Probabilistic surface reconstruction from multiple data sets: An example for the Australian Moho

T. Bodin,[1] M. Salmon,[2] B. L. N. Kennett,[2] and M. Sambridge[2]

Received 21 June 2012; revised 13 September 2012; accepted 16 September 2012; published 19 October 2012.

[1] Interpolation of spatial data is a widely used technique across the Earth sciences. For example, the thickness of the crust can be estimated by different active and passive seismic source surveys, and seismologists reconstruct the topography of the Moho by interpolating these different estimates. Although much research has been done on improving the quantity and quality of observations, the interpolation algorithms utilized often remain standard linear regression schemes, with three main weaknesses: (1) the level of structure in the surface, or smoothness, has to be predefined by the user; (2) different classes of measurements with varying and often poorly constrained uncertainties are used together, and hence it is difficult to give appropriate weight to different data types with standard algorithms; (3) there is typically no simple way to propagate uncertainties in the data to uncertainty in the estimated surface. Hence the situation can be expressed by Mackenzie (2004): "We use fantastic telescopes, the best physical models, and the best computers. The weak link in this chain is interpreting our data using 100 year old mathematics". Here we use recent developments made in Bayesian statistics and apply them to the problem of surface reconstruction. We show how the reversible jump Markov chain Monte Carlo (rj-McMC) algorithm can be used to let the degree of structure in the surface be directly determined by the data. The solution is described in probabilistic terms, allowing uncertainties to be fully accounted for. The method is illustrated with an application to Moho depth reconstruction in Australia.

Citation: Bodin, T., M. Salmon, B. L. N. Kennett, and M. Sambridge (2012), Probabilistic surface reconstruction from multiple data sets: An example for the Australian Moho, *J. Geophys. Res.*, *117*, B10307, doi:10.1029/2012JB009547.

1. Introduction

[2] Surface fitting, spatial prediction, regular griding or data interpolation are problems which often occur in many fields of geosciences. They are all specific cases of a general well known problem in statistics: the regression problem. Some noisy records of a continuous function (e.g., temperature, gravity, concentration of stable isotopes) are recorded at some discrete locations in space or time, and the problem consists of recovering this unknown function.

[3] Recent one-dimensional examples include time series analysis of geochemical proxies [e.g., *Large et al.*, 2009; *Burton et al.*, 2010; *Kylander et al.*, 2010]. Two-dimensional examples (i.e. surface reconstruction) include gridding of satellite measurements for mapping gravity anomalies [*Sandwell and Smith*, 1997], interpolation of digital elevation data for landscape reconstruction [*Atkinson and Lloyd*, 2007], interpolation of monthly precipitation data [*Lloyd et al.*, 2010], interpolation of aeromagnetic data [*Billings et al.*, 2002], or reconstruction of the Moho discontinuity from geophysical data [*Kennett et al.*, 2011; *Di Stefano et al.*, 2011].

[4] There are a large number of surface fitting algorithms, with different features and limitations (for a comparative review within the geosciences, see *El Abbas et al.* [1990]). For example, Kriging [*Stein*, 1999] is based on several assumptions. First, the data are seen as random realizations of a normal distribution whose mean and variance are constant over the 2D field. Second, the spatial variability of data (i.e. the data covariance or variogram function) is also assumed constant over the 2D field. These stationarity assumptions make Kriging only adapted to a certain range of problems in geosciences.

[5] Generally, most 2D interpolation methods used in Earth sciences estimate surface values from weighted averages of nearby data points, a procedure justified by the assumption that the surface varies smoothly with distance. Weighted-average schemes differ in how they assign weights to the constraining values. The simplest methods use a polynomial or power law in distance [*Smith and Wessel*, 1990]. The requirement is that the surface should minimize some global norm (e.g., distance to a reference surface, level of smoothness, spline tension, …) while fitting the data in a least square

[1]Earth and Planetary Science, University of California, Berkeley, California, USA.
[2]Research School of Earth Sciences, Australian National University, Canberra, ACT, Australia.

Corresponding author: T. Bodin, Earth and Planetary Science, University of California, 173 McCone Hall, Berkeley, CA 94720, USA. (thomas.bodin@berkeley.edu)

0148-0227/12/2012JB009547

sense. A well known limitation of these weighted average methods is that the solution strongly depends on the choice of a global norm. In other words, the number of free parameters in the solution (i.e. level of structure) has to be determined by the user in advance. Furthermore, these standard optimization schemes do not allow propagation of data uncertainties towards confidence limits in the surface [*Aster et al.*, 2005].

[6] One common problem in Earth sciences is that the level of information provided by observations is unevenly distributed geographically. Thus, the density of observations and the data noise are spatially variable. For example, in an airborne geophysical survey, data are collected along roughly parallel transects across the area of interest. In the inference process this suggests that the level of resolvable detail in the model will also vary spatially, however standard regression algorithms often involve only a few tuneable parameters, the selection of which is always a global compromise between data fit and model complexity. Inevitably the use of globally tuned damping parameters is likely to mean that sub-regions of the spatial domain may be under or over damped indicating that data information content has not been fully utilized. Regression techniques are the subject of much study, with literally hundreds of publications proposing new approaches to specific problems. For a literature review the reader is referred to text books on spatial data analysis [*Banerjee et al.*, 2004; *Kanevski et al.*, 2009; *Lloyd*, 2010, and references therein]. For a discussion on statistics in model inference, see *Mackenzie* [2004].

[7] In this study we propose an alternative approach to surface reconstruction by applying the reversible jump Markov chain Monte Carlo (rj-McMC) algorithm [*Geyer and Møller*, 1994; *Green*, 1995]. The rj-McMC is a fully non-linear stochastic parameter search scheme developed in the area of Bayesian statistics. It allows simultaneous inference on both model and parameter space, i.e., both the number of basis functions and the functions themselves are free to vary. With growing computational power in the last decade, this new class of sampling algorithm has been applied to a wide range of areas such as signal processing [*Andrieu and Doucet*, 1999], genetics [*Huelsenbeck et al.*, 2004], medical imaging [*Bertrand et al.*, 2001], image analysis [*Descombes et al.*, 2001], or computer vision [*Mayer*, 2008].

[8] In this short paper we show how the rj-McMC approach can be used for regression analysis in Earth sciences, and how it allows the complexity of the recovered surface to be spatially variable and directly determined by the data. Instead of seeking a best fitting model within an optimization framework, the full state of knowledge is represented in probabilistic terms, thus allowing inference on constraints, resolutions, and trade-offs. The method is presented and illustrated with an application to Moho depth reconstruction for Australia.

1.1. The Thickness of the Australian Crust

[9] The Moho discontinuity defines the base of the Earth's crust. It was first observed in 1909 by Mohorovičić, when he noticed that seismograms from shallow-focus earthquakes had two sets of P-waves and S-waves, one direct and one refracted back from a higher velocity medium. The crust-mantle boundary is therefore generally defined through a transition in the velocity of seismic waves. Above the Moho the velocity of P-waves is in the range 6.7–7.2 km s^{-1}, and below is 7.6–8.6 km s^{-1} corresponding to ultramafic materials.

[10] Modes of Moho topography are usually constructed by seismologists by interpolating compilations of local measures of Moho depth obtained from different types of seismic data. For a global model, see *Mooney et al.* [1998]. For examples of regional models, see *Marone et al.* [2003] for the Mediterranean, *Lloyd et al.* [2010] for South America, and *Kennett et al.* [2011] for Australia. A recent local model for Italy was constructed by *Di Stefano et al.* [2011].

[11] Improved knowledge of crustal thickness helps understanding of the geodynamical evolution of a continent. The thickness of the crust directly determines the rate at which heat is released to the Earth's surface, influences the location of earthquakes, and more generally defines the rules for plate tectonic processes. Maps of crustal thickness have numerous applications in geophysics. For example, in potential field studies, the crustal thickness (together with density) is needed to correct long-wavelength gravity data in order to infer lateral variations in mantle density. Crustal density and thickness are also used to calculate crustal isostasy [*Mooney et al.*, 1998].

[12] In this paper we show how to construct a probabilistic Moho topography model for Australia, exploiting the use the same data sets as *Kennett et al.* [2011]. The problem requires simultaneous inferences from different data types, characterized by different sensitivities to structure and different levels of noise. With multiple sources of data we usually would need to make somewhat arbitrary decisions about how to weigh the relative contributions of each data set to the final solution. However, using the rj-McMC formulation, we show the choices of parameterization, level of data fit and weighting between data types can be constrained by the data themselves, rather than having to be specified in advance.

2. Data

[13] The results compiled by *Kennett et al.* [2011] consist of an ensemble of data points providing the depth to the Moho at a large number of geographical locations across Australia (Figure 1). These local estimates of Moho depth do not represent direct measurements; they have been derived from a wide range of different seismological studies carried out in the last few decades (see *Kennett et al.* [2011, and references therein] for the complete description of the data used here). As can be clearly seen in Figure 1, although there is good coverage of the continent, the data sampling is non-uniform and also highly anisotropic in places. The poorest sampling occurs in remote areas where logistics make experiments difficult such as the Simpson and Great Sandy Deserts.

[14] The different sets of Moho estimates can be divided into 6 classes, based on the nature of the seismic observations employed: refraction (339 points), reflection (652 points), broad-band receiver functions (225 points), short-period receiver functions (41 points), and historical reflection studies (32 points). A seventh data set of Moho depths derived from gravity measurements (11 points) was also included to help define continent-ocean margins. We represent the full set of depth estimates by combining 7 vectors of different length: $\mathbf{d} = [\mathbf{d}_1, \mathbf{d}_2, \ldots \mathbf{d}_7]$, where each data set $\mathbf{d}_i$ is a vector

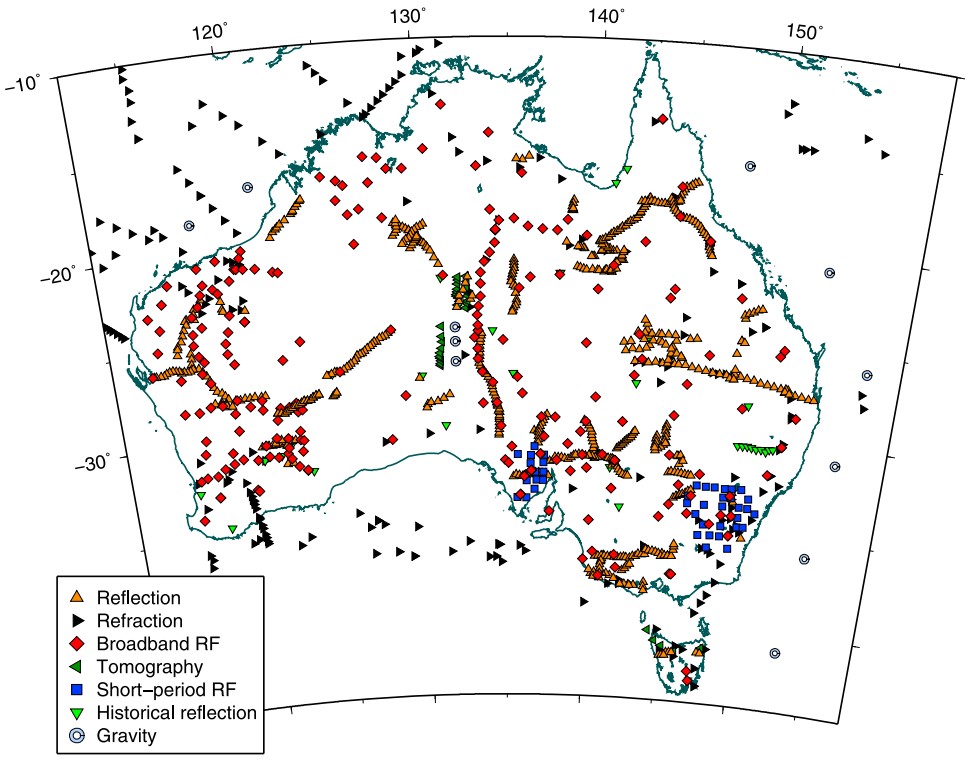

Figure 1. Location of Moho estimates used for the construction of the Moho surface from *Kennett et al.* [2011]. There are 7 classes of data with different levels of uncertainty, spatial sampling, and consistency.

of Moho estimates measured at a number of geographical locations with method *i.*

[15] Since the seismic wavetype, frequency band, and processing technique vary between the different classes of observations, the level of constraint on Moho depth differs between data types. The depth of Moho may be clear in certain experiments, such as reflection profiling, which are sensitive to sharp discontinuities; whereas in smoothed tomographic images the expression of Moho tends to be through a gradient in wavespeed. In order to keep consistency between the different classes of constraints, the Moho was defined by *Kennett et al.* [2011] as the boundary where the velocities on the lower side are greater than 7.8 km s^{-1} for P waves and 4.4 km s^{-1} for shear waves. Where the crust-mantle transition occurs as a gradient zone, the base of the transition was taken. The individual estimates of Moho depth have associated measures of quality. But, since the inverse methods used to interpret seismic records often provide just a single best fitting Earth model to a often highly non-unique and ill-posed problem, there is scant information about the true level of uncertainty on the Moho depths employed. Further, the relative uncertainties for results from different data types are not well constrained.

3. Regression With B-Spline Interpolation

[16] As a first step we reconstruct a Moho topography model in the same way as in *Kennett et al.* [2011]. First, a uniform grid of 0.5 deg × 0.5 deg cells is constructed and the data points are average over each cell. This preliminary smoothing process restricts the likelihood of close points with highly varying values. Second, average points are interpolated with a standard B-spline interpolating scheme [*Wessel and Smith*, 1998], where the user defines a priori a spline-tension parameter (see *Smith and Wessel* [1990] for details). Here all data-type are equally weighted in the minimization scheme, since there is no a priori information about relative quality between different experiments.

[17] In Figure 2 we show results for two different tension values. Figure 2 (left) might be over-complicated and may contain features due to data errors. Conversely, Figure 2 (right) is missing details present in the data. *Kennett et al.* [2011] use the spline tension shown in Figure 2 (left) but the weakness of this approach is that it relies on the judgment of the user to determine the appropriate level of smoothing. There is no clear quantitative way to choose the appropriate level of detail, or data fit. We acknowledge that a range of statistical techniques have been developed for judging whether the choice of the model dimension is warranted by the data, for example, the Bayesian information criterion [*Schwarz*, 1978], the Akaike information criterion [*Akaike*, 1974] or F-tests [*Aster et al.*, 2005]. However, such procedures rely on an accurate knowledge of data error statistics, which is not the case here, nor in a wide range of regression problems across the Earth sciences.

4. Transdimensional Regression

[18] In this section we show how the rj-McMC algorithm can be used to allow the number of free parameters in the representation of the surface of Moho depth to be a free parameter in the regression process. This situation is known as a transdimensional inversion, that is, one where the

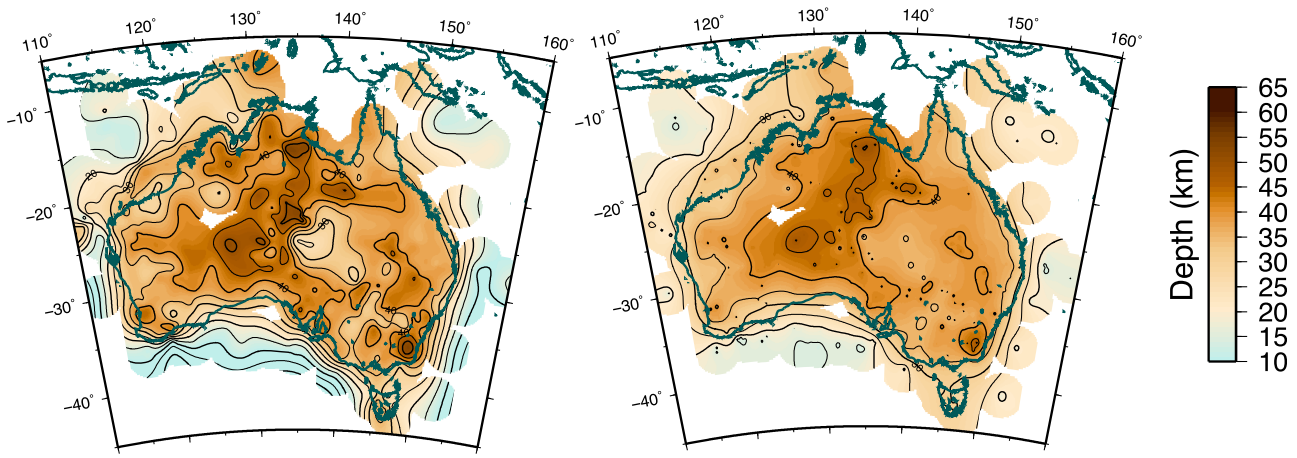

Figure 2. Moho depth surfaces constructed with a B-spline interpolation of averages over 5° × 5° cells, with two different spline tension values. A white mask is applied to all points that are further than 250 km from a data constraint. This illustration shows how an arbitrary user-defined global norm (here the minimum tension for B-splines) influences the solution in an standard linear optimization regression scheme. Without accurate knowledge of data uncertainties, there is no way to objectively discriminate between these two solutions.

dimension of the parameter space is itself variable [*Sisson*, 2005; *Sambridge et al.*, 2006].

4.1. Surface Parameterization

[19] The surface is parameterized with an irregular mesh consisting on a variable number of Voronoi cells [*Voronoi*, 1908; *Okabe et al.*, 1992] as shown in Figures 3 and 4. Although Voronoi cells seem complex structures, the mesh is uniquely controlled by a small number of nodes (blue squares in Figure 3). Any point inside a cell is closer to the node of that cell than any other node, so the shape of the parameterization is entirely defined by the location of nodes $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k]$. Boundaries between neighboring nodes are simply the perpendicular bisectors of the direct line between the center of the cells.

[20] The position $\mathbf{C}$ and number k of nodes are unknown variables to be directly inferred from the data. Here we use only the simplest possible representation of a surface within each cell, that is, a constant. This means that a single Moho depth parameter is assigned to each Voronoi cell $\mathbf{v} = [v_1, v_2, \ldots, v_k]$, yielding a surface made of piecewise constant polygons (Figure 4). Higher order polynomials are possible, for example, a linear gradient or quadratic, which would require additional unknowns for each cell. We expect this parameterization of the surface to self-adapt to the geometry of the problem. At a first glance this way of describing the surface seems coarse, as we only allow infinite gradients at boundaries. However, we shall show that in a probabilistic framework the expected model tends to be a continuous surface.

[21] The number of cells, and hence the fit to data becomes an unknown in the problem. It is worth noting that the data can be perfectly fitted by simply placing a Voronoi node at the location of each data point (i.e. with the number of cells equal to the number of data points). However, contrary to optimization schemes (where the goal is to minimize a misfit measure), the rj-McMC algorithm is able to automatically adjust the number of model parameters, in order to fit the data up to the level of data noise. For example, given a choice between a simple model with fewer cells and a more complex model that provides a similar fit to data, the simpler one will be favored. This is due to a fundamental property of Bayesian inference called 'natural parsimony' which gives preference for the least complex explanation for an observation (see *Malinverno* [2002] and *MacKay* [2003] for a discussion).

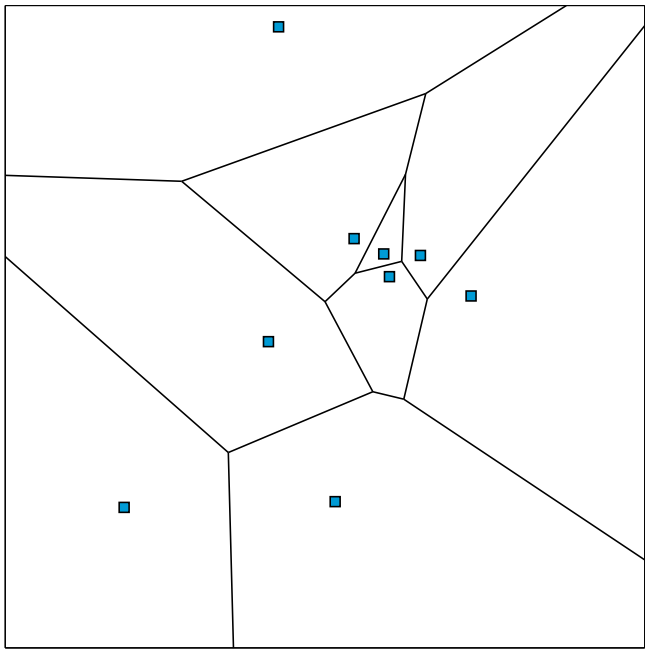

Figure 3. Construction of a surface with a Voronoi tessellation. The boundaries of each cell are defined by the perpendicular bisector of each pair of nodes. A constant surface value is assigned to each cell. As the number and position of the nodes changes the Voronoi diagram corresponds to a multi-scale parameterization of the surface.

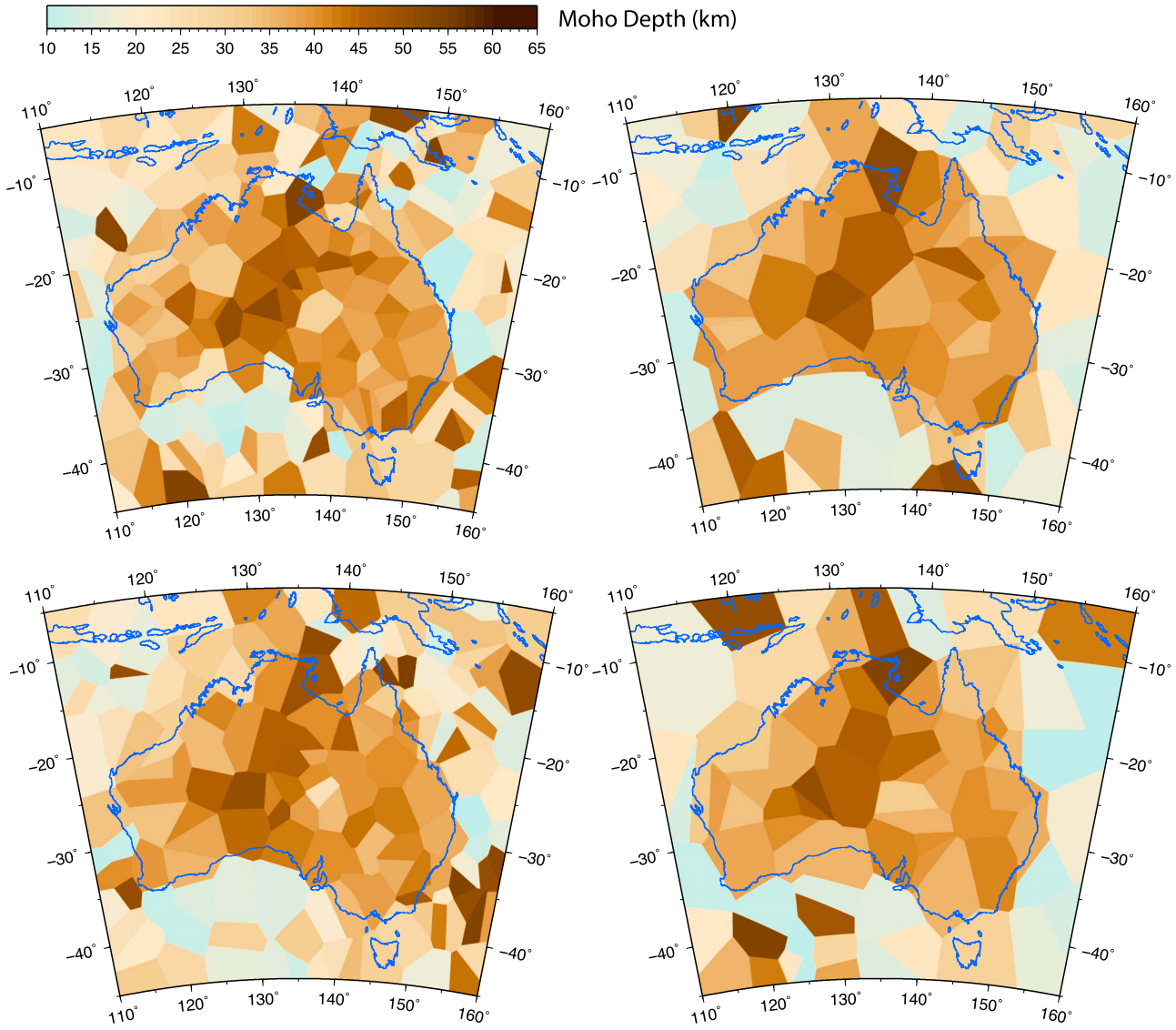

Figure 4. Four Voronoi models for Moho depth randomly drawn out of the ensemble solution provided by the reversible jump Markov chain Monte Carlo algorithm (rj-McMC). The full ensemble solution consists of 10^6 such models with variable complexity and parameterization. Each of these models $\mathbf{m}_i$ is associated with a set of noise estimates $\mathbf{h}_i$. The statistical distribution of these models is proportional to the posterior PDF, and hence meaningful statistical information can be extracted from the ensemble for interpretation of the nature of the Moho surface (see Figures 5–7).

In this way the level of structure in the solution is directly determined by estimates of the data errors.

[22] In our problem the uncertainty associated with each class of data is poorly constrained. This lack of knowledge on errors statistics can be accounted for in a probabilistic framework by using a hierarchical formulation [*Malinverno and Briggs*, 2004; *Malinverno and Parker*, 2006; *Dettmer et al.*, 2012] where the data noise is parameterized and also becomes an unknown to be inverted for in the regression problem. Here we assume an independent, random Gaussian noise with a different variance for each data type. That is, within a data type i, errors are assumed to be normally distributed with variance σ_i^2, and we define 7 parameters $\mathbf{h} = [\sigma_1, \sigma_2, \ldots, \sigma_7]$, each representing the level of uncertainty for a data class.

[23] We recognize that the Gaussian assumption may itself be questionable in some cases. Furthermore, by assuming a normal distribution has zero mean, we do not account for systematic errors. With this representation, the full set of unknown model parameters can be described by:

$$\mathbf{m} = [k, \mathbf{C}, \mathbf{v}, \mathbf{h}], \tag{1}$$

giving $3k + 8$ unknowns in total.

[24] Note that it is possible to include a priori information about data errors. For example one might have some indication about relative uncertainties between data points. In this case, weights can be applied to different measurements in the likelihood function, and the standard deviation of data errors σ_i can be written as proportional to the weighting

factor, with the constant of proportionality being the parameter to invert for [*Bodin et al.*, 2012a].

4.2. Probabilistic Bayesian Inference

[25] In a Bayesian framework the solution is a probability distribution representing the level of knowledge one has about model parameters, after combining a priori information with observations [*Tarantola and Valette*, 1982]. This formulation relies on Bayes' rule [e.g., *Bernardo et al.*, 1994] which, up to a constant of proportionality, can be written as

$$\text{posterior} \propto \text{likelihood} \times \text{prior} \tag{2}$$

$$p(\mathbf{m} \mid \mathbf{d}) \propto p(\mathbf{d} \mid \mathbf{m}) p(\mathbf{m}) \tag{3}$$

where $p(\mathbf{m}|\mathbf{d})$ is the posterior probability density function (PDF) of the model parameters $\mathbf{m}$, given the data vector, $\mathbf{d}$. $p(\mathbf{m})$ is the prior PDF on the model, which represents what we consider reasonable for the values of the model parameters. The prior information $p(\mathbf{m})$ used in this paper consists of bounded, uniform distributions with a range of parameter values chosen to represent physically reasonable limits, wide enough that the data dominate the posterior PDF (e.g., a uniform distribution for Moho depth between 10 and 50 km).

[26] The likelihood function, $p(\mathbf{d}|\mathbf{m})$, quantifies the probability of obtaining the data, $\mathbf{d}$, given the model, $\mathbf{m}$. This is a measure of the data fit and increases as the model fits the data better. Effectively the likelihood updates the prior information, transforming it to the posterior. If the prior and posterior distributions are the same, then we have learnt nothing from the data. Here we define it as the product of seven multivariate Gaussian distributions, one for each data type

$$p(\mathbf{d}|\mathbf{m}) = \prod_{i=1}^{7} \left[\frac{1}{\sqrt{(2\pi\sigma_i^2)^{n_i}}} \exp\left\{ \frac{-\|\mathbf{d}_i - g_i(\mathbf{m})\|^2}{2\sigma_i^2} \right\} \right] \tag{4}$$

where $g_i(\mathbf{m})$ is the vector of Moho depths estimated by the model $\mathbf{m}$ for data type $\mathbf{d}_i$ of size n_i. Note that the level of data uncertainty σ_i for a data type $\mathbf{d}_i$ determines the width of the probability distribution in this direction, and hence the importance given to this particular data set in the final solution. Therefore, by treating the different noise levels as unknowns, we implicitly let the weights between data types to be variable and directly determined by the data.

[27] By setting data errors to be unknown parameters in the problem, one would expect a tendency for high values to be preferred, since a larger σ_i will increase the likelihood by increasing the exponent in equation (4). However, this effect is counter-balanced by the normalizing constant of the Gaussian distribution which also contains σ_i in the denominator. Hence, when inverting for noise levels, the algorithm will be driven towards values which are a compromise between these competing forces. Here we define measurement errors to as all factors contributing to an inability to fit the data including theoretical errors due to parameterization.

4.3. The Reversible Jump Algorithm

[28] Since the unknown model is described with a spatially variable parameterization with variable number of cells, the regression problem is highly non-linear, and there is no analytical formula for the posterior PDF. Instead we use the rj-McMC algorithm [*Green*, 1995, 2003] to sample the posterior distribution. This algorithm is able to generate a collection of models $\mathbf{m}$, whose statistical distribution is proportional to the posterior PDF (Figure 4). The rj-McMC algorithm is a generalization of the well-known Metropolis-Hastings algorithm [*Metropolis et al.*, 1953; *Hastings*, 1970] to the case where the dimension of the solution space is variable. A sequence of models are generated in a chain, where typically each is a random perturbation of the last. Recent descriptions of the algorithm have been given by *Sambridge et al.* [2006] and *Gallagher et al.* [2009]. Examples of applications in Earth sciences can be found in *Malinverno* [2002], *Stephenson et al.* [2004, 2006], *Jasra et al.* [2006], *Bodin and Sambridge* [2009], *Hopcroft et al.* [2009], *Charvin et al.* [2009], *Gallagher et al.* [2011], *Dettmer et al.* [2010], *Luo* [2010], *Piana Agostinetti and Malinverno* [2010], *Bodin et al.* [2012b], *Minsley* [2011], *Dettmer et al.* [2012], and *Bodin et al.* [2012a].

[29] The implementation used here is identical to that described by *Bodin et al.* [2012a] applied to the joint tomographic inversion of different seismic data types with unknown noise levels. In this work, the same inversion scheme is used, but instead of fitting travel times of seismic rays, we fit regression points. Below we give a brief description of the algorithm, but refer the reader to *Bodin and Sambridge* [2009] and *Bodin et al.* [2012a] for a complete description of the algorithm.

[30] Having randomly initialized the model parameters $\mathbf{m} = [k, \mathbf{C}, \mathbf{v}, \mathbf{h}]$ by drawing values from the prior distribution of each parameter, the algorithm proceeds iteratively. Each step of the Markov chain is divided into two stages:

First, propose a new model by drawing from a probability distribution $q(\mathbf{m}'|\mathbf{m})$ such that the proposed model $\mathbf{m}'$ is conditional only on the current model $\mathbf{m}$. This involves one randomly selected type of change, with probability 1/5, out of five possible:

1. Change a depth value: Randomly pick one Voronoi cell (from a uniform distribution) and change the depth value assigned to this cell according to a Gaussian probability distribution $q(v_i'|v_i)$ centered at the current value v_i.

2. Move a Voronoi node: Randomly pick on Voronoi cell, and randomly perturb the position of its node according to a 2D Gaussian proposal probability density $q(\mathbf{c}_i'|\mathbf{c}_i)$ centered at the current position $\mathbf{c}_i$.

3. Birth: Create a new Voronoi cell by randomly drawing a point in the 2D map. A depth value needs to be assigned to the new cell. This is drawn from a Gaussian proposal probability centered on the current depth value where the birth takes place.

4. Death: Delete a Voronoi node chosen randomly from the current set of k cells.

5. Change one noise parameter: Randomly pick one component of the vector $\mathbf{h}$, and randomly perturb its value according to a Gaussian proposal probability density $q(h_i'|h_i)$ centered at the current value h_i.

Second, randomly accept or reject the proposed model (in terms of replacing the current model) with probability $\alpha(\mathbf{m}'|\mathbf{m})$ given by :

$$\alpha(\mathbf{m}'|\mathbf{m}) = \min\left[1, \frac{p(\mathbf{m}')}{p(\mathbf{m})} \times \frac{p(\mathbf{d}_{obs}|\mathbf{m}')}{p(\mathbf{d}_{obs}|\mathbf{m})} \times \frac{q(\mathbf{m}|\mathbf{m}')}{q(\mathbf{m}'|\mathbf{m})}\right] \tag{5}$$

[*Bodin and Sambridge*, 2009]. When the proposed model is rejected, the current model is retained for the next step and also added again to the output ensemble.

[31] The first part of the chain (called the burn-in period) is discarded, after which the random walk is assumed to be stationary and models generated by the chain are asymptotically distributed according to the posterior probability distribution $p(\mathbf{m}|\mathbf{d})$.

5. Appraising the Ensemble Solution

[32] The solution to the self-adaptive regression is represented by an ensemble of 10^6 Voronoi models $\mathbf{m} = [k,\mathbf{C},\mathbf{v},\mathbf{h}]$ with variable parameterizations (Figure 4), and noise estimates, distributed according to the posterior PDF. This large multivariate probability distribution can be visualized and interpreted by looking at the distribution of marginals, i.e., by projecting inferred quantities on a small set of meaningful variables. For example, at any geographical location, the ensemble of sampled Moho depths give the 1-dimensional posterior probability of Moho depth at this location.

[33] Thus one can extract at any point of the map, the posterior mean and standard deviation for Moho depth. In this way, an expected Moho surface as well as a point by point error map can be constructed by simply stacking individual models. We show in Figure 5 maps for the first four statistical moments (mean, standard deviation, skewness and kurtosis) for the Moho distribution. These are quantitative measures of the shape of the posterior solution at each "pixel" of the map. Note that when these posterior expectations are computed, models with variable geometries overlap providing continuous smooth maps that have an effective spatial resolution higher than any single model in the ensemble. This way, information is extracted from the ensemble as a whole, and the approach provides a parsimonious solution with no need for explicit smoothing.

[34] The first moment (Figure 5a) is the mean, and can be seen as the expected Moho surface. The square root of the second moment is the standard deviation (Figure 5b) and can be interpreted as an error map. As expected, surface errors are correlated with data coverage (Figure 1), although they also contain information about data consistency. The skewness in Figure 5c gives information about the asymmetry of the depth distribution at each point of the map. Qualitatively, a negative skew (green in Figure 5c) indicates that the tail on the shallower side of the probability density function is longer than the deep side and the bulk of the values lies shallower than the mean. The 4th moment or Kurtosis (Figure 5c) is a measure of 'peakedness' of the distribution (a high kurtosis indicates heavy tails). For this measure, higher kurtosis means more of the variance is the result of infrequent extreme deviations (heavy tails), as opposed to frequent modestly sized deviations. In our case, high kurtosis appears to be a an indicator of data control. When no data is available, the posterior is equal to the uniform prior distribution which has a low kurtosis. Low kurtosis is also an indicator of a bi-modal distribution, which happens at sharp discontinuities in the Moho surface (see Figure 7).

[35] Some aspects of the moment distribution appear to link directly to the availability of information. A particular case is in the offshore environment where in the absence of data the Moho tends to drift back to a value of 30 km (the center of the allowed uniform distribution). The effect could be reduced by introducing a spatially varying prescription of the prior distribution on the Moho. The largest values of the standard deviation for the ensemble occur where the data controls are weakest and takes very large values in the offshore zone where there is no data. With only weak constraints the Voronoi cellular representations have more freedom in their representation and this is reflected in a larger standard deviation.

[36] Changes in the skewness and kurtosis can be linked to the presence of dense directional data particularly from reflection profiling, e.g., in Western Australia around 120°E, and again near 20°S, 130°E. Skewness and kurtosis are higher order moments, and hence are mostly determined by values far from the mean (they are sensitive to outliers). They are sensitive to poorly constrained tails of the posterior, and also to the limits chosen for the prior distribution. Therefore, interpretation of these quantities should always be undertaken with caution, although here it would appear that they have information content.

[37] It is important to emphasize that, contrary to optimization schemes, here there is no unique solution for the surface, but rather the level of information is described probabilistically, and different statistical measures can be extracted. We illustrate this in Figure 6 where four measures of the characteristic value of the Moho distribution are plotted at each pixel, namely the arithmetic mean, the harmonic mean (i.e. the inverse of the average of the inverses), the median and the mode or maximum a posteriori (MAP). The arithmetic mean is appropriate to a purely Gaussian PDF and the other forms give different weighting to the outliers in the distribution. The median is a robust statistic that can be effectively applied to a wide range of PDFs. The map for the mode (maximum) in Figure 6d) has the most distinctive character, as it tends to preserve the sharp jumps in the underlying Voronoi models. It also preserves some significant information about the nature of the controls on the Moho distribution.

[38] We show in Figure 7 the complete posterior PDF at locations along two sections of constant longitude ϕ = 130 deg and constant latitude Θ=−21 deg (red lines in Figure 6a) For each point on the cross-section, the entire posterior marginal is plotted in a color scale. White dashed lines show the 95% credible interval. These 2D probability marginals seem 'patchy', and one can easily recognize the underlying parameterization with constant cells. However, the red line following the mean of the Moho depth distribution at each geographical location is smooth and does not exhibit obvious artifacts of the parameterization. It is clear that this plot of the posterior PDF contains much more information than is given by the average map in Figure 6a. It is interesting to see that in some areas, the marginal posterior is far from being a normal distribution. In this case, the mean and the standard deviation are not so representative of the nature of the function.

[39] As an example, Figure 7 (bottom) shows the probability distribution for Moho depth where the two cross-sections intersect. At this particular location, it is striking how the distribution is bi-modal. The mean of the distribution falls

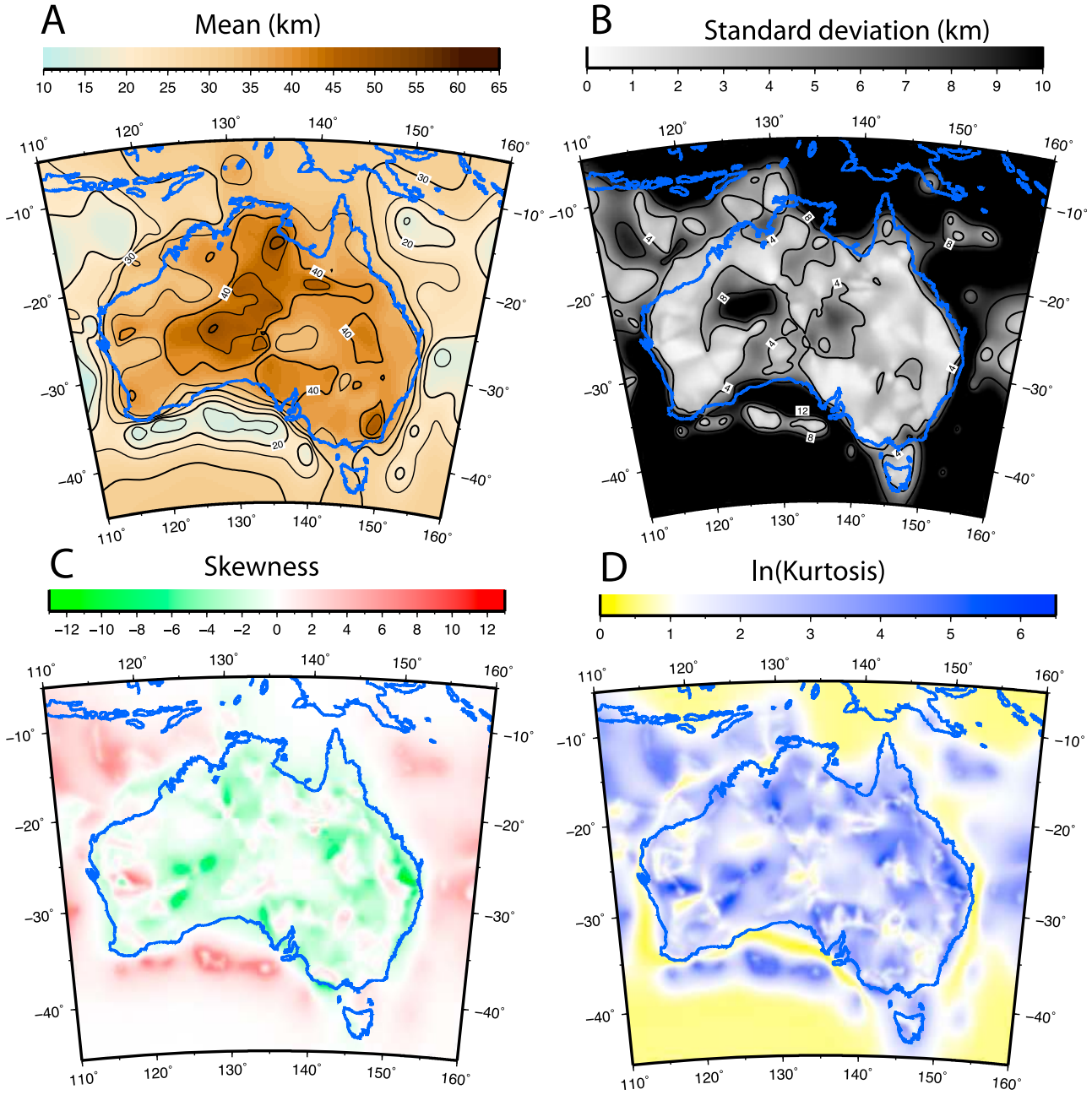

Figure 5. Maps of statistical moments extracted at each point of the 2D field of Moho depth. (a) Average map simply obtained by stacking all Voronoi models in the ensemble solution. This map can be interpreted as the expected Moho surface. (b) Map of standard deviation than can be seen as an error map for the expected surface. (c) The skewness map shows the asymmetry of the distribution at each points. Points with a symmetric distribution (0 Skewness) are shown in white. (d) Kurtosis map describing the"peakedness" of the distribution. Distributions less 'spiky' than a Gaussian are in yellow, 'spikier' than a Gaussian are in blue, and Gaussian-like are in white.

between the two modes and has a very low predictive power, and hence the maximum of the distribution (Figure 6d) might be better suited to describe the solution. In fact, in the presence of spatial discontinuities in data due to a sharp topological gradient in the Moho, the Voronoi cells alternatively take the depth values of each side of the topographic transition. This results in having a marginal posterior that has 2 modes that lie at the depths on each side of the topographic discontinuity. Another potential cause of bi-modality could be inconsistency between data points from different sources, where the solution is jumping between them to alternatively try to maximize the likelihood.

[40] We can also make inferences about the level of errors for each of the seven data types through the noise parameters $\mathbf{h} = [\sigma_1, \sigma_2, \ldots, \sigma_7]$. The posterior solution for each standard deviations $p(\sigma_i|\mathbf{d})$ is given by a 1D non-parametric probability

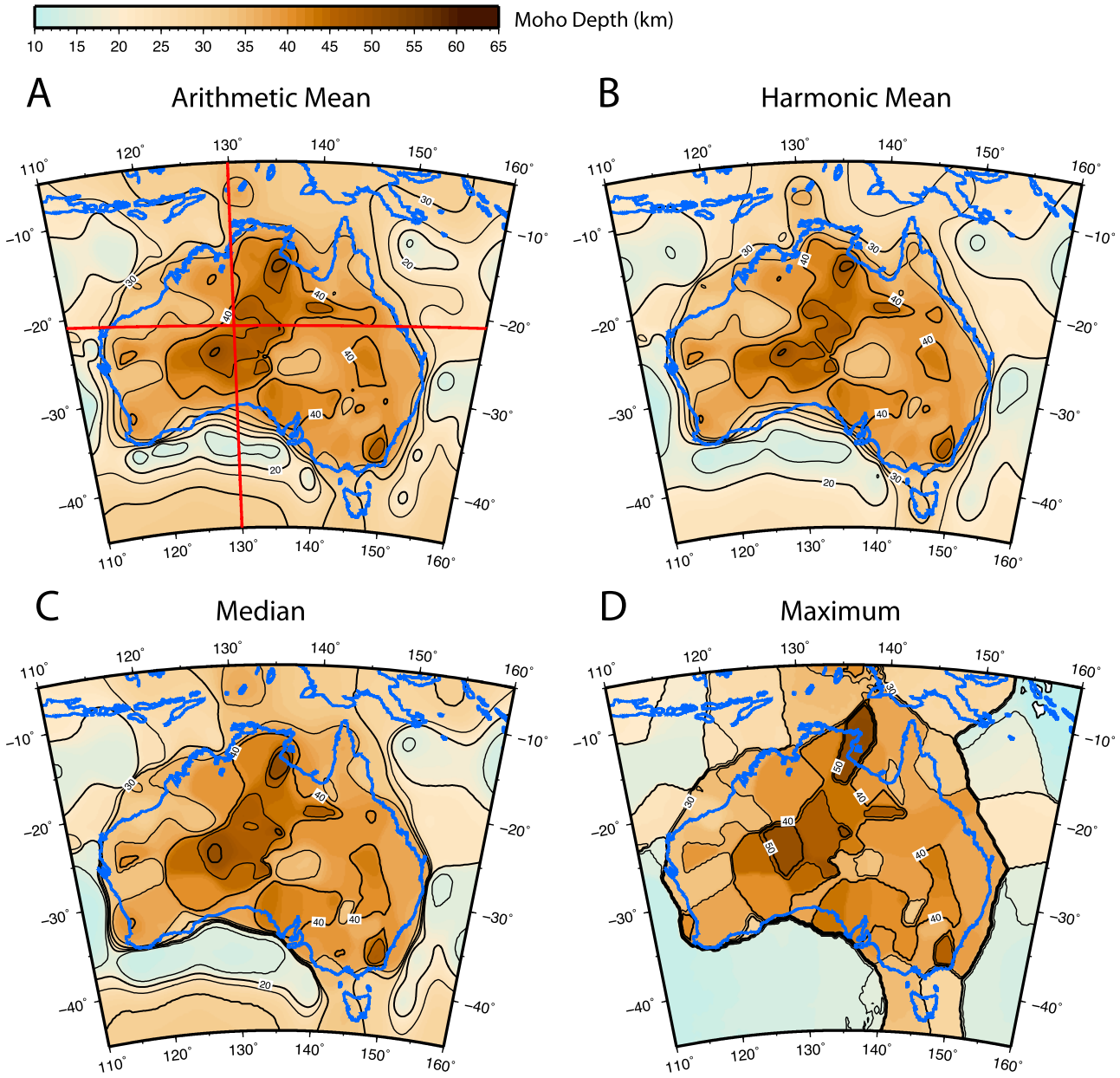

Figure 6. Different measures of the characteristic value of the Moho depth at a point can be extracted from the ensemble solution, and the corresponding maps can be used for interpretation. (a) The arithmetic mean is the standard average measure (equivalent to Figure 5a). The two red lines show the cross-sections where the full posterior solution is shown in Figure 7. (b) The harmonic mean is the inverse of the averages of inverses. (c) Median map. (d) Maximum map following the maximum (or mode) of the distribution at each point.

density function plotted in Figure 8. Again, these distributions are simply obtained from the histogram of sampled values of noise parameters during the search algorithm. The differences of uncertainties between data types can be explained by the nature of the different processing schemes used to interpret the seismic wave-field in terms of Moho depths. The smaller standard deviation for reflection profiling is probably linked to the relative close spatial sampling (Figure 1), which yields a high data consistency, and also to the fact that noisy records have already been discarded (there has to be something to pick before a value is declared). The gravity spread is also linked to the wide geographic dispersal.

[41] One difficulty in the interpretation of the PDF's for the different data classes is that the same method can have been applied in regions with very different Moho depth and character. Thus, the results for tomography mix two distinct data sets, one from central Australia where the Moho is deep and the other from Tasmania where the Moho is quite

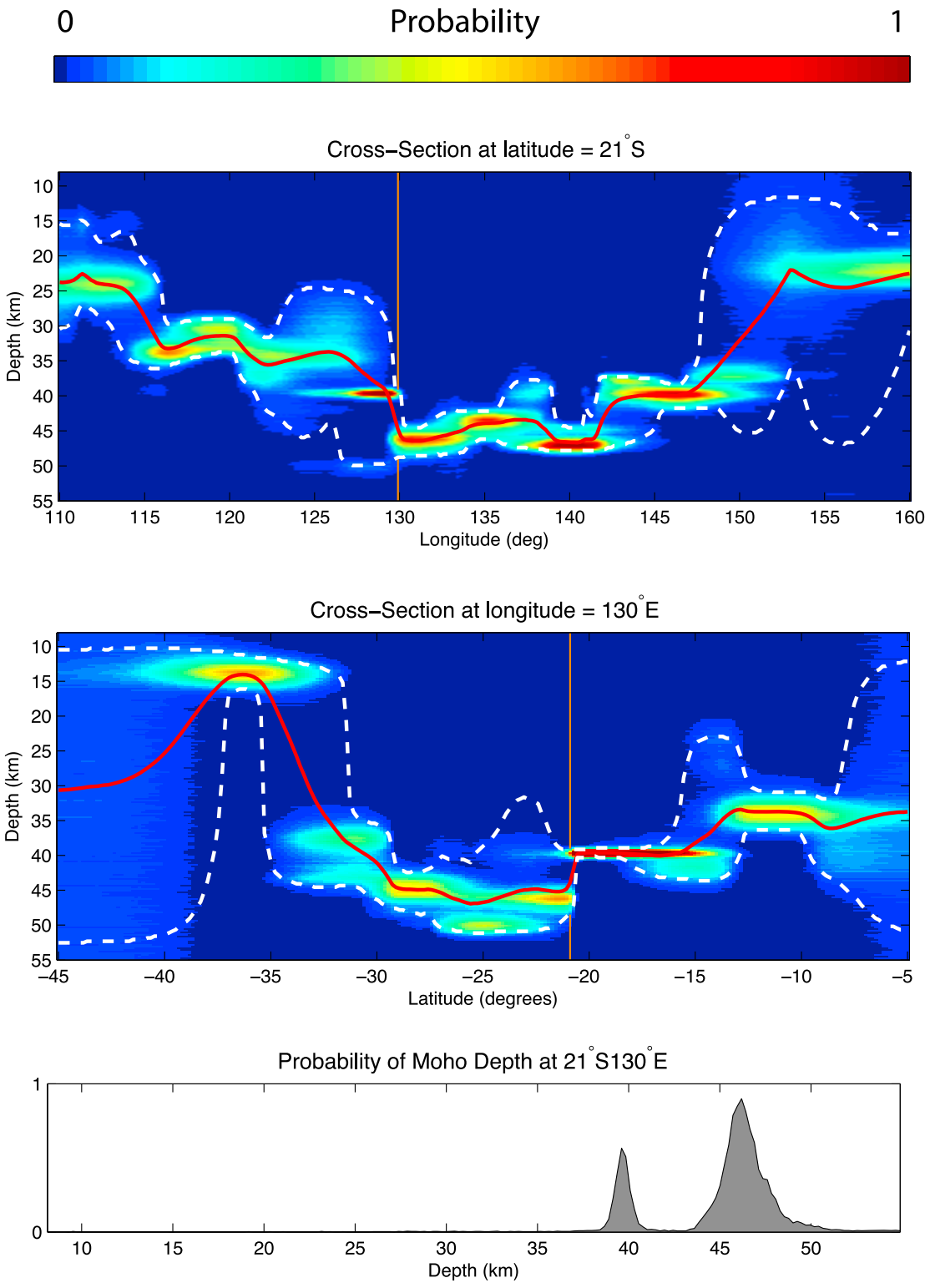

Figure 7. (top and middle) Full posterior probability distribution for Moho depth along the lines shown in red in Figure 6a (red is high probability and blue is low). White dashed lines show the 95% credible interval, and red lines follow the mean of the distribution at each location. (bottom) The full probability distribution where the 2 lines intersect.

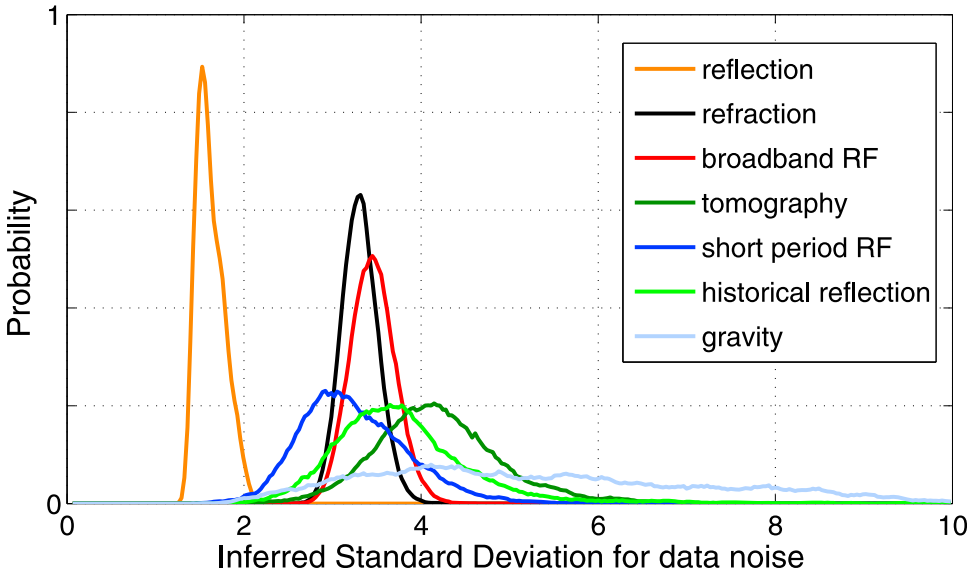

Figure 8. Posterior probability distribution for the standard deviation of errors associated with each class of Moho estimate: $\mathbf{h} = [\sigma_1, \sigma_2, \ldots, \sigma_7]$.

shallow, and so the broad distribution is not surprising. Similarly the historical reflection results are point measurements scattered across the continent so little coherence is to be expected.

[42] There is an interesting difference in the nature of the PDF for the two different classes of receiver function information. The high-frequency receiver functions show a skewed distribution with a long tail to large values, even though they are concentrated in southeastern Australia. This approach will perform well where the Moho is sharp, but will be less effective where the Moho is transitional and this may well lead to estimates that are less consistent with other information. With a full broad-band signal gradients in the crust-mantle transition are better recovered and there is a much more clustered PDF.

[43] How does the result for the self adaptive approach compare with that published by *Kennett et al.* [2011]? We can get a good idea by comparing Figures 2a and 5a, since the B-spline interpolation in Figure 2a is carried out using the same spline tension values as in the *Kennett et al.* [2011] study. The self-adaptive surface is smoother, but displays the same long wavelength distribution of deeper Moho. The rapid transition from thick to relatively shallow Moho near 135°E in southern central Australia is preserved with a very similar geometry. Similarly the two promontories of thicker crust into the west of Western Australia have a similar configuration. The merits of the more sophisticated procedure are that we have a much more complete characterization of the properties of the Moho distribution across the continent, particularly with respect to reliability. In addition we gain insight into the statistical properties of the different classes of data constraints.

[44] Furthermore, one important result of this method is the ability to recover sharp discontinuities in surfaces. Figure 7 illustrates one such discontinuity in continental Australia. At 21 S 130 E there is a ~8 km step in the Moho. This region is Proterozoic in origin implying that Moho discontinuities may have considerable longevity.

6. Discussion

[45] The approach developed in this work can be viewed as the combination of two separate and independent procedures. The first is transdimensional Bayesian inference, where the problem is to produce a large number of models with different parameterizations and variable complexities that describe the posterior PDF. This model distribution represents the complete solution of the inverse problem from the Bayesian viewpoint, in that it contains all the available information and uncertainty about the Earth model, as well as the correlation between parameters (not shown in this paper). However, the final goal of regression is to produce an interpretable solution, and a single Earth surface is required for practitioners who are not familiar with Bayesian methods. This is a reason why optimization methods are often preferred by Earth scientists (the true Earth is unique after all). As shown in Figure 4, the individual members of the ensemble do not have a simple character, and cannot be directly interpreted.

[46] Hence, the second part of the procedure consists in extracting interpretable information (e.g., a solution surface and an error map) from this ensemble of models, that is to construct a comprehensive view of the Bayesian solution. In Bayesian studies, the standard way of extracting information from the ensemble is to use marginal and conditional distributions on model parameters [*Box and Tiao*, 1973; *Sivia*, 1996]. To obtain a 'solution model' for analysis purposes, the Voronoi models are simply averaged, i.e., a solution surface is constructed by taking the mean of the distribution of values at each point across the Earth model. Instead of the mean value, other possibilities include taking the median or the mode of the distribution of values at each pixel (Figure 6).

[47] There is a relative freedom in the design of solution schemes. Different choices may lead to different Earth models, and hence to different interpretations. Therefore it can be argued that there is an inherent contradiction in the proposed method, as the initial philosophy is to remove any subjective choices made at the outset (e.g. number of free parameters, level of smoothing). Indeed, as we introduce parameters such as the number of cells and produce a parsimonious ensemble solution that accounts for all states of uncertainty, the geometry of model space becomes very complex, and then arbitrary projections of the posterior PDF need to be chosen for interpretation. It is important to emphasize that in a Bayesian formulation, the only true solution to the problem is the posterior distribution, and any single model derived from the ensemble must be seen by interpreters just as a projection of the posterior solution.

[48] Furthermore, it must be noted that we gain our extra information on the nature of the Moho distribution at a high computational price. If the true surface is too complex, the number of Voronoi models needed to sample the posterior distribution becomes colossal. Since the predicted data have to be computed each time a Voronoi model is proposed, the algorithm may become computationally expensive. In section 3, we have compared our results to solutions obtained with a conventional regression scheme, and showed the advantages of transdimensional sampling. However we did not compare computational times, and here it is necessary to recognize that, even with a parallelized and optimized code, the method used here is between 1 and 3 orders of magnitude slower than standard linear regression. Yet, with available computing power, it now becomes possible to handle such a problem using modern McMC techniques, a feat which would have been infeasible even 10 years ago.

[49] As we have demonstrated above, the self-adaptive parameterization procedure and the ensemble of solutions interpreted as a Bayesian posterior PDF, provide significantly more information about the nature of the Moho surface than is obtained in any form of simple surface fitting. *Kennett et al.* [2011] were forced to use the variability in nearby data estimates to provide a measure of the reliability of their Moho surface. In contrast the ensemble approach can give direct estimates and even recognize the likelihood of local discontinuities that cannot be incorporated in any scheme that forces a fit to a smooth surface.

[50] **Acknowledgments.** We thank Kerry Gallagher for discussions on Hierarchical Bayes method. T.B. wishes to acknowledge support from the Miller Institute for Basic Research at the University of California, Berkeley. Aspects of this study were supported by Australian Research Council Discovery project grant DP110102098. Calculations were performed on the TerraWulf II, a computational facility supported through the AuScope inversion laboratory. AuScope Ltd is funded under the National Collaborative Research Infrastructure Strategy (NCRIS) and the Education Infrastructure Fund (EIF), both Australian Commonwealth Government programmes. Software for 2-D probabilistic surface reconstruction is available from the AuScope inversion laboratory via the authors. We would like to thank the members of the AusMoho Working group whose efforts produced the Moho depth compilation exploited in this work.

References

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Trans. Autom. Control*, *19*(6), 716–723.

Andrieu, C., and A. Doucet (1999), Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC, *IEEE Trans. Signal Process.*, *47*(10), 2667–2676.

Aster, R., B. Borchers, and C. Thurber (2005), *Parameter Estimation and Inverse Problems*, Elsevier, Burlington, Mass.

Atkinson, P., and C. Lloyd (2007), Non-stationary variogram models for geostatistical sampling optimisation: An empirical investigation using elevation data, *Comput. Geosci.*, *33*(10), 1285–1300.

Banerjee, S., B. Carlin, and A. Gelfand (2004), *Hierarchical Modeling and Analysis for Spatial Data*, vol. 101, Chapman and Hall, Boca Raton, Fla.

Bernardo, J., A. Smith, and M. Berliner (1994), *Bayesian Theory*, vol. 62, John Wiley New York.

Bertrand, C., M. Ohmi, R. Suzuki, and H. Kado (2001), A probabilistic solution to the MEG inverse problem via MCMC methods: The reversible jump and parallel tempering algorithms, *IEEE Trans. Biomed. Eng.*, *48*(5), 533–542.

Billings, S., R. Beatson, and G. Newsam (2002), Interpolation of geophysical data using continuous global surfaces, *Geophysics*, *67*(6), 1810–1822.

Bodin, T., and M. Sambridge (2009), Seismic tomography with the reversible jump algorithm, *Geophys. J. Int.*, *178*(3), 1411–1436.

Bodin, T., M. Sambridge, N. Rawlinson, and P. Arroucau (2012a), Transdimensional tomography with unknown data noise, *Geophys. J. Int.*, *189*, 1536–1556, doi:10.1111/j.1365-246X.2012.05414.x.

Bodin, T., M. Sambridge, H. Tkalčić, P. Arroucau, and K. Gallagher (2012b), Transdimensional inversion of receiver functions and surface wave dispersion, *J. Geophys. Res.*, *117*, B02301, doi:10.1029/2011JB008560.

Box, G., and G. Tiao (1973), *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, Mass.

Burton, K., A. Gannoun, and I. Parkinson (2010), Climate driven glacial-interglacial variations in the osmium isotope composition of seawater recorded by planktic foraminifera, *Earth Planet. Sci. Lett.*, *295*(1), 58–68.

Charvin, K., G. Hampson, K. Gallagher, and R. Labourdette (2009), A Bayesian approach to inverse modelling of stratigraphy, part 2: Validation tests, *Basin Res.*, *21*(1), 27–45.

Descombes, X., R. Stoica, L. Garcin, and J. Zerubia (2001), A RHMCMC algorithm for object processes in image processing, *Monte Carlo Methods Appl.*, *7*(1–2), 149–156.

Dettmer, J., S. Dosso, and C. Holland (2010), Trans-dimensional geoacoustic inversion, *J. Acoust. Soc. Am.*, *128*, 3393–3405.

Dettmer, J., S. Molnar, G. Steininger, S. Dosso, and J. Cassidy (2012), Trans-dimensional inversion of microtremor array dispersion data with hierarchical autoregressive error models, *Geophys. J. Int.*, *188*, 719–734.

Di Stefano, R., I. Bianchi, M. G. Ciaccio, G. Carrara, and E. Kissling (2011), Three-dimensional Moho topography in Italy: New constraints from receiver functions and controlled source seismology, *Geochem. Geophys. Geosyst.*, *12*, Q09006, doi:10.1029/2011GC003649.

El Abbas, T., C. Jallouli, Y. Albouy, and M. Diament (1990), A comparison of surface fitting algorithms for geophysical data, *Terra Nova*, *2*(5), 467–475.

Gallagher, K., K. Charvin, S. Nielsen, M. Sambridge, and J. Stephenson (2009), Markov chain Monte Carlo (MCMC) sampling methods to determine optimal models, model resolution and model choice for Earth science problems, *Mar. Pet. Geol.*, *26*(4), 525–535.

Gallagher, K., T. Bodin, M. Sambridge, D. Weiss, M. Kylander, and D. Large (2011), Inference of abrupt changes in noisy geochemical records using transdimensional changepoint models, *Earth Planet. Sci. Lett.*, *311*, 182–194.

Geyer, C., and J. Møller (1994), Simulation procedures and likelihood inference for spatial point processes, *Scand. J. Stat.*, *21*, 359–373.

Green, P. (1995), Reversible jump MCMC computation and Bayesian model selection, *Biometrika*, *82*, 711–732.

Green, P. (2003), Trans-dimensional Markov chain Monte Carlo, *Highly Struct. Stochastic Syst.*, *27*, 179–198.

Hastings, W. (1970), Monte Carlo simulation methods using Markov chains and their applications, *Biometrika*, *57*, 97–109.

Hopcroft, P., K. Gallagher, and C. Pain (2009), A Bayesian partition modelling approach to resolve spatial variability in climate records from borehole temperature inversion, *Geophys. J. Int.*, *178*(2), 651–666.

Huelsenbeck, J., B. Larget, and M. Alfaro (2004), Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo, *Mol. Biol. Evol.*, *21*(6), 1123–1133.

Jasra, A., D. Stephens, K. Gallagher, and C. Holmes (2006), Bayesian mixture modelling in geochronology via Markov chain Monte Carlo, *Math. Geol.*, *38*(3), 269–300.

Kanevski, M., A. Pozdnoukhov, and V. Timonin (2009), *Machine Learning for Spatial Environmental Data: Theory, Applications and Software*, EFPL Press, Lausanne, Switzerland.

Kennett, B., M. Salmon, E. Saygin, and AusMoho Working Group (2011), AusMoho: The variation of Moho depth in Australia, *Geophys. J. Int.*, *187*, 946–958.

Kylander, M., J. Klaminder, R. Bindler, and D. Weiss (2010), Natural lead isotope variations in the atmosphere, *Earth Planet. Sci. Lett.*, *290*(1–2), 44–53.

Large, D., et al. (2009), The influence of climate, hydrology and permafrost on Holocene peat accumulation at 3500 m on the eastern Qinghai-Tibetan plateau, *Quat. Sci. Rev.*, *28*(27–28), 3303–3314.

Lloyd, C. (2010), *Local Models for Spatial Analysis*, CRC Press, Boca Raton, Fla.

Lloyd, S., S. van der Lee, G. França, M. Assumpção, and M. Feng (2010), Moho map of South America from receiver functions and surface waves, *J. Geophys. Res.*, *115*, B11315, doi:10.1029/2009JB006829.

Luo, X. (2010), Constraining the shape of a gravity anomalous body using reversible jump Markov chain Monte Carlo, *Geophys. J. Int.*, *180*(3), 1067–1079.

MacKay, D. (2003), *Information Theory, Inference, and Learning Algorithms*, Cambridge Univ. Press, Cambridge, U. K.

Mackenzie, D. (2004), Vital statistics, *New Sci.*, *2453*, 36.

Malinverno, A. (2002), Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem, *Geophys. J. Int.*, *151*(3), 675–688.

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

Malinverno, A., and R. Parker (2006), Two ways to quantify uncertainty in geophysical inverse problems, *Geophysics*, *71*, W15–W27.

Marone, F., M. Van Der Meijde, S. Van Der Lee, and D. Giardini (2003), Joint inversion of local, regional and teleseismic data for crustal thickness in the Eurasia-Africa plate boundary region, *Geophys. J. Int.*, *154*(2), 499–514.

Mayer, H. (2008), Object extraction in photogrammetric computer vision, *ISPRS J. Photogramm. Remote Sens.*, *63*(2), 213–222.

Metropolis, N., et al. (1953), Equations of state calculations by fast computational machine, *J. Chem. Phys.*, *21*(6), 1087–1091.

Minsley, B. (2011), A trans-dimensional Bayesian Markov chain Monte Carlo algorithm for model assesment using frequency-domain electromagnetic data, *Geophys. J. Int.*, *187*, 252–272.

Mooney, W., G. Laske, and T. Masters (1998), Crust 5.1: A global crustal model at 5° × 5°, *J. Geophys. Res.*, *103*(B1), 727–747.

Okabe, A., B. Boots, and K. Sugihara (1992), *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, John Wiley, New York.

Piana Agostinetti, N., and A. Malinverno (2010), Receiver function inversion by trans-dimensional Monte Carlo sampling, *Geophys. J. Int.*, *181*(2), 858–872.

Sambridge, M., K. Gallagher, A. Jackson, and P. Rickwood (2006), Trans-dimensional inverse problems, model comparison and the evidence, *Geophys. J. Int.*, *167*(2), 528–542.

Sandwell, D. T., and W. H. F. Smith (1997), Marine gravity anomaly from Geosat and ERS 1 satellite altimetry, *J. Geophys. Res.*, *102*, 10,039–10,054.

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Stat.*, *6*(2), 461–464.

Sisson, S. (2005), Transdimensional Markov chains: A decade of progress and future perspectives, *J. Am. Stat. Assoc.*, *100*(471), 1077–1090.

Sivia, D. (1996), *Data Analysis: A Bayesian Tutorial*, Oxford Univ. Press, New York.

Smith, W., and P. Wessel (1990), Gridding with continuous curvature splines in tension, *Geophysics*, *55*(3), 293–305.

Stein, M. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, Springer, New York.

Stephenson, J., K. Gallagher, and C. Holmes (2004), Beyond kriging: Dealing with discontinuous spatial data fields using adaptive prior information and Bayesian partition modelling, *Geol. Soc. Spec. Publ.*, *239*(1), 195–209.

Stephenson, J., K. Gallagher, and C. Holmes (2006), Low temperature thermochronology and strategies for multiple samples 2: Partition modelling for 2D/3D distributions with discontinuities, *Earth Planet. Sci. Lett.*, *241*(3–4), 557–570.

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

Voronoi, G. (1908), Nouvelles applications des parametres continus a la theorie des formes quadratiques, *J. Reine Angew. Math.*, *134*, 198–287.

Wessel, P., and W. Smith (1998), New, improved version of generic mapping tools released, *Eos Trans. AGU*, *79*, 579–579.