} and T_{Bot} are the depth and temperature at the bottom of the model, respectively. T_{Bot} is a free parameter and is obtained from the inversion.

[91] The conductive geotherms so obtained represent our first-order estimation of the temperature distribution with depth. These initial geotherms are then perturbed during the inversion (section 3.2.1), to obtain the final temperature structure consistent with the data.

A2. Elevation

[92] The application of hydrostatic equilibrium to the Earth's crust is commonly known as crustal isostasy (strictly, local crustal isostasy). Likewise, its extension to include the Earth's lithosphere [e.g., *Lachenbruch and Morgan*, 1990] may be referred to as lithospheric isostasy. Lateral pressure (and density) gradients associated with either temperature, topography, or compositional variations can exist within the lithosphere over long time scales due to its high viscosity. Indeed, several authors have demonstrated that the topography of the outer surface in mantle-like fluids is primarily controlled by the physical conditions and buoyancy of the uppermost thermal boundary layer [*McKenzie*, 1977; *Parsons and Daly*, 1983; *Marquart and Schmeling*, 1989]. In contrast, the relatively low viscosity of the shallow sub-lithospheric mantle (i.e., asthenosphere) allows lateral pressure gradients to be relaxed by flow over much shorter time scales. This creates a mechanical and temporal decoupling between the long-term rigid lithosphere and the much less viscous sub-lithospheric mantle, which explains the success of lithospheric isostatic models in reproducing absolute elevation.

[93] The formulae used in the isostatic calibration and calculation of absolute elevations have been described in detail elsewhere [*Afonso et al.*, 2008; *Fullea et al.*, 2009]. The only difference here is that we explicitly decouple contributions to topography from sub-lithospheric density anomalies. Therefore, the elevation of any column can be

written as [*Lachenbruch and Morgan*, 1990; *Afonso et al.*, 2008; *Fullea et al.*, 2009]

$$E = L_s - \frac{\overline{\rho_L}}{\overline{\rho_a}} L - \prod \tag{A2}$$

where

$$\overline{\rho_L} = \frac{1}{L} \int_{\text{top}}^{\text{bot}} \rho_L(z)\, \mathrm{d}z \tag{A3}$$

$$\overline{\rho_a} = \frac{1}{\text{bot} - \prod} \int_{\prod}^{\text{bot}} \rho_a(z)\, \mathrm{d}z \tag{A4}$$

and $\rho_L(z)$ is the depth-dependent density of the column, $\rho_a(z)$ the depth-dependent density of a reference adiabatic column, $L = bot - top$ the total height of the column (including water depth, w for columns with negative topography), L_s the thickness of the solid part of the column (e.g., $L - w$), and $\prod$ is a calibration parameter depending on the assumed reference column [see *Afonso et al.*, 2008]. To remove the contribution of sub-lithospheric density anomalies, $\rho_L(z)$ is divided into two parts:

$$\overline{\rho_L} = \frac{1}{\text{LAB}} \int_{\text{top}}^{\text{LAB}} \rho_L(z)\, \mathrm{d}z + \frac{1}{\text{bot} - \text{LAB}} \int_{\text{LAB}}^{\text{bot}} \rho_a(z)\, \mathrm{d}z \tag{A5}$$

where *LAB* is the depth to the lithosphere-asthenosphere boundary (i.e., 1315°C isotherm). By doing this, we only account for lithospheric contributions to elevation, while at the same time considering the effect of depth-dependent density in the sub-lithospheric mantle (i.e., buoyancy force depends on the total lithospheric thicknesses). Note that $\rho_a(z)$ is used *only* for elevation computations; all other observables are computed with the real density distribution of the sub-lithospheric mantle. This decoupling of sub-lithospheric contributions to elevation is necessary because density anomalies in the sub-lithospheric mantle are generally not compensated locally but dynamically. In practice, dynamic effects can be indirectly introduced into the inversion by filtering the signal or by assigning additional uncertainties to the elevation forward calculator. Future formulations could include a first-order approximation of dynamic effects by solving a simplified 1-D Stokes problem (work in progress). Given the area of the 1-D columns (typically > 6400 km^2), flexural effects are not included at this stage.

A3. Geoid Height

[94] We combine the gravity potential equation for cylindrical density anomalies [e.g., *Turcotte and Schubert*, 1982] with Brun's formula to obtain

$$\Delta N = \frac{2\pi G}{g_0} \int_0^h \Delta\rho(z)\, [(R^2 + z^2)^{1/2} - z]\, \mathrm{d}z \tag{A6}$$

where h is the maximum depth of the model, G the universal gravitational constant, and g_0 the reference acceleration of gravity on the reference geoid. For all practical purposes, it suffices to take $g_0 = 9.81$ m s^{-2} everywhere. We numerically integrate equation (A6) (trapezoidal rule) using the same internode distance as for the temperature calculation (computation nodes). We solve equation (A6) instead of the more familiar isostatic approximation [*Turcotte and Schubert*, 1982] due to the fact that not all density anomalies in our models are necessarily isostatically compensated (e.g., sub-lithospheric anomalies). The only unbounded parameter in equation (A6) is the radius of the column R. It is crucial to choose a value for R that guarantees that the 1-D assumption is a good first-order approximation of the real 3-D effects. Otherwise, when passing from the 1-D to 3-D stages of the inversion, the density structure constrained by the 1-D stage may not represent an acceptable model for the 3-D stage. There is no simple rule of thumb to determine the value of R, albeit intuitively it should be comparable to the actual radius of the column making up the model. Given the sensitivity of equation (A6) to the dipole moment of the density distribution, low values of R tend to underestimate the effects of shallow density anomalies compared to their real 3-D signature. Similarly, large values of R overestimate the effect of deep density anomalies. In practice, we find that a value of 80–100 km provides a good estimate of the true 3-D geoid (± 2.5 m).

A4. Surface Waves

[95] For the calculation of surface waves dispersion curves, we use a modified version (to account for frequency-dependent dispersion) of the code *disp96* [*Herrmann*, 2002], based on the Thompson-Haskell eigenvalue problem [cf. *Aki and Richards*, 2002]. For this problem, we use a variable thickness discretization (i.e., number of layers with constant density, *Vs* and *Vp*) based on the finest mesh (computation nodes); we subdivide each nodal distance of the finest mesh into 10 sub-layers with constant properties obtained by linear interpolation of the values at the computation nodes. Therefore, a typical discretization scale for the calculation of dispersion curves is ∼ 200 m. For depths below the 410 discontinuity (limit of our physical/numerical domain), we use the parameters from the *AK135* model [*Kennett et al.*, 1995].

[96] We use the laboratory-based model of *Jackson and Faul* [2010] to compute the quality factor for shear waves *Qs* as function of period, grain size, T, and P. The corresponding *Vs* velocities used in the calculation of dispersion curves are computed from [*Minster and Anderson*, 1981]

$$Vs(T,P,t,d) = Vs_0 \left[1 - \left(\frac{1}{2}\right) \cot\left(\frac{\pi\alpha}{2}\right) Qs^{-1}(T,P,t,d)\right] \tag{A7}$$

where Vs_0 is the anharmonic, unrelaxed velocity (obtained from the free energy minimization operator), and T, P, t, and d the temperature, pressure, period, and grain size, respectively. For the examples in this paper, we use a constant grain size $d = 5$ mm and a activation volume of 1.3×10^{-5} m^3 mol^{-1} [*Faul and Jackson*, 2005; *Afonso et al.*, 2008; *Jackson and Faul*, 2010] . *Vp* velocities can be computed similarly by replacing Qs^{-1} with $Qp^{-1} = L\ Qs^{-1} + (1 - L)Q_k^{-1}$, where $L = (4/3)(Vp/Vs)^2$ and Q_k is the bulk quality factor [*Anderson*, 1989]. We make two standard simplifications to solve for $Vp(T,P,t,d)$ in the upper mantle. First, we assume that Q_k^{-1} is identically zero (all losses occur in shear); second, we assume the Poisson solid approximation, which results in the simple relation $Qp = (9/4)\ Qs$ [*Anderson*, 1989]. *Vp* is then easily computed as [e.g., *Afonso et al.*, 2005]

$$Vp(T,P,t,d) = Vp_0 \left[1 - \left(\frac{2}{9}\right) \cot\left(\frac{\pi\alpha}{2}\right) Qs^{-1}(T,P,t,d)\right] \tag{A8}$$

where Vp_0 is the anharmonic, unrelaxed, compressional velocity obtained from the free energy minimization operator.

A5. Maxwell's Equations

[97] Theoretical 1-D MT responses are computed according to *Wait*'s [1954] recursion formula for an *N*-layered conductivity structure. Wait's formula is based on the continuity of the field components at the transition between the *n*th to (*n* + 1)th conductivity layer. This continuity property (strictly, a continuity condition for the Maxwell equations) allows us to compute the transfer function at the top of the *n*th layer provided that the transfer function at the top of the (deeper) (*n* + 1)th layer is known. Therefore, assuming that the lowermost *N*th layer is a homogeneous half-space with constant conductivity (i.e., the transfer function can be easily computed), it is possible to propagate the transfer function upwards by means of Wait's formula. In the final step, the apparent resistivity and the phase at the surface of the Earth are computed from the transfer function and compared with the measured MT data.

A6. Equilibrium Assemblages

[98] Equilibrium assemblages at specific P-T-C conditions and all their relevant physical properties are computed with components of the free energy minimization software Perple_X [*Connolly*, 2009]. The algorithms used in the energy minimization problem as well as the methods to compute the bulk properties of the equilibrium assemblages have been described in detail elsewhere [*Connolly*, 2005, 2009]; the reader is referred to these references for details.

Appendix B: Removing Information in the NA

[99] The NA generates new samples in each iteration based on the fitting properties of *all* previous samples. Therefore, the NA keeps track of these samples and their associated Voronoi cells in order to generate new samples in each iteration. This is an important feature that makes the NA more robust [*Sambridge*, 1999a] than other search algorithms, particularly for problems with multiple minima. In our problem, however, most of the original samples generated in the first two to four iterations are never sampled again in the course of the search (the reason for generating a large number of samples at the beginning of the search has been explained in section 3.2) due to their significantly higher misfit values. Keeping this (useless) information during the search deteriorates the performance of the NA, since it has to loop over all previous samples at the beginning of each iteration. In order to accelerate the search and take full advantage of our strategy for solving the free energy minimization problem (section 3.2.2), we remove all samples that have not been re-sampled after 10 iterations.

Appendix C: 3-D Forward Problem

C1. Temperature

[100] The only difference between the 1-D and 3-D forward problems for temperature is that the conductive part (lithosphere) is solved in 3-D Cartesian geometry. The model is discretized in a 3-D grid with *Nx*, *Ny*, and *Nz* nodes in the *X*, *Y*, and *Z* axes, respectively. While *Nz* is the same as in the 1-D problem, *Nx* and *Ny* are typically 3–4 times larger than the number of nodes defining the lateral extension of the 1-D columns. This is necessary for numerical reasons only; the original number of *Nx* and *Ny* nodes is too small to allow an accurate determination of the 3-D temperature structure than can be compared with the 1-D approximation.

[101] The boundary conditions are (i) no lateral heat flow at the lateral limits of the model, (ii) constant temperature at the surface, and (iii) constant temperature at the bottom of the lithosphere. The current numerical implementation uses a least-square iterative scheme to solve the linear system of equations resulting from the discretization of the heat transfer equation [*Fullea et al.*, 2009]. No attempts have been made to test direct solvers, although this is part of ongoing work.

C2. 3-D Maxwell's Equations

[102] The computational cost associated with the solution of this forward problem is notoriously large and renders it impractical for probabilistic formulations. Other approximations, such as 2-D solutions embedded in the 3-D model or post-inversion consistency control, may be worth exploring in future implementations of probabilistic methods.

C3. Gravity and Geoid Anomalies

[103] Gravity and geoid anomalies in every surface point of the model are calculated adding the effect of all individual columns. The vertical gravitational attraction produced by a right rectangular prism whose density varies linearly with depth can be expressed analytically in Cartesian coordinates as [*Gallardo-Delgado et al.*, 2003]

$$\begin{aligned}\Delta g(\rho) = G\rho_0 \Big|\Big|\Big| & x\,\ln(y+r) + y\,\ln(x+r) \\ & - z\,\arctan\left(\frac{xy}{zr}\right)\Big|_{x1}^{x2}\Big|_{y1}^{y2}\Big|_{z1}^{z2} \\ & + G\gamma \Big|\Big|\Big| -xy\,\ln(r+z) - \frac{z^2}{2}\arctan\left(\frac{xy}{zr}\right) \\ & + \frac{x^2}{2}\arctan\left(\frac{yz}{xr}\right) + \frac{y^2}{2}\arctan\left(\frac{xz}{yr}\right)\Big|_{x1}^{x2}\Big|_{y1}^{y2}\Big|_{z1}^{z2}\end{aligned} \tag{C1}$$

where

$$r = \sqrt{x^2 + y^2 + z^2} \tag{C2}$$

$$\gamma = (\rho(z) - \rho_0)/z \tag{C3}$$

G is the universal gravitational constant and ρ_0 is a constant reference density. The first term on the right-hand side of equation (C1) is equivalent to the contribution of a constant density prism [e.g., *Nagy et al.*, 2000]. The second term introduces the effect of the linear density gradient γ. The corresponding geoid anomaly for the same prism is [*Fullea et al.*, 2009]

$$\begin{aligned}\Delta N(\rho) = \frac{G\rho_0}{g_0} \Big|\Big|\Big| & xy\,\ln(r+z) \\ & + zy\,\ln(r+x) + xz\,\ln(r+y) - F_g\Big|_{x1}^{x2}\Big|_{y1}^{y2}\Big|_{z1}^{z2} \\ & + \frac{G\gamma}{3g_0} \Big|\Big|\Big| xyr + \frac{y}{2}(y^2+3z^2)\,\ln(r+x) \\ & + \frac{x}{2}(x^2+3z^2)\,\ln(y+r) - z^3\arctan\left(\frac{xy}{zr}\right)\Big|_{x1}^{x2}\Big|_{y1}^{y2}\Big|_{z1}^{z2}\end{aligned} \tag{C4}$$

where

$$F_g = \frac{z^2}{2}\arctan\left(\frac{xy}{zr}\right) + \frac{x^2}{2}\arctan\left(\frac{zy}{xr}\right) + \frac{y^2}{2}\arctan\left(\frac{xz}{yr}\right) \quad \text{(C5)}$$

[104] Columns located at the edges of the model are extended 10^6 km to avoid border effects. The way in which regional components and reference models are subtracted follows the description in *Fullea et al.* [2009]; the reader is referenced to this work for more details.

[105] The corresponding anomalies in every surface point of the model are calculated adding the effect of all individual prisms:

$$PF(x,y) = \sum_{\text{prisms}} f(\rho)\, g(x,y,z) \quad \text{(C6)}$$

where $PF(x,y)$ is the geoid or gravity anomaly over a point at the surface of the model, $f(\rho)$ is a function of the density distribution within the prism, and $g(x,y,z)$ represents a geometrical contribution that depends only on the spatial coordinates of each prim. This decomposition is particularly convenient when computing the 3-D gravity field a large number of times, where only the density of one or two prisms (i.e., columns) changes at a time. The rationale is that the geometrical contribution g of each prism needs to be computed only once at the beginning of the 3-D phase (provided that the grid is kept fixed). Once these geometrical contributions are computed for each prism and saved in memory, the 3-D effect of changing the density structure of a particular column during the inversion process can be computed quickly, without having to consider the entire 3-D density structure.

Appendix D: Convergence of the Chains

[106] The crucial problem of estimating the minimum number of samples (iterations) in MCMC simulation needed to achieve a specified accuracy of the posterior has no unique answer. Many different methods (and criticisms) can be found in the statistics literature [cf. *Gelman*, 1996; *Gelman et al.*, 2003; *Wilkinson*, 2005; *Congdon*, 2006], but the problem is typically ignored in the geophysical literature. Although a thorough discussion of the benefits and limitations of such methods is beyond the scope of this paper, we briefly describe our preferred approach.

[107] The Gelman-Rubin scale-reduction factor [*Gelman and Rubin*, 1992] is an attractive option for our case, since it provides information on convergence and consistency of multiple chains. The idea is that parameter samples from poorly identified models will exhibit significant divergence in the sample paths between different chains, and thus, the variability of sampled parameter values between chains will be larger than the variability within any one chain. Given J chains with T iterations each (after burn-in), the variability of samples ψ_{jt} within the jth chain is defined as

$$w_j = \frac{\left(\psi_{jt} - \overline{\psi_j}\right)^2}{T-1} \quad \text{(D1)}$$

where $\overline{\psi_j}$ is the average of all ψ_{jt} defined over T iterations. The variability within chains W is simply the average of the w_j. Similarly, the variance between chains is defined as

$$B = \frac{T}{J-1}\sum_{j=1}^{J}\left(\overline{\psi_j} - \overline{\psi}\right)^2 \quad \text{(D2)}$$

where $\overline{\psi}$ is the average of $\overline{\psi_j}$. With these definitions, a simple pool estimator of the variance of ψ is $V = B/T + TW/(T-1)$ [*Congdon*, 2006]. The so-called potential scale reduction factor, *GRSRF*, [*Gelman and Rubin*, 1992] compares V with W as $GRSRF = \sqrt{(V/W)}$, which reduces to 1 as $\psi \to \infty$. Convergence of the chains can be assumed when $GRSRF < 1.2$. In our synthetic case studies, *GRSRF* values ~ 1.2 are obtained $T \lesssim 1000000$.

[108] **Acknowledgments.** This manuscript benefited from discussions with A. Khan, N. Rawlinson, M. Fernàndez, M. Sambridge, B. Kennett, B. Griffin, and S. Lebedev. We thank M. Sambridge, M. Laine, and H. Haario for making their MCMC and visualization codes available. Suggestions by S. Fishwick, Uli Faul, and an anonymous reviewer helped to improve this paper. We also thank R. Herrmann and U. Faul for making their dispersion and attenuation codes available. Some figures were made with the free packages GMT (http://gmt.soest.hawaii.edu/) and Visit (https://wci.llnl.gov/codes/visit/). The work of JCA and YY have been supported by an ARC (Australian Research Council) Discovery grant (DP120102372). AGJ acknowledges financial support from Science Foundation Ireland (grants 07/RFP/GEOF759 and 10/IN.1/I3022-STTF 11). JF was supported by an SFI PI grant (10/IN.1/I3022) to AGJ for IRETHERM (http://www.iretherm.ie www.iretherm.ie), by the JAE-DOC programme (CSIC-Spain), and by Spanish Ministry of Economy and Competitiveness grant CGL2012-37222. This is contribution 218 from the ARC Center of Excellence for Core to Crust Fluid Systems (http://www.ccfs.mq.edu.au) and 856 in the GEMOC Key Center (http://www.gemoc.mq.edu.au).

References

Afonso, J. C., G. Ranalli, and M. Fernàndez (2005), Thermal expansivity and elastic properties of the lithospheric mantle: Results from mineral physics of composites , *Phys. Earth Planet. Int.*, *149*, 279–306., doi: http://dx.doi.org/10.1016/j.pepi.2004.10.003.

Afonso, J. C., M. Fernàndez, G. Ranalli, W. L. Griffin, and J. A. D. Connolly (2008), Integrated geophysical-petrological modelling of the lithospheric-sublithospheric upper mantle: Methodology and applications, *Geochem. Geophys. Geosyst.*, *9*, Q05008, doi: 10.1029/2007GC001834.

Afonso, J. C., and D. Schutt (2012), The effects of polybaric partial melting on the density and seismic velocities of mantle restites, *Lithos*, *134–135*, 289–303., doi:http://dx.doi.org/10.1016/j.lithos.2012.01.009.

Afonso, J. C., G. Ranalli, M. Fernàndez, W. L. Griffin, S. Y. OReilly, and U. Faul (2010), On the Vp/Vs-Mg# correlation in mantle peridotites: Implications for the identification of thermal and compositional anomalies in the upper mantle, *Earth Planet. Sci. Lett.*, *289*, 606–618., doi: http://dx.doi.org/10.1016/j.epsl.2009.12.005.

Aki, K., and P. G. Richards (2002), *Quantitative Seismology* 2nd Ed., University Science Books, Sausalito, California.

Anderson, D. L. (1989), *Theory of the Earth*, vol. 299, Blackwell Scientific Publications, Boston. http://resolver.caltech.edu/CaltechBOOK: 1989.001 02/02/2012.

Becker, T. W. (2012), On recent seismic tomography for the western United States, *Geochem. Geophys. Geosyst.*, *13*, Q01W10, doi: 10.1029/2011GC003977..

Bensen, G. D., M. H. Ritzwoller, M. P. Barmin, A. L. Levshin, F. Lin, M. P. Moschetti, N. M. Shapiro, and Y. Yang (2007), Processing seismic ambient noise data to obtain reliable broad-band surface wave dispersion measurements, *Geophys. J. Int.*, *169*, 1239–1260, doi:10.1111/j.1365-246X.2007.03374.X.

Biegler, L. et al. (eds.) (2011), *Large-Scale Inverse Problems and Quantification of Uncertainty*, Wiley and Sons, UK.

Brandenburg, J. P., E. H. Hauri, P. E. van Keken, and C. J. Ballentine (2008), A multiple-system study of the geochemical evolution of the mantle with force-balanced plates and thermochemical effects, *Earth Planet. Sci. Lett.*, *276*, 1–13., doi: http://dx.doi.org/10.1016/j.epsl.2008.08.027.

Brockwell, A. E. (2006), Parallel Markov Chain Monte Carlo simulation by pre-fetching, *J. Comp. Graph. Stat.*, *15* (1), 246–261., doi: 10.1198/106186006X100579.

Cammarano, F., P. Tackley, and L. Boschi (2011), Seismic, petrological and geodynamical constraints on thermal and compositional structure of the upper mantle: Global thermo-chemical models, *Geophys. J. Int.*, *187*, 1301–1318., doi:10.1111/j.1365-246X.2011.05223.x.

Carlin, B. P., and T. A. Louis (2000), *Bayes and Empirical Bayes Methods for Data Analysis* 2nd ed, Chapman and Hall/CRC, USA.

Claerbout, J. F., and F. Muir (1973), Robust modelling with erratic data, *Geophysics*, *38*, 826–844.

Congdon, P. (2006), *Bayesian Statistical Modelling* 2nd ed., Wiley & Sons, Sussex, UK.

Connolly, J. A. D. (2005), Computation of phase equilibria by linear programming: A tool for geodynamic modelling and its application to subduction zone decarbonation, *Earth Planet. Sci. Lett.*, *236*, 524–541., doi:10.1016/j.epsl.2005.04.033.

Connolly, J. A. D. (2009), The geodynamic equation of state: What and how, *Geochem. Geophys. Geosys.*, *10*, Q10014, doi:10.1029/2009GC002540.

Craiu, R. V., J. Rosenthal, and C. Yang (2009), Learn from thy neighbor: Parallel-chain and regional adaptive MCMC, *J. Am. Stat. Assoc*, *104*, 1454–1466., doi:10.1198/jasa.2009.tm08393.

Deschamps, F., J. Trampert, and R. Snieder (2002), Anomalies of temperature and iron in the uppermost mantle inferred from gravity data and tomographic models, *Phys. Earth Planet Int.*, *129*, 245–264.

Eltoft, T., T. Kim, and T. Lee (2006), On the multivariate Laplace distribution, *IEEE Signal Proc. Lett.*, *13*, 300–303., doi:10.1109/LSP.2006.870353.

Faul, U. H., and I. Jackson (2005), The seismological signature of temperature and grain size variation in the upper mantle, *Earth Planet. Sci. Lett.*, *234*, 119–134., doi:10.1016/j.epsl.2005.02.008.

Fishwick, S. (2010), Surface wave tomography: Imaging of the lithosphere-asthenosphere boundary beneath central and southern Africa? *Lithos*, *120*, 63–73., doi:10.1016/j.lithos.2010.05.011.

Forte, A. M., and H. K. C. Perry (2000), Geodynamic evidence for a chemically depleted continental tectosphere, *Science*, *290*, 1940–1944.

Frederiksen, A. W., and M. G. Bostock (2000), Modelling teleseismic waves in dipping anisotropic structures, *Geophys. J. Int.*, *141*, 401–412.

Fullea, J., J. C. Afonso, J. A. D. Connolly, M. Fernàndez, D. Garcia-Castellanos, and H. Zeyen (2009), LitMod3D: An interactive 3-D software to model the thermal, compositional, density, seismological, and rheological structure of the lithosphere and sublithosheric upper mantle, *Geochem. Geophys. Geosyst.*, *10*, Q08019, doi:10.1029/2009GC002391.

Gallardo-Delgado, L. A., M. A. Pérez-Flores, and E. Gómez-Treviño (2003), A versatile algorithm for joint 3D inversion of gravity and magnetic data, *Geophysics*, *68*, 949–959, doi:10.1190/1.1581067.

Gelman, A. (1996), Inference and monitoring convergence, in *Markov Chain Monte Carlo in Practice*, edited by W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, Chapman and Hall, USA, 131-144.

Gelman, A., and D. Rubin (1992), Inference from iterative simulation using multiple sequences, *Stat. Sci.*, *7*, 457–511.

Gelman, A., J. Carlin, H. Stern, and D. B. Rubin (2003), *Bayesian Data Analysis* 2nd ed., Chapman and Hall/CRC, New York, USA.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (eds.) (1996), *Markov Chain Monte Carlo in Practice*, Chapman and Hall, USA.

Gouveia, W. P., and J. A. Scales (1998), Bayesian seismic waveform inversion: Parameter estimation and uncertainty analysis, *J. Geophys. Res.*, *103*(B2), 2759–2779, doi:10.1029/97JB02933.

Gregory, P. (2005), *Bayesian Logical Data Analysis for the Physical Sciences*, Cambridge Univ. Press, Cambridge, UK.

Haario, H., E. Saksman, and J. Tamminen (2005), Componentwise adaptation for high dimensional MCMC, *Comp. Stat.*, *20*, 265–273., doi:10.1007/BF02789703.

Haario, H., M. Laine, A. Mira, and E. Saksman (2006), DRAM: Efficient adaptive MCMC, *Stat. Comp.*, *16*, 339–354., doi:10.1007/s11222-006-9438-0.

Hasenbusch, M., and S. Schaefer (2010), Speeding up parallel tempering simulations, *Phys. Rev. E*, *82*, 046707., doi: 10.1103/PhysRevE.82.046707.

Hasterok, D., and D. S. Chapman (2011), Heat production and geotherms for the continental lithosphere, *Earth Planet. Sci. Lett.*, *307*, 59–70., doi: http://dx.doi.org/10.1016/j.epsl.2011.04.034.

Herrmann, R. B. (2002), *Computer Program in Seismology,* Version 3.3, Saint Louis University . http://www.eas.slu.edu/eqc/eqcsoftware.html.

Herzberg, C. (1999), Phase equilibrium constraints on the formation of mantle, in *Mantle Petrology: Field Observations and High Pressure Experimentation: A Tribute to Francis F. (Joe) Boyd. Geochemical Society*, edited by Fei, Y., C. M. Bertka, and B. O. Mysen, Special Publication 6, Houston: Geochemical Society, 241–258.

Holland, T. J. B., and R. Powell (2011), An improved and extended internally consistent thermodynamic dataset for phases of petrological interest, involving a new equation of state for solids, *J. Metamorph. Geol.*, *29*, 333–383., doi:10.1111/j.1525-1314.2010.00923.x.

Jackson, I., and U. H. Faul (2010), Grainsize-sensitive viscoelastic relaxation in olivine: Towards a robust laboratory-based model for seismological application, *Phys. Earth Planet Int.*, *183*, 151–163., doi: http://dx.doi.org/10.1016/j.pepi.2010.09.005.

Jaupart, C., and J.-C. Mareschal (2011), *Heat Generation and Transport in the Earth*, Cambridge Univ. Press, Cambridge.

Jokinen, J., and I. T. Kukkonen (1999), Random modelling of the lithospheric thermal regime: Forward simulations applied in uncertainty analysis, *Tectonophysics*, *306*, 277–292.

Jones, A. G., R. L. Evans, and D. W. Eaton (2009), Velocity-conductivity relationships for mantle mineral assem1 blages in Archean cratonic lithosphere based on a review of laboratory data and HashinShtrikman extremal bounds, *Lithos*, *109*, 131–143., doi: http://dx.doi.org/10.1016/j.lithos.2008.10.014.

Kennett, B. L. N., E. R. Engdahl, and R. Buland (1995), Constraints on seismic velocities in the Earth from travel times, *Geophys. J. Int.*, *122*, 108–124.

Khan, A., J. A. D. Connolly, J. Maclennan, and K. Mosegaard (2007), Joint inversion of seismic and gravity data for lunar composition and thermal state, *Geophys. J. Int.*, *168*, 243, doi: 10.1111/j.1365-246X.2006.03200.X.

Khan, A., L. Boschi, and J. A. D. Connolly (2011a), Mapping the Earth's thermochemical and anisotropic structure using global surface wave data, *J. Geophys. Res.*, *116*, B01301, doi:10.1029/2010JB007828.

Khan, A., A. Zunino, and F. Deschamps (2011b), The thermo-chemical and physical structure beneath the North American continent from Bayesian inversion of surface-wave phase velocities, *J. Geophys. Res.*, *116*, B09304, doi:10.1029/2011JB008380.

Kind, R., X. Yuan, and P. Kumar (2012), Seismic receiver functions and the lithosphere-asthenosphere boundary, *Tectonophysics*, *536–537*, 25–43., doi:http://dx.doi.org/10.1016/j.tecto.2012.03.005.

Kitanidis, P. K. (2011), Bayesian and geostatistical approaches to inverse problems in *Large-Scale Inverse Problems and Quantification of Uncertainty*, edited by L. Biegler et al., 71–85, Wiley and Sons, UK.

Lachenbruch, A. H., and P. Morgan (1990), Continental extension, magmatism and elevation; formal relations and rules of thumb, *Tectonophysics*, *174*, 39–62, doi:10.1016/0040-1951(90)90383-J.

Lin, F. C., M. H. Ritzwoller, and R. Snieder (2009), Eikonal tomography: Surface wave tomography by phase-front tracking across a regional broad-band seismic array, *Geophys. J. Int.*, *177*, 1091–1110., doi:10.1111/j.1365-246X.2009.04105.x.

Malinverno, A., and V. A. Briggs (2004), Expanded uncertainty quantification in inverse problems: Hierarchical Bayes and empirical Bayes, *Geophysics*, *69*, 1005–1016.

Marquart, G., and H. Schmeling (1989), Topography and geoid undulations caused by small scale convection beneath continental lithosphere of variable elastic thickness, *Geophys. J. Int.*, *97*, 511–527, doi:10.1111/j.1365-246X.1989.tb00520.X.

McKenzie, D. (1977), Surface deformation, gravity anomalies and convection, *Geophys. J. R. Astron. Soc.*, *48*, 211–238.

Minster, B., and D. L. Anderson (1981), A model of dislocation controlled rheology for the mantle , *Philos. Trans. R. Soc. London, Ser. A*, 299–319

Mosegaard, K., and A. Tarantola (1995), Monte Carlo sampling of solutions to inverse problems, *J. Geophys. Res.*, *100*(B7), 12431–12447.

Mosegaard, K., and A. Tarantola (2002), *Probabilistic Approach to Inverse Problems: Chapter for the International Handbook of Earthquake and Engineering Seismology*, 237–265, Academic Press, Amsterdam, Elsevier. ISBN 0-12-440652-1.

Nagy, D., G. Papp, and J. Benedek (2000), The gravitational potential and its derivatives for the prism, *J. Geod.*, *74*, 552–560, doi:10.1007/s001900000116.

Nakagawa, T., P. J. Tackley, F. Deschamps, and J. A. D. Connolly (2010), The influence of MORB and harzburgite composition on thermo-chemical mantle convection in a 3-D spherical shell with self-consistently calculated mineral physics, *Earth Planet. Sci. Lett.*, *296*, 403–412., doi:http://dx.doi.org/10.1016/j.epsl.2010.05.026.

Oh, S.-H., and B.-D. Kwon (2001), Geostatistical approach to Bayesian inversion of geophysical data: Markov chain Monte Carlo method, *Earth Planets Space*, *53*, 777–791.

O'Reilly, S. Y., and W. L. Griffin (2010), Rates of magma ascent: Constraints from mantle-derived xenoliths, in *Timescales of Magmatic Processes: From Core to Atmosphere*, edited by A. Dosseto, S.P. Turner, and J. A. Van Orman, Blackwell Publishing Ltd., Chichester, West Sussex, UK ; Hoboken, NJ : Wiley-Blackwell, 116–124.

Pasyanos, M. E. (2010), Lithospheric thickness modeled from long-period surface wave dispersion, *Tectonophysics*, *481*, 38–50., doi: 10.1016/j.tecto.2009.02.023.

Parsons, B., and S. Daly (1983), The relationship between surface topography, gravity anomalies, and the temperature structure of convection, *J. Geophys. Res.*, *88*, 1129–1144, doi:10.1029/JB088iB02p01129.

Rawlinson, N., and M. Sambridge (2004), Wavefront evolution in strongly heterogeneous layered media using the fast marching method, *Geophys. J. Int.*, *156*, 631–647., doi:10.1016/j.epsl.2006.01.008.

Rawlinson, N., A. M. Reading, and B. L. N. Kennett (2006), Lithospheric structure of Tasmania from a novel form of teleseismic tomography, *J. Geophys. Res.*, *111*, B02301, doi:10.1029/2005JB003803.

Rickwood, P., and M. Sambridge (2006), Efficient parallel inversion using the Neighbourhood Algorithm, *Geochem. Geophys. Geosys.*, *7*, Q11001, doi:10.1029/2006GC001246.

Ringwood, A. E. (1975), *Composition and Petrology of the Earth's Mantle*, McGraw-Hill, New York.

Roberts, G. O., and J. S. Rosenthal, (2006), Examples of adaptive MCMC, *Technical Report*, University of Toronto.

Rosenthal, J. S. (2000), Parallel computing and Monte Carlo algorithms, *Far East Journal of Theoretical Statistics*, *4*, 207–236.

Sambridge, M. (1998), Exploring multidimensional landscapes without a map, *Inverse Probl.*, *14*, 427–440.

Sambridge, M. (1999a), Geophysical inversion with a neighbourhood algorithm—I. Searching a parameter space, *Geophys. J. Int.*, *138*, 479–494.

Sambridge, M. (1999b), Geophysical inversion with a neighbourhood algorithm—II. Appraising the ensemble, *Geophys. J. Int.*, *138*, 727–746.

Sambridge, M., K. Gallagher, A. Jackson, and P. Rickwood (2006), Trans-dimensional inverse problems, model comparison and the evidence, *Geophys. J. Int.*, *167*, 528–542., doi:10.1111/j.1365-246X.2006.03155.x.

Scales, J. A., and L. Tenorio (2001), Prior information and uncertainty in inverse problems, *Geophysics*, *66*, 389–397.

Scales, J. A., and R. Snider (1997), Bayes or not to Bayes? *Geophysics*, *62*, 1045–1046.

Shapiro, N. M., M. Campillo, L. Stehly, and M. H. Ritzwoller (2005), High-resolution surface wave tomography from ambient seismic noise, *Science*, *307*, 1615–1618., doi:10.1126/science.1108339.

Solomatov, V. S., and L.-N. Moresi (2000), Scaling of time-dependent stagnant lid convection: Application to small-scale convection on the Earth and other terrestrial planets, *J. Geophys. Res.*, *105*, 21795–21818.

Stixrude, L., and C. Lithgow-Bertelloni (2011), Thermodynamics of mantle minerals—II. Phase equilibria, *Geophys. J. Int.*, *184*, 1180–1213, doi: 10.1111/j.1365-246X.2010.04890.x.

Tackley, P. J., and S. Xie (2002), The thermochemical structure and evolution of Earth's mantle: Constraints and numerical models, *Phil. Trans. R. Soc. Lond. A*, *360*, 2593–2609., doi:10.1098/rsta.2002.1082.

Tarantola, A. (2005), *Inverse Problem Theory and Model Parameter Estimation*, Soc. for Ind. and Appl. Math., Philadelphia.

Tarantola, A., and B. Valette (1982), Inverse problems = quest for information, *J. Geophys.*, *50*, 159–170.

Tirone, M., G. Sen, and J. P. Morgan (2012), Petrological geodynamic modeling of mid-ocean ridges, *Physics Earth Planet Int.*, *190–191*, 51–70, doi:10.1016/j.pepi.2011.10.008.

Turcotte, D. L., and G. Schubert (1982), *Geodynamics: Applications of Continuum Physics to Geological Problems*, John Wiley, New York.

van Gerven, M., B. Cseke, R. Oostenveld, and T. Heskes (2009), Bayesian source localization with the multivariate Laplace prior , in *Advances in Neural Information Processing Systems*, edited by Bengio, Y., D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, vol. 22, 1901–1909.

Venables, W. N., and B. D. Ripley (1999), *Modern Applied Statistics with S-PLUS* 3rd ed., Springer-Verlag, New York.

Vilá, M., M. Fernàndez, and I. Jiménez-Munt (2010), Radiogenic heat production variability of some common lithological groups and its significance to lithospheric thermal modeling, *Tectonophysics*, *490*, 152–164., doi:http://dx.doi.org/10.1016/j.tecto.2010.05.003.

Wait, J. R. (1954), On the relation between telluric currents and the Earth's magnetic field, *Geophysics*, *19*, 281–289.

Walter, M. J. (2004), Melt extraction and compositional variability in mantle lithosphere, in *The Mantle and Core (ed Carlson, R.W.), Vol 2, Treatise on Geochemistry*, edited by Holland, H. D., and K. K Turkeian, 363–394, Elsevier-Pergamon, Oxford.

Wilkinson, D. J. (2005), Parallel Bayesian computation, in *Handbook of Parallel Computing and Statistics*, edited by E. J. Kontoghiorghes, Marcel Dekker/CRC Press, Chapman and Hall/CRC, Broken Sound Park, NW, FL., 481–512.

Yang, Y., M. H. Ritzwoller, F.-C. Lin, M. P. Moschetti, and N. M. Shapiro (2008), Structure of the crust and uppermost mantle beneath the western United States revealed by ambient noise and earthquake tomography, *J. Geophys. Res.*, *113*, B12310, doi:10.1029/2008JB005833.

Zaranek, S. E., and E. M. Parmentier (2004), Convective cooling of an initially stably stratified fluid with temperature-dependent viscosity: Implications for the role of solid-state convection in planetary evolution, *J. Geophys. Res.*, *109*, B03409, doi:10.1029/2003JB002462.

Zhdanov, M. S. (2002), *Geophysical Inverse Theory and Regularization Problems, Methods in Geochemistry and Geophysics*, vol. 36, Elsevier, Amsterdam, Elsevier.

Zhu, L., and H. Kanamori (2000), Moho depth variation in southern California from teleseismic receiver functions, *J. Geophys. Res.*, *105* (B2), 2969–2980.