Transdimensional inversion of receiver functions and surface wave dispersion

Thomas Bodin, Malcolm Sambridge, Hrvoje Tkalcic, P. Arroucau, Kerry Gallagher, N. Rawlinson

Transdimensional inversion of receiver functions and surface wave dispersion

T. Bodin,[1] M. Sambridge,[1] H. Tkalčić,[1] P. Arroucau,[2] K. Gallagher,[3] and N. Rawlinson[1]

Received 5 June 2011; revised 17 November 2011; accepted 18 November 2011; published 3 February 2012.

[1] We present a novel method for joint inversion of receiver functions and surface wave dispersion data, using a transdimensional Bayesian formulation. This class of algorithm treats the number of model parameters (e.g. number of layers) as an unknown in the problem. The dimension of the model space is variable and a Markov chain Monte Carlo (McMC) scheme is used to provide a parsimonious solution that fully quantifies the degree of knowledge one has about seismic structure (i.e constraints on the model, resolution, and trade-offs). The level of data noise (i.e. the covariance matrix of data errors) effectively controls the information recoverable from the data and here it naturally determines the complexity of the model (i.e. the number of model parameters). However, it is often difficult to quantify the data noise appropriately, particularly in the case of seismic waveform inversion where data errors are correlated. Here we address the issue of noise estimation using an extended Hierarchical Bayesian formulation, which allows both the variance and covariance of data noise to be treated as unknowns in the inversion. In this way it is possible to let the data infer the appropriate level of data fit. In the context of joint inversions, assessment of uncertainty for different data types becomes crucial in the evaluation of the misfit function. We show that the Hierarchical Bayes procedure is a powerful tool in this situation, because it is able to evaluate the level of information brought by different data types in the misfit, thus removing the arbitrary choice of weighting factors. After illustrating the method with synthetic tests, a real data application is shown where teleseismic receiver functions and ambient noise surface wave dispersion measurements from the WOMBAT array (South-East Australia) are jointly inverted to provide a probabilistic 1D model of shear-wave velocity beneath a given station.

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

1. Introduction

[2] The coda of teleseismic P-waves contains a large number of direct and reverberated phases generated at interfaces beneath the receiver that contain a significant amount of information on seismic structure. However, these phases are difficult to identify as they are buried in micro-seismic and signal-generated noise [*Lombardi*, 2007]. The signal to noise ratio is usually improved by stacking seismograms from different records from a single station, but a major drawback is the introduction of different source time functions generated by multiple earthquakes. This problem is overcome by a method developed in the 1970s now widely used in seismology and follows the pioneering work by *Phinney* [1964]. The idea is to deconvolve the vertical component from the horizontal components to produce a time series called a "receiver function" (RF) [*Langston*, 1979]. In a receiver function the influence of source and distant path effects are eliminated, and hence one can enhance conversions from P to S generated at boundaries beneath the recording site.

[3] The RF waveform can be inverted in the time domain for a 1D S-wave velocity model of the crust and uppermost mantle beneath the receiver. In this paper we present a novel RF inversion methodology where the number of layers defining the velocity model as well as the variance and correlation of data noise are treated as unknowns in the problem. We also show how independent data of different character and with different sensitivities (e.g. surface wave dispersion measurements) can be included in a consistent manner. The result is a general probabilistic joint inversion methodology, where no explicit "tuning" is needed to weight different data sets.

1.1. A Brief History of RF Inversion

[4] The RF inverse problem is highly non-linear and non-unique [*Ammon et al.*, 1990]. *Owens et al.* [1984] carried out an iterative linearized inversion where partial derivatives were computed numerically with a finite difference scheme. The inversion was stabilized with truncation of small eigenvalues

[1]Research School of Earth Sciences, Australian National University, Canberra, ACT, Australia.
[2]Environmental, Earth and Geospatial Sciences, North Carolina Central University, Durham, North Carolina, USA.
[3]Géosciences Rennes, Université de Rennes 1, Rennes, France.

0148-0227/12/2011JB008560

after singular value decomposition of the system of equations. *Kosarev et al.* [1993] and *Kind et al.* [1995] used a linearized Tikhonov inversion and stabilized the algorithm by penalizing solutions far from a given reference model. It is well known that linear inversion procedures based on partial derivatives are easily trapped by local minima, and hence solutions may be strongly dependent on initial models.

[5] As increased computational power became available, Monte Carlo parameter search methods became a practical alternative for RF inversion. These include global optimization techniques such as genetic algorithms [*Shibutani et al.*, 1996; *Levin and Park*, 1997; *Clitheroe et al.*, 2000; *Chang et al.*, 2004], niching genetic algorithm [*Lawrence and Wiens*, 2004], simulated annealing [*Vinnik et al.*, 2004, 2006], very fast simulated annealing [*Zhao et al.*, 1996], and also the neighborhood algorithm [*Sambridge*, 1999a; *Bannister et al.*, 2003; *Nicholson et al.*, 2005]. These techniques are able to efficiently search a large multidimensional model space and provide complex Earth models that minimize a misfit measure without need of linearization or computation of derivatives. They were applied in the hope that the solution avoided entrapment in local minima of the objective function.

[6] The inherent non-uniqueness of RF inversion means that two Earth models that are far apart in the model space (i.e which have different parameter values) can provide a similar level of data fit. Non-linear optimization algorithms, to varying degrees, are able to search a large parameter space and find global minima, however they usually only provide a single solution, i.e. the best one in some sense. This leaves open the possibility that other Earth models, which are far from this solution, might also fit the data within errors. Hence a single solution is often not representative of the information contained in the data. To reduce dependence on single "best fit" models, global optimization techniques have been used to perform an ensemble inference, where one obtains an ensemble of models satisfying some predefined criteria (e.g. the best 1000 data fitting models generated by the algorithm). This ensemble of "acceptable" models are thus plotted together for visualization [e.g., *Piana Agostinetti et al.*, 2002; *Reading et al.*, 2003; *Hetényi and Bus*, 2007].

[7] However, the ensemble obtained in this way is rather arbitrary and there is no guarantee that these models are representative of all acceptable models. Furthermore, the statistical distribution of models within the ensemble generally does not represent the acceptable range in the objective function and therefore cannot be directly used to infer trade-off, constraints or resolution on model parameters. These issues arise because most non-linear optimization algorithms do not perform importance sampling (i.e where the frequency distribution of sampled models is proportional to the objective function, or posterior distribution in a Bayesian framework), and hence the ensemble solution strongly depends on user choices, or on the class of algorithm employed.

[8] A typical example has been the use of the neighborhood algorithm [*Sambridge*, 1999a] for RF inversion [e.g., *Piana Agostinetti et al.*, 2002; *Reading et al.*, 2003; *Bannister et al.*, 2003; *Frederiksen et al.*, 2003; *Hetényi and Bus*, 2007]. In a second paper, *Sambridge* [1999b] invoked the Bayesian philosophy and showed how to calculate standard Bayesian outputs using an arbitrary distributed ensemble, i.e. one generated by any ensemble technique. However, most studies which employ the neighborhood algorithm do so only in an optimization context.

[9] In a Bayesian framework the objective is to create an ensemble of models that represent the posterior probability distribution quantifying the degree of belief we have about the Earth's structure and composition. This probability distribution combines "*a priori*" knowledge with information contained in the observed data. Models most consistent with both data and prior information correspond to the maxima of this distribution. The tails are described by poorly fitting models in the ensemble, and the "width" or the variance quantifies the constraints we have on model parameters, i.e the uncertainty on the inferred solution. The covariance of the posterior distribution provides information on the correlation or trade-off between model parameters.

[10] The "Bayesian neighborhood algorithm" [*Sambridge*, 1999b] was used for RF inversion by *Lucente et al.* [2005] and *Piana Agostinetti and Chiarabba* [2008]. Subsequently, *Piana Agostinetti and Malinverno* [2010] expanded the Bayesian formulation to the case where the number of layers is not fixed in advance but is treated as an unknown in the problem. At first sight this may sound like an unrealistic prospect, as there would seem to be little to prevent an algorithm introducing ever more detail into a model to improve data fit. However in a transdimensional Bayesian formulation, high dimensional, many layers, models are naturally discouraged [*Malinverno*, 2002]. This results from a convenient property of Bayesian inference referred to as "natural parsimony," i.e. preference for the least complex explanation for an observation. Overly complex models suffer from over-fitting and so have poor predictive power. Therefore, given a choice between a simple model with fewer unknowns and a more complex model that provide similar fits to data, the simpler one will be favored in Bayesian sampling (see *MacKay* [2003] for a discussion). The preference for models with fewer unknowns is an intrinsic feature of transdimensional sampling algorithms.

[11] Transdimensional inversion, i.e. where the dimension of the model space is an unknown, was first used in Earth Science by *Malinverno* [2002] for DC resistivity sounding. Since then, it has rapidly become popular, and has been introduced to a wide range of areas such as geostatistics [*Stephenson et al.*, 2004], thermo-chronology [e.g., *Stephenson et al.* 2006], geochronology [*Jasra et al.*, 2006], palaeoclimate inference [e.g., *Hopcroft et al.*, 2007, 2009], inverse modeling of stratigraphy [*Charvin et al.*, 2009a, 2009b], seismic tomography [*Bodin and Sambridge*, 2009], wire-line log data interpretation [*Reading et al.*, 2010], change point modeling of geochemical records [*Gallagher et al.*, 2011], geoacoustic inversion [*Dettmer et al.*, 2010], potential fields studies [*Luo*, 2010], and inversion of electromagnetic data [*Minsley*, 2011].

[12] *Piana Agostinetti and Malinverno* [2010] appears to be the first application of a transdimensional algorithm to the receiver function problem. In this paper, we extend their scheme to the hierarchical case where data noise levels are also treated as unknowns. Our scheme also solves a longstanding problem in geophysical inversion, i.e. how to determine the relative weights applied to different data types (e.g. receiver functions and dispersion measurements) during an inversion.

1.2. Receiver Function Variance

[13] As shown by *Gouveia and Scales* [1998], the level of data uncertainty estimated prior to inversion (i.e. the covariance

matrix of data noise) plays a critical role in Bayesian inference. In an optimization framework the solution does not depend on the level of data noise (since the best fitting model remains the same as we rescale all error bars of the data). In contrast, with a Bayesian framework, the data uncertainty directly determines the form of the posterior probability distribution and hence the posterior samples generated from it.

[14] In the context of a transdimensional inversion, the variance of data noise becomes even more important. *Piana Agostinetti and Malinverno* [2010] showed a clear relationship between the data errors and number of interfaces in the sampled models. A transdimensional Bayesian procedure automatically adapts the complexity of the solution in order to fit the data up to the level of noise determined by the user. Of course, as more model parameters (e.g. more layers) are introduced, the data could be fitted better, but the procedure naturally prevents the data to be fitted more than the given level of noise [for a recent example, see *Dettmer et al.*, 2010].

[15] For receiver functions the noise is correlated from sample to sample by the band-limited nature of the waveforms. The uncertainty can be characterized into three types. Firstly, observational errors on the seismic waveform result from background seismic noise (micro-seisms) and from the instrumental noise affecting the recording. Often, outlier RFs in the stack are eliminated from visual inspection and hence there is no clear quantification of the degree of observational noise. Secondly, processing errors occur in the deconvolution between components of the seismogram, which is an unstable operation. The frequency domain deconvolution is stabilized with a water-level scheme, whose parameter is chosen by trial-and-error [*Clayton and Wiggins*, 1976].

[16] In addition, there are assumptions made about the Earth (e.g. horizontal homogeneous isotropic layers) that prevent us from reproducing the observed RFs. We refer to the part of the data that cannot be modeled by our physical approximation of the Earth as "theory errors." This type of noise is coherent and fully reproducible, and following the definition of *Scales and Snieder* [1998], it is a part of the signal we choose not to explain. For example, the complex structures near the receiver produce a scattered wavefield that is not taken in account in our forward model and which is thus considered as data noise in the inversion. A Gaussian filter is applied to limit the final frequency band, in order to reduce the sensitivity to fine structure.

[17] As shown by *Di Bona et al.* [1998], all these contributions to the RF variance may not be simply additive and an overall quantification of the noise in terms of magnitude and correlation is often difficult.

1.3. The Covariance Matrix of Data Errors

[18] In most Bayesian studies, the data noise is assumed to be normally distributed, and represented by a covariance matrix of data errors $\mathbf{C}_e$. *Ammon* [1992] estimated the noise level from the power-spectral density in a pre-signal time window (the segment which precedes the direct P-wave arrival). *Sambridge* [1999b] estimated a noise covariance from multiple realizations of correlated noise waveforms. *Piana Agostinetti and Malinverno* [2010] derived the noise correlation from the averaging function which is calculated by deconvolving the vertical component of motion from itself, using the chosen water-level parameter. If the water-level fraction is zero, the result is a perfect Gaussian (from the low-pass filter included in the procedure). *Di Bona et al.* [1998] evaluated the noise involved in the frequency domain deconvolution by using the residuals of a time domain deconvolution of the averaging functions from the computed RFs, in the portion preceding the P-pulse.

[19] These different schemes approximate different effects, and it is clear here that there is no consensus to date on the way of measuring noise in RFs. Although these techniques can be used to infer the level of observational and processing noise, they do not estimate uncertainties due to theoretical errors. Note however that *Gouveia and Scales* [1998] accounted for model discretization errors in the case of Bayesian seismic waveform inversion. In the case of geoacoustic inversion, *Dettmer et al.* [2007, 2008] propose to estimate the covariance of data errors from analysis of data residuals obtained from a maximum likelihood solution.

[20] In this work we address the issue of noise estimation with a Hierarchical approach [*Malinverno and Briggs*, 2004; *Malinverno and Parker*, 2006]. The Hierarchical Bayes model is so named because it has two levels of inference. At the higher level are "hyper-parameters" such as the noise variances of the data. At the lower level are the physical parameters that represent Earth properties, e.g. seismic velocities. Information on physical parameters at the lower level is conditional on the values of hyper-parameters selected at the outer level. Overall, a joint posterior probability distribution is defined both for hyper-parameters and Earth parameters.

[21] Here we use a Hierarchical Bayes formulation where both the variances and correlation parameters of data noise are treated as unknowns in the inversion. In this way we fully take account of the complex combination of effects contributing to the misfit. In the context of a variable number of layers, we shall show that this can be a major advantage over having fixed noise estimates.

1.4. Joint Inversion With Surface Wave Dispersion Measurements

[22] Although RFs are particularly suited to constrain the depth of discontinuities, they are only sensitive to relative changes in S-wave velocities in different layers, and poorly constrain absolute values. Conversely, surface waves dispersion (SWD) measurements are sensitive to absolute S-wave velocities but cannot constrain sharp gradients, and are poor at locating interfaces [*Juliá et al.*, 2000]. The difficulty of quantitatively utilizing data sets with different sensitivities have resulted in most models of shear-wave velocity being based only on either RFs or SWD.

[23] However, with increasing computational power, methods to jointly invert RF and SWD are gaining in popularity [e.g., *Özalaybey et al.*, 1997; *Du and Foulger*, 1999; *Juliá et al.*, 2000, 2003; *Chang et al.*, 2004; *Lawrence and Wiens*, 2004; *Yoo et al.*, 2007; *Tkalčić et al.*, 2006; *Moorkamp et al.*, 2010; *Tokam et al.*, 2010; *Salah et al.*, 2011]. The motivation for this approach is to improve resolution, and reduce the non-uniqueness of the problem and influence of noise. Furthermore, if different types of data are inverted together, the complementary constraints are likely to better resolve structure.

[24] In this work we propose to invert RFs jointly with observations based on the cross-correlation of ambient noise recorded at nearby receivers [*Campillo and Paul*, 2003; *Shapiro and Campillo*, 2004; *Stehly et al.*, 2009], which

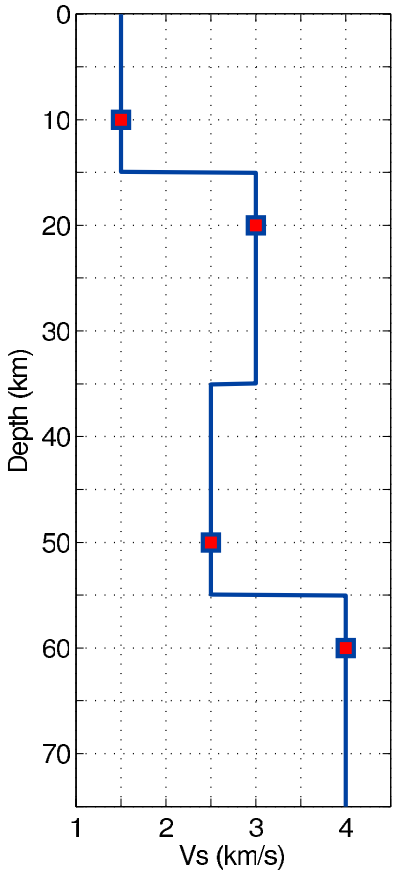

Figure 1. The model is parameterized with a variable number of Voronoi nuclei (red squares) which define the seismic structure (blue line). The vertical location of nuclei define the geometry of layers which Vs value is given by horizontal positioning of nuclei. Note that Voronoi nuclei are not necessary at the center of layers but rather boundaries are defined as equidistant to adjacent nuclei.

provides apparent travel times of surface waves at periods ~1–30 s (mostly sensitive to the crust). In the context of joint inversions, assessment of data uncertainty becomes crucial in the construction and evaluation of a misfit function. This is because data sets of different nature have different levels of noise, and their relative uncertainty determines their relative contribution to the misfit. Often, some arbitrary weighting factor is chosen which begs the question of whether maximum benefit is being obtained from the joint inversion. In this paper we show that a Hierarchical Bayes procedure appears to be effective in this situation, as it is able to quantify the level of information brought by different data types in a self consistent manner.

2. Methodology

2.1. Model Parameterization

[25] In this study the radial RF and SWD are assumed to be dominated by the response of homogeneous horizontal layers beneath the receiver. The geometry of layers is described by a variable number, k, of Voronoi nuclei as shown in Figure 1.The layer boundaries are defined as equidistant between adjacent nuclei, with the lowest layer a half space. Each layer i (with $i\epsilon[1k]$) is therefore determined by the depth of its nucleus c_i and by a shear wave velocity value v_i. By allowing the number of layers, k, to be variable as well as both the position of the nuclei, $\mathbf{c}$, and velocities, $\mathbf{v}$, we have a highly flexible parameterization of variable thickness layers (see Figure 1). In our transdimensional approach, this dynamic parametrization will adapt to the spatial variability in the velocity structure information provided by the data.

[26] We also make inference on the covariance matrix of data noise $\mathbf{C}_e$, which can be expressed with a number of hyper-parameters $\mathbf{h} = (h_1, h_2, \ldots h_m)$, treated as unknowns in the inversion. Therefore, the complete model to be inverted for is defined as $\mathbf{m} = [\mathbf{c}, \mathbf{v}, k, \mathbf{h}]$.

[27] As described by *Lombardi* [2007], the timing of RFs are relative to the first P arrival and thus very sensitive to the variation of Vs relative to Vp. Note that RFs are also sensitive to crustal attenuation. However here, we consider the Vp/Vs ratio as well as attenuation coefficients fixed to a reference model and we only invert for Vs in each layer. Furthermore, we use only the simplest possible representation of velocity within each layer, i.e. a constant, although higher order polynomials are possible, e.g. a linear gradient or quadratic.

[28] In our inversion study, inappropriate modeling assumptions (e.g. no dipping layers or anisotropy) may manifest themselves in a poor data fit. In a conventional Bayesian framework, these theory errors have to be taken into account in the data noise covariance matrix which is often impractical (see *Gouveia and Scales* [1998] for details). For example, how would one quantify the magnitude and correlation of data noise generated by approximating a dipping layer as horizontal, or a complex anisotropic medium as isotropic ? An advantage of the Hierarchical Bayes formulation is that we let the data infer its own degree of uncertainty, and hence theory errors, while still present, are allowed for in the estimation of the data noise and acceptable levels of data fit.

2.2. The Forward Calculation

[29] Our direct search algorithm requires solving the forward problem a large number of times, that is to compute the RF predicted by a given Earth model parameterized as described above. We use the Thomson-Haskell matrix method [*Thomson*, 1950; *Haskell*, 1953] to compute the spectral response of a stack of isotropic layers to an incident planar P-wave. Multiple reflections are considered with this method. Since this way of solving the forward model is achieved without slowness integration, it is fast and has been widely used in Monte Carlo algorithms [e.g., *Shibutani et al.*, 1996; *Sambridge*, 1999a]. Once synthetic seismograms have been computed for different components, receiver functions are made via frequency domain deconvolution of the vertical component from the radial component using water-level spectral division [*Langston*, 1979] with a water-level of 0.0001. In the case of a joint inversion, the forward method used to calculate surface wave dispersion is DISPER80 developed by *Saito* [1988], this algorithm does not consider seismic attenuation. Since the proposed inversion algorithm is a direct parameter search, the forward calculations are separate routines independent of the main algorithm, and hence they can be easily replaced by alternative algorithms.

2.3. Bayesian Inference

[30] In a Bayesian approach all information is represented in probabilistic terms [*Box and Tiao*, 1973; *Smith*, 1991; *Gelman et al.*, 2004]. Geophysical applications of Bayesian inference are described by *Tarantola and Valette* [1982], *Duijndam* [1988a, 1988b], and *Mosegaard and Tarantola* [1995]. The aim of Bayesian inference is to quantify the *a posteriori* probability distribution (or posterior distribution) which is the probability density of the model parameters, $\mathbf{m}$, given the observed data, $\mathbf{d}_{obs}$, written as $p(\mathbf{m}|\mathbf{d}_{obs})$ [*Smith*, 1991]. In a transdimensional formulation, the number of unknowns is not fixed in advance, and so the posterior is defined across spaces with different dimensions. This transdimensional probability distribution is taken as the complete solution of the inverse problem. In practice one tends to use computational methods to generate samples from the posterior distribution, i.e. an ensemble of vectors $\mathbf{m}$ whose density reflects that of the posterior distribution.

[31] Bayes' theorem [*Bayes*, 1763] is used to combine prior information on the model with the observed data to give the posterior probability density function:

$$\text{posterior} \propto \text{likelihood} \times \text{prior}, \tag{1}$$

$$p(\mathbf{m} \mid \mathbf{d}_{obs}) \propto p(\mathbf{d}_{obs} \mid \mathbf{m}) p(\mathbf{m}), \tag{2}$$

where $x|y$ means x given, or conditional on, y, i.e. the probability of having x when y is fixed. $\mathbf{m}$ is the vector of the model parameters and $\mathbf{d}_{obs}$ is a vector defined by the set of observed data. The term $p(\mathbf{d}_{obs}|\mathbf{m})$ is the likelihood function, which is the probability of observing the measured data given a particular model. $p(\mathbf{m})$ is the *a priori* probability density of $\mathbf{m}$, that is, what we (think we) know about the model $\mathbf{m}$ before measuring the data $\mathbf{d}_{obs}$.

[32] Hence, the posterior distribution represents how our prior knowledge of the model parameters is updated by the data. Clearly, if the prior and the posterior distributions are the same, then the data add no new information.

[33] From an ensemble of models distributed according to the posterior, it is straightforward to determine special properties like the best or average model, or to construct marginal probability distributions for individual model parameters. Correlation coefficients between pairs of parameters can also be extracted [*Gelman et al.*, 2004].

2.4. The Likelihood Function

[34] The likelihood function $p(\mathbf{d}_{obs}|\mathbf{m})$ quantifies how well a given model with a particular set of parameter values can reproduce the observed data.

[35] The observed receiver function can be written as

$$\mathbf{d}_{obs}(i) = \mathbf{d}_{True}(i) + \epsilon(i)\ i = [1, n], \tag{3}$$

where n is the size of the data vector, and $\epsilon(i)$ represents errors that are distributed according to a multivariate normal distribution with zero mean and covariance $\mathbf{C}_e$, which may be unknown. We recognize that the Gaussian assumption may itself be questionable in some cases. Furthermore, by assuming the normal distribution has zero mean, we do not account for systematic errors.

[36] In the case of correlated data noise, the fit to observations, $\Phi(\mathbf{m})$, is no longer defined as a simple "least-square" measure but is the Mahalanobis distance [*Mahalanobis*, 1936] between observed, $\mathbf{d}_{obs}$, and estimated, $g(\mathbf{m})$, data vectors:

$$\Phi(\mathbf{m}) = (g(\mathbf{m}) - \mathbf{d}_{obs})^T \mathbf{C}_e^{-1} (g(\mathbf{m}) - \mathbf{d}_{obs}). \tag{4}$$

In contrast to the Euclidean distance, this measure takes in account the correlation between data (equality being obtained where $\mathbf{C}_e$ is diagonal). The general expression for the likelihood probability distribution is hence:

$$p(\mathbf{d}_{obs} \mid \mathbf{m}) = \frac{1}{\sqrt{(2\pi)^n \mid \mathbf{C}_e \mid}} \times \exp\left\{\frac{-\Phi(\mathbf{m})}{2}\right\}. \tag{5}$$

Note that this expression requires both the inverse $\mathbf{C}_e^{-1}$ of the data noise covariance matrix and also its determinant $|\mathbf{C}_e|$.

[37] When treating the data noise as a variable, one might intuitively expect the algorithm to choose high values for the variance of data noise (i.e. the diagonal elements of $\mathbf{C}_e$) because this would reduce the misfit in (4). However, the Gaussian likelihood function is normalized by $|\mathbf{C}_e|$ in (5) and here high data errors implies a low likelihood. Hence the value taken by the magnitude of data noise has two competing effects on the likelihood.

2.5. The Prior

[38] The Bayesian formulation enables one to account for prior knowledge, provided that this information can be expressed as a probability distribution $p(\mathbf{m})$ [*Gouveia and Scales*, 1998]. All inferences from the data are then relative to this prior. In our 1D seismic inverse problem, this prior information is what we think is reasonable for the shear-wave velocity model we want to infer, according to previous studies.

[39] In a stimulating short essay, *Scales and Snieder* [1997] reviewed the philosophical arguments that have been invoked for and against Bayesian inversion. The principal criticism made to Bayesian inversion is that users often "tune" the prior in order to get the solution they expect. In other words, the *a priori* knowledge of the model is often used as a control parameter to tune the properties of the final model produced. Therefore, one can easily argue that in a Bayesian framework, the solution is influenced by the form of the prior distribution, whose choice is subjective.

[40] However in the examples shown here, the final models will be dominated by the data rather than by prior information and so we do not consider this to be a major limitation. This is because we assume unobtrusive prior knowledge by setting priors to uniform distribution with relatively wide bounds, although we acknowledge that uniform distributions are very informative about their bounds, and hence it is not possible to have a completely uninformative prior.

[41] The complete mathematical form of our prior distribution is detailed in Appendix A.

2.6. Transdimensional Inference

[42] Given the Bayesian formulation described above, our goal is to generate a collection or ensemble, of Earth models distributed according to the posterior function. In our problem,

we do not know the number of layers, i.e. the dimension of the model space is itself a variable, and hence the posterior becomes a transdimensional function. This can be sampled with a generalization of the well known Metropolis-Hasting algorithm [*Metropolis et al.*, 1953; *Hastings*, 1970] termed the reversible-jump Markov chain Monte Carlo (rj-McMC) sampler [*Geyer and Møller*, 1994; *Green*, 1995, 2003] which allows inference on both model parameters and model dimensionality.

[43] A general review of transdimensional Markov chains is given by *Sisson* [2005] and *Gallagher et al.* [2009] present an overview of the general methodology and its application to Earth Science problems. The reversible jump algorithm is described in previous studies [e.g., *Malinverno*, 2002; *Gallagher et al.*, 2011]. Here we only give a brief overview of the procedure, and present the mathematical details of our particular implementation in Appendices.

[44] The rj-McMC method is iterative in which a sequence of models are generated in a chain, where typically each is a perturbation of the last. The starting point of the chain is selected randomly and the perturbations are governed by a proposal probability distribution which only depends on the current state of the model. The procedure for a given iteration can be described as follows: (1) Randomly perturb the current model, to produce a proposed model, according to some chosen proposal distribution (see Appendix B). (2) Randomly accept or reject the proposed model (in terms of replacing the current model), according to the acceptance criterion ratio (see Appendix C).

[45] The first part of the chain (called the burn-in period) is discarded, after which the random walk is assumed to be stationary and starts to produce a type of "importance sampling" of the model space. This means that models generated by the chain are asymptotically distributed according to the posterior probability distribution (for a detailed proof, see *Green* [1995, 2003]). If the algorithm is run long enough, these samples should then provide a good approximation of the posterior distribution for the model parameters, i.e. $p(\mathbf{m}|\mathbf{d}_{obs})$.

[46] This "ensemble solution" contains many models with variable parameterization, and inference can be carried out with ensemble averages over the structure [see *Piana Agostinetti and Malinverno*, 2010]. For example, the posterior probability of the shear wave velocity at a given depth can be visualized simply by plotting the histogram of the values selected for the ensemble solution.

[47] In terms of choosing a single model for interpretation, we can consider the average over the ensemble of sampled models. This is known as the expected model, and is in fact a weighted average, in which the weighting is through the posterior distribution (sampled by the rj-McMC algorithm). All the models sampled have a particular parameterization defined by the number and position of their interfaces. When a large number of models are averaged, the positions of well defined interfaces will tend to overlap while less well defined ones will tend to cancel out. This spatial average model is effectively a continuous line which will capture the well resolved parts of the model [*Bodin and Sambridge*, 2009].

[48] Here there is no need for statistical tests or regularization procedures to choose the adequate model complexity or smoothness corresponding to a given degree of data uncertainty. Instead, the reversible jump technique automatically adjusts the underlying parametrization of the model to produce solutions with appropriate level of complexity to fit the data to statistically meaningful levels.

2.7. Data Uncertainty Quantification: Hierarchical Bayes

[49] The covariance matrix of data noise $\mathbf{C}_e$ in (4) imposed at the outset has a direct effect on the solution of the reversible jump algorithm and implicitly acts as a smoothing parameter. This can be seen as an advantage over optimization based schemes where the level of smoothing is chosen *a priori* or interactively. However, in seismology, assessment of measurement errors can be difficult to achieve *a priori*. Without any reliable information about the data uncertainty, it is impossible to give a preference between two solutions obtained with different values of $\mathbf{C}_e$.

[50] Fortunately, an expanded Bayesian formulation can take into account the lack of knowledge we have about data errors. Following current statistical terminology, the data noise covariance matrix is expressed with a number of hyper-parameters (i.e. $\mathbf{C}_e = f(h_1, h_2, \ldots)$), and the method used is known as Hierarchical Bayes [*Gelman et al.*, 2004]. The model to be inverted for is defined by the combined set $\mathbf{m} = [\mathbf{c}, \mathbf{v}, k, \mathbf{h}]$, where $\mathbf{c}$ and $\mathbf{v}$ are the vectors containing the nuclei locations and velocity values, and $\mathbf{h} = (h_1, h_2, \ldots)$ is a vector of hyper-parameters defining the unknown data errors. The Hierarchical algorithm is implemented in the same manner as the conventional reversible jump, the only difference being that here we add an extra type of model perturbation, i.e. a change in the hyper-parameter vector $\mathbf{h}$. As for other model parameters, each time $\mathbf{h}$ is perturbed, a new value is randomly proposed from a given distribution, and the new value of data noise is either accepted or rejected according to the acceptance criterion ratio (see Appendix C).

2.8. Parameterizing the Covariance Matrix of Data Noise

[51] As explained before, our philosophy is to consider the level of data noise as an unknown in the inversion. Therefore the main issue here is to "parameterize" the noise covariance matrix $\mathbf{C}_e$, i.e. to express it as a function of a given number of hyper-parameters. This is a symmetric $n \times n$ matrix defined with $(n^2 + n)/2$ values which are obviously impossible to estimate separately from only n data, and hence some assumptions need to be made. The noise covariance can be written in terms of a matrix of correlation $\mathbf{R}$ and a vector of standard deviations $\mathbf{s}$:

$$\mathbf{C}_e = \mathbf{s}^t \mathbf{R} \mathbf{s}. \tag{6}$$

With this decomposition, one can separate two properties of the noise, i.e its magnitude and correlation [*Piana Agostinetti and Malinverno*, 2010]. For simplicity, in this study the noise is considered stationary, i.e. its magnitude and correlation are constant along the time series (although we acknowledge that this might not be always the case in RFs), then $\mathbf{C}_e$ can be written:

$$\mathbf{C}_e = \sigma^2 \mathbf{R}, \tag{7}$$

where σ^2 is the constant noise variance, i.e. the magnitude of data noise. (In a case of a non-stationary time series, σ can be

parameterized as a linear function of time $\sigma(t) = h_1 \times t + h_2$). $\mathbf{R}$ is a symmetric diagonal-constant or Toeplitz matrix:

$$\mathbf{R} = \begin{bmatrix} 1 & c_1 & c_2 & \dots & c_{n-1} \\ c_1 & 1 & c_1 & \dots & c_{n-2} \\ c_2 & c_1 & 1 & \dots & c_{n-3} \\ & & & \vdots & \\ c_{n-1} & c_{n-2} & c_{n-3} & \dots & 1 \end{bmatrix}, \quad (8)$$

where $c_i (i = [1,n])$ describes the noise correlation between points of the series. Thus c_1 defines the correlation between two adjacent points, and more generally c_i is the noise correlation between points that are i samples apart in the series. The key properties that we need are that the correlation function decreases with distance, with limiting values of 1 at $i = 0$ and of 0 at $i = \infty$. This is the most common kind of association found in times series. Then, the main question is how to parameterize the correlation function c_i and with how many unknowns? Below we present two types of parameterization for the noise correlation.

[52] We first propose a parameterization which is convenient to implement for our particular problem. The correlation function is simply assumed to decay exponentially and is thus given by

$$c_i = r^i, \quad (9)$$

where $r = c_1$ is a constant number between 0 and 1. The major advantage of such a parameterization is that the inverse and determinant of $\mathbf{C}_e$ needed in the likelihood in (5) have simple analytical forms, i.e. they can be expressed in terms of our two hyper-parameters $\mathbf{h} = [\sigma,r]$ describing the magnitude and correlation of data noise (see Appendix D).

[53] A second type of parameterization that is commonly used to model the noise in RFs is a Gaussian correlation law:

$$c_i = r^{(i^2)}. \quad (10)$$

Compared to the first type of correlation, here there are no high frequency components, and hence this form of noise clearly seems closer to what is observed in receiver functions before the first P-arrival (see Appendix D). This is because a Gaussian filter is used in the deconvolution process to remove high frequency noise that has high amplitude and which blurs the signal. Although this type of noise parameterization appears to be more realistic in the case of RF, there are no stable analytical formulations for the inverse and determinant of $\mathbf{C}_e$. Therefore $\mathbf{C}_e^{-1}$ and $|\mathbf{C}_e|$ have to be numerically computed, which is computationally expensive and cannot be carried out each time $\mathbf{C}_e$ is perturbed along the random walk. Therefore, here one can only invert for the magnitude of noise $h = [\sigma]$ whereas its correlation r needs to be chosen by the user.

2.9. Adding Surface Wave Dispersion Data Into the Problem

[54] Given the framework described above, it is straightforward to invert jointly independent data types with different units and levels of noise. This is done simply by defining the data vector $\mathbf{d}$ and covariance matrix of data errors $\mathbf{C}_e$ in (5) as a concatenation of the data vectors and noise covariance matrices. In the case of a joint inversion of RF and SWD, this yields

$$\mathbf{d} = \left[\mathbf{d}^{RF}, \mathbf{d}^{SWD}\right], \quad (11)$$

$$\mathbf{C}_e = \begin{bmatrix} \mathbf{C}_d^{RF} & 0 \\ 0 & \mathbf{C}_d^{SWD} \end{bmatrix}, \quad (12)$$

where $\mathbf{C}_e^{RF}$ is constructed with two parameters σ^{RF} and r^{RF}. For surface wave dispersion data, we assume the measurement error is constant with period and $\mathbf{C}_e^{SWD}$ is also constructed with $\sigma^{SWD}$and r^{SWD} (note that instead of being constant, σ^{SWD} could also be parameterized as a linear function of period).

[55] In this way, $\mathbf{C}_e$ can be parameterized with different noise parameters for each data type. As can be seen in (4), the level of noise in different data types controls their contribution in the misfit function, and hence their influence in the solution. Therefore, by inverting for data noise, we let the data themselves infer the contribution of each data type in the misfit, without having to define a scale factor to weight independent data sets.

[56] An alternative to this joint Bayesian inversion would be a two-step inversion where the posterior distribution obtained after inverting a first data set $\mathbf{d}_1$ would determine the prior for a second inversion based on the second data set $\mathbf{d}_2$. In the case where the two data sets are independent, we can write $p(\mathbf{d}_1,\mathbf{d}_2|\mathbf{m}) = p(\mathbf{d}_1|\mathbf{m})p(\mathbf{d}_2|\mathbf{m})$, and from (2) the posterior solution after the two step inversion would be identical to the posterior solution for the joint inversion.

3. Inversion of Synthetic Data

[57] We first test our algorithm with synthetic data computed from a known velocity model made of 6 horizontal layers. The true model (red line in Figure 3) presents two major features often targeted by RF studies: a low S-wave velocity layer in the crust (between 10–20 km) and a strong velocity increase at the Moho (at about 30 km depth). A synthetic receiver function (red line in Figure 2) is calculated from the true model with the forward method described in section 2.2, and a correlated random noise is added, which results in the "observed" receiver function (blue line in Figure 2). An alternate approach would be to add noise to the seismic waveforms before the deconvolution. However, here the receiver function waveform is the data vector to be inverted. By directly adding the correlated noise to this vector, we know the exact form of the data noise, and can verify that the proposed algorithm is able to recover it. The synthetic noise is generated according to a covariance matrix $\mathbf{C}_e$ defined with the first type of correlation (i.e. $c_i = r^i$) with values $\sigma_{true} = 2.5 \times 10^{-2}$ and $r_{true} = 0.85$, and hence inversions carried out in this section assume an exponential correlation law. We recognize that this type of noise contains high frequency signals that are normally filtered out in real data. Therefore the results presented in this section should only be seen as a "proof-of-concept," and not as a statement of the optimal noise parameterization.

[58] In order to illustrate the different features of the algorithm, we first present results for a conventional transdimensional RF inversion with fixed noise estimates. Then, we

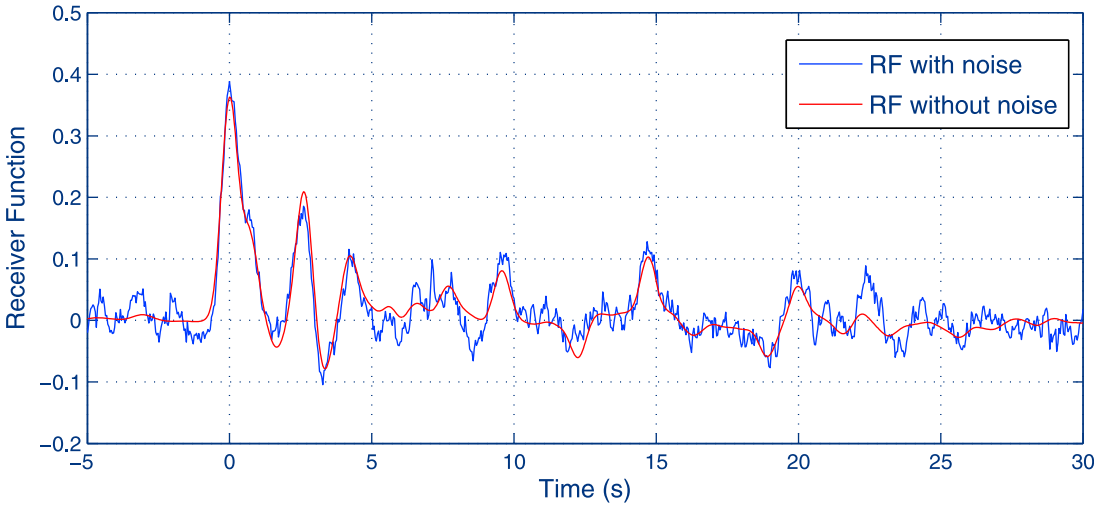

Figure 2. Simulated receiver function. Red: Synthetic data estimated from the true model in red in Figure 3. Blue: RF with added Gaussian random noise generated with an exponential correlation law (i.e. $c_i = r^i$).

extend the formulation to Hierarchical models, i.e. treat the data noise parameters (σ and r) as variables in the inversion. Finally, surface wave dispersion measurements with unknown errors are added into the problem for a joint inversion.

3.1. Transdimensional Inversion of RF With Fixed Noise Parameters

[59] The purpose of this section is to show that in a standard transdimensional inversion, i.e. where the estimated data noise is fixed to some values given by the user prior to the inversion, the form of the solution strongly depends on the choice of the covariance matrix of data errors. Hence the situation in this section is similar to that considered by *Piana Agostinetti and Malinverno* [2010]. The RF shown in Figure 2 has been inverted twice with a different matrix $\mathbf{C}_e$ at each run. The inversion is first carried out with the correct noise estimates (Figures 3a–3c), and then using incorrect values for σ and r (Figures 3d–3f). In the second case, σ has been underestimated by 40% (we use 0.015 instead of 0.025), and r has been overestimated by 8% (we use 0.92 instead of 0.85). Hence in this test we observe the effects of misestimating the data noise.

[60] For the two cases, the transdimensional sampling has been carried out allowing between 2 and 50 interfaces ($n_{\min} = 2$ and $n_{\max} = 50$). Bounds for the uniform prior distribution were set to 2–5 km/s for S-wave velocity values. Posterior inference was made using an ensemble of 10^5 models. The algorithm was implemented on 200 parallel cpus to allow large numbers of independent chains, starting at different random points, and sampling the model space simultaneously and independently. Each chain was run for 2×10^5 steps in total. The first half was discarded as burn-in steps, only after which the sampling algorithm was judged to have converged. To eliminate dependent samples in the ensemble solution, every 200th model visited in the second half was selected for the ensemble. The convergence of the algorithm is monitored with a number of indicators such as acceptance rates, and sampling efficiency is optimized by adjusting the variance of the Gaussian proposal functions (see Appendix B). (For details on convergence and independence of sampled models in Markov chains, see *MacKay* [2003].)

[61] The solution is given by the transdimensional posterior distribution which is represented by an ensemble of 1D models with variable number of layers and thicknesses. In order to visualize the final ensemble, the collected models can be projected into a number of physical spaces that are used for interpretation. For example, Figures 3a and 3d show the marginal distribution for S-wave velocities as a function of depth. At each depth, local information about the velocity model is represented by a complete distribution which can be seen as a marginal distribution of the posterior in this "interpretation space." These marginal posteriors are shown as a color density map in Figure 3. In practice, the marginal posterior is simply constructed from the density plot (i.e. the histogram) of the ensemble of models in the solution. This density plot is convenient to visualize the ensemble solution, and it is particularly useful to demonstrate the constraints on Vs.

[62] If one is interested in assessing the resolution (number and position) of seismic discontinuities beneath the seismic station, it is possible to examine the ensemble solution from a different point of view and to plot the marginal posterior distribution on the location of interfaces. Figures 3b and 3e also show histograms of interface depths in the ensemble of models. For each depth, this function represents the probability density of having a discontinuity, given the data. This provides useful information on the inferred locations of transitions, which can be unclear in other plots. Note that the positions of interfaces are not direct model parameters, and hence this marginal distribution is again constructed by projecting the ensemble of sampled earth models into a visualization space.

[63] Since the models in the ensemble solution have varying number of cells, the complexity of the solution cannot be described with a single number k. However we plot Figures 3c and 3f the histogram of k across the ensemble solution that is directly proportional to the marginal posterior

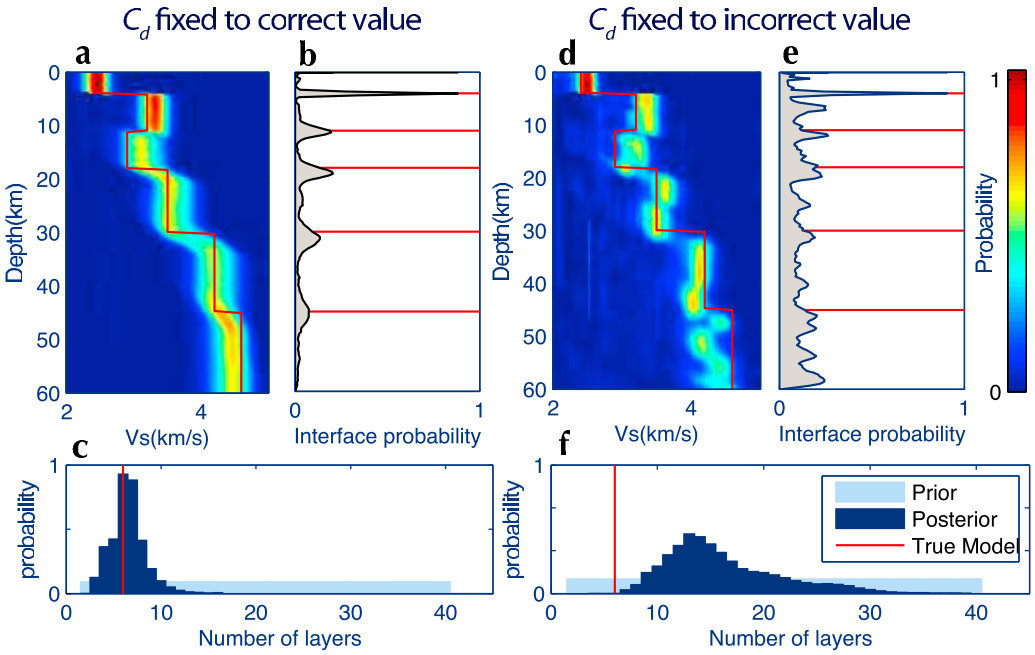

Figure 3. Transdimensional inversion of the synthetic RF (in blue in Figure 2). Here the noise parameters σ and r are kept fixed during the inversion to predefined values. (a–c) Results when noise estimates are set equal to the values used to construct the synthetic noise. (d–f) Results when σ and r are respectively under- and over-estimated relative to their "true" values. In this case the posterior approximation of the true model in red is clearly worsened. Figures 3a and 3d show posterior probability distribution for Vs at each depth. Red shows high probabilities and blue low probabilities. The synthetic true velocity model is plotted as a red line. Figures 3b and 3e show posterior probability for the position of discontinuities. Red lines show depth of interfaces in the true model. Figures 3c and 3f show posterior probability on the number of cells. The red line shows the number of cells in the true model.

$p(k|\mathbf{d}_{obs})$. The number of layers in the true model is shown by the red line, and the uniform prior distribution on k, $p(k)$ is shown in light blue.

[64] Clearly, the solution obtained with correct noise estimates gives better results (Figures 3a–3c) than when the noise is misestimated (Figures 3d–3f). Since the magnitude of noise has been underestimated, the algorithm automatically adds more layers than necessary and "overfits" the observed RF by fitting the unattributed noise. The expected number of layers in the model in Figure 3f is 15, which is twice the true value. Figure 3e shows that location of transitions are not recovered as well. However, even though features are generally degraded in comparison with Figure 3a, the most resolvable elements in Figure 3d (i.e. location of shallow discontinuities) are recovered. We verified in a separate experiment (not shown here) that if the data are assumed as noisier than they really are (i.e. σ in $\mathbf{C}_e$ is overestimated), then this would tend to fit a model that has too few cells (i.e. the model would be too simple).

[65] Therefore, the number of model parameters used, and hence the complexity of the solution, clearly depends on the estimated data noise imposed at the outset. The reversible jump technique automatically adjusts the underlying parametrization of the model to produce an average solution with just enough complexity to fit the data. This is a potential advantage over optimization based inversions. However, as seen before, assessment of measurements errors in RFs (and also in SWD) can be difficult a priori. Without any reliable information about the data uncertainty, it is impossible to give a preference between two solutions in Figure 3 obtained with different values of $\mathbf{C}_e$.

3.2. Hierarchical Bayes Inversion of RF

[66] In this section we consider the situation where little is known about data noise. We repeat the experiment of section 3.1 with exactly the same data vector (Figure 2) but instead treat the noise parameters σ and r as unknowns. Interestingly, Figure 4 shows results for this test which are virtually identical to those obtained when noise parameters are fixed to their correct values (as in Figures 3a and 3b).

[67] In Figure 4a, we show a random sample of 60 models in the ensemble solution. This is a very small subset of the ensemble solution and cannot be used to infer statistical properties, although it can be useful to visualize the solution. Note that some of these profiles are far from the true model in red and may not fit the data well, and reflect models from low probability regions of the posterior distribution. We show how particular 1D models for interpretation can be constructed from the ensemble solution. Firstly, we plot (in blue) the posterior mean model, simply constructed by taking the average Vs at each depth across the ensemble solution. We call this model the "average solution." When models with different transition locations are added, the sharp changes present in individual models are smoothed out, while those at similar locations are reinforced. In this way, the average solution can exhibit at the same time sharp

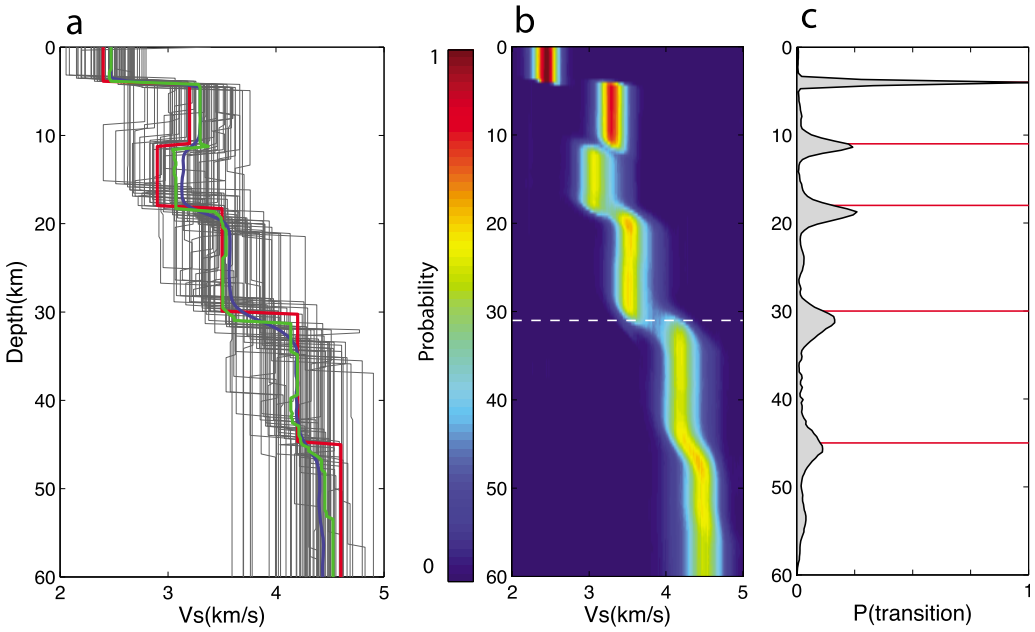

Figure 4. Hierarchical Bayes inversion of the synthetic RF in blue in Figure 2. Here the noise parameters σ and r are treated as unknowns in the inversion. (a) Black lines show a random subset of 60 models in the ensemble solution, which contains 10^5 models, and which is fully represented with (b) a color density plot. The synthetic true velocity model in Figure 4a is plotted as a red line. The blue line shows the posterior mean model (or average solution) constructed by taking the average Vs at each depth across the ensemble solution. The green line shows the maximum of marginal posterior model (or maximum solution) which follows the maximum of the distribution on Vs with depth. (c) Posterior probability for the position of discontinuities. Red lines show depth of interfaces in the true model and correspond well to the peaks in the distribution.

discontinuities and low gradients (blue line in Figure 4b). Instead of being predefined in advance by a single regularization parameter, the level of smoothness in the average solution is variable with depth and directly inferred by the data.

[68] A second model that can be constructed is the mode of marginal posterior which follows the maximum of the marginal with depth (shown in green in Figure 4a). We call this model the "maximum solution." Note that these two models are merely properties of an ensemble of models that have variable parameterizations, and hence do not correspond to any particular individual member of the ensemble. Note also that the maximum solution model is different from the best fitting model in the ensemble, and from the model that maximizes the posterior distribution overall. By inspection it is clear that those models provide a good estimation of the true model in red. Furthermore, all five transitions present in the true model are well recovered in Figure 4c.

[69] Figure 5 shows posterior inference on the number of layers and noise hyper-parameters, together with prior distributions. The number of layers in the true model, as well as true values σ_{true} and r_{true} used to generate the synthetic noise are showed in red. With scant information on data errors, and on the complexity of the true model prior to the inversion, the Hierarchical Bayes procedure has been able to infer the magnitude and correlation of data noise, which quantified the required level of data fit, and thus the number of model parameters needed in the inversion. Therefore this example demonstrates that, by allowing the user to formulate the full state of uncertainties she has about data noise, a Hierarchical Bayesian procedure enables one to correct for a lack of knowledge about data noise.

[70] The RF inverse problem is highly non linear, and hence the posterior is far from being a unimodal Gaussian distribution. To illustrate this, we have plotted in Figure 6 the marginal distribution on Vs at 31 km depth. This cross-section corresponds to the dashed line in Figure 4b, and it is close to the "Moho transition" in the true model. As a result, the marginal distribution is influenced by velocity values in the two layers on each side of the Moho, and it has two maxima about these two values. In this case, one can see that the average solution in blue is not representative of the true model whereas the maximum solution in green is closer to the true velocity in the lower interface. Although the average solution model is smooth, the maximum solution model (jumping from one value to the other) is better at showing sharp transitions.

[71] Finally, we give an example of trade-off assessment between two model parameters, that is Moho depth *vs* S-wave velocity in the last layer of the crust. Again, here the depth of an interface is not strictly a model parameter but a useful feature that can be picked in any sampled model. The crust-mantle transition is defined in the Voronoi models as the closest discontinuity to 30 km. (We acknowledge that this definition for the Moho is somewhat arbitrary, however, here the main purpose is to illustrate trade-off assessment between selected seismic properties.) Figure 7 shows the 2D marginal posterior for the selected pair of parameters, which

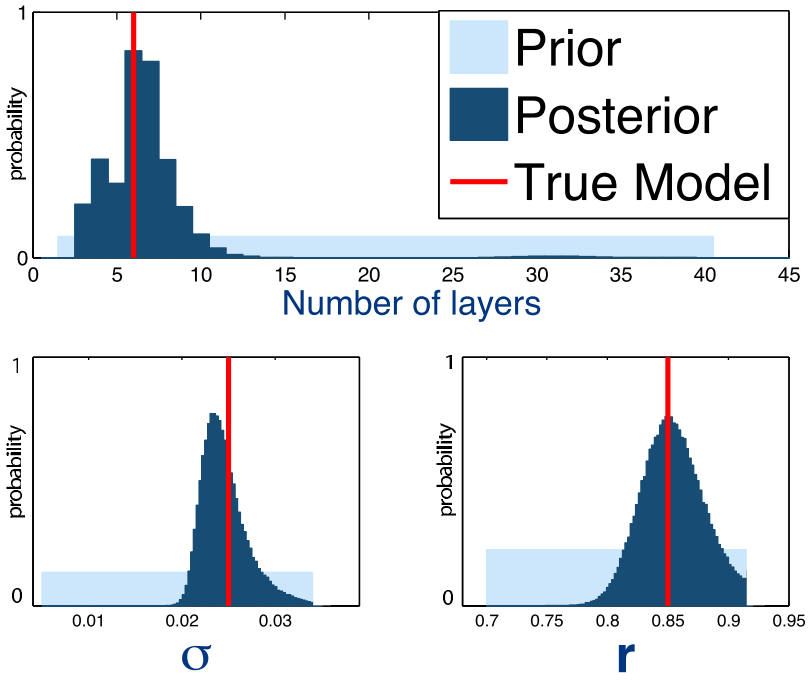

Figure 5. Hierarchical Bayes inversion. (top) Posterior distribution for the number of layers, (bottom left) the standard deviation of data noise σ, and (bottom right) correlation of data noise r. The uniform prior is showed in light blue, and true values (used to construct the observed RF in blue in Figure 2) are shown in red.

is obtained from the 2D histogram over the ensemble of models. In this way one can extract accurate and quantifiable information from the ensemble about the constraints and correlation for these parameters. This trade-off means unsurprisingly, that data are fit equally well when the Moho is deeper and Vs is higher or *vice-versa*, and reassuringly this limitation of the resolving power of the data is clearly evident in the analysis.

[72] Another trade-off that can be quantified is the correlation between the number of cells k and the magnitude of data noise σ. Figure 8 shows the joint posterior distribution on these two parameters. As expected, as the model complexity increases, the data can be better fit, and the inferred value of data errors decrease. However, the degree of trade-off is limited and the data clearly constrains the joint distribution of the two parameters reasonably well.

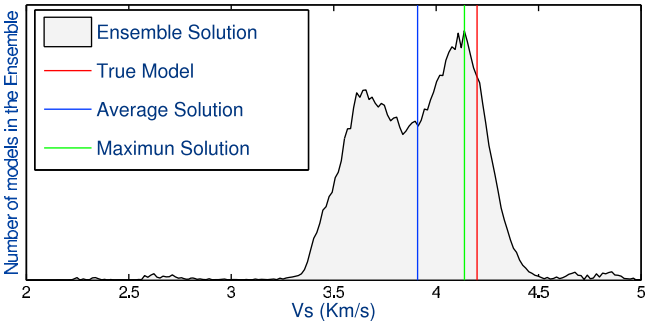

Figure 6. Marginal posterior for Vs at 31 km depth (i.e. slightly after the Moho discontinuity). The distribution is clearly influenced by both Vs value taken above and under the discontinuity. Blue line shows the mean value (average solution).Green line is the maximum of the marginal (maximum solution), and the red line is the true value (i.e. the Vs value in the first layer of the mantle in the true model).

3.3. Joint Inversion of RF and SWD

[73] In this last section of synthetic experiments, we show how surface wave dispersion measurements with unknown errors can be added to the problem without having to predefine the weight given to different data types in the inversion. The same synthetic model as sections 3.2 and 3.1 is used to construct a synthetic RF and SWD (red lines in Figure 9), with forward methods mentioned in section 2.2. A synthetic noise randomly generated from (12) is added to data to produce the two observed data sets shown in blue in Figure 9. Since dispersion data are not time series but travel-time measurements at different periods, we consider the

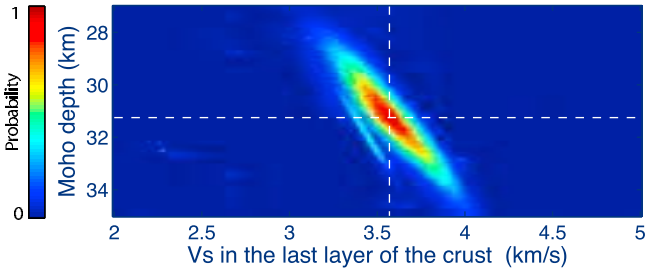

Figure 7. Posterior 2D marginal for the parameters representing depth of Moho and Vs in the last layer of the crust. (The Moho is defined as the closest interface to 30 km in the Voronoi models.) White dashed lines show true values for both parameters.

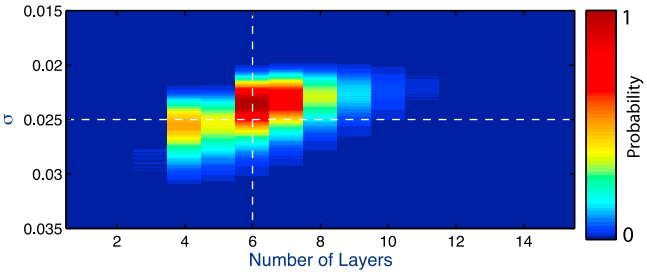

Figure 8. Joint posterior distribution for the number of layers k and the magnitude of data noise σ (note here that k is a discrete variable whereas σ is continuous). White dashed lines show the true values used to construct the synthetic RF. As expected, a clear negative correlation can be seen between the two variables.

noise independent (i.e. not correlated), and use a diagonal covariance matrix to generate SWD noise (i.e. $\sigma_{true}^{SWD} = 0.1$ and $r_{true}^{SWD} = 0$). The two values used to generate random noise for RF are $\sigma_{true}^{RF} = 4 \times 10^{-2}$, $r_{true}^{RF} = 0.85$ with an exponential correlation law.

[74] In the inversion we assume for simplicity $\mathbf{C}_e^{SWD}$ diagonal in (12), and $\mathbf{C}_e^{RF}$ with the first type of parameterization (exponential correlation law), and hence invert for three hyper-parameters σ^{RF}, r^{RF}, and σ^{SWD}. In order to show the constraints brought by each of the two data sets, we first show results obtained after inverting RF data alone (Figures 10a and 10b) and SWD data alone (Figures 10c and 10d). The posterior distributions in Vs are wide in Figure 10a and only the shallowest discontinuity is recovered in 10b. This is because the RF is more noisy than in sections 3.2 and 3.1, and hence the model is poorly recovered. The RF alone contains little absolute velocity information, and this gives rise to a non-uniqueness problem known as the velocity-depth ambiguity (see Figure 7).

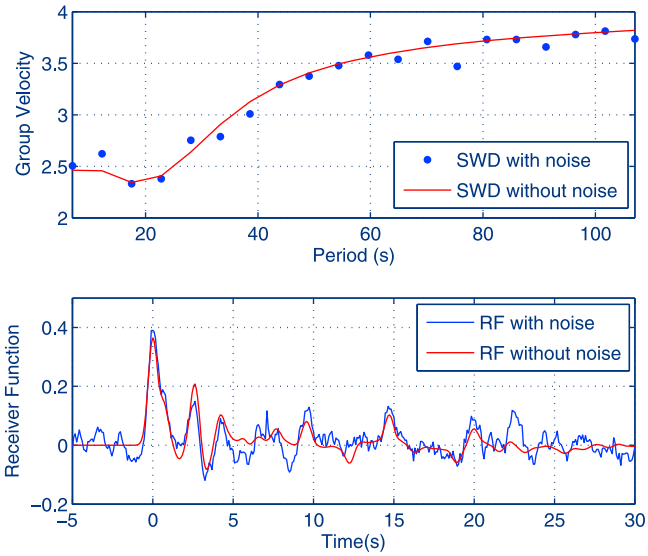

Figure 9. Simulated data with and without added noise for (top) SWD and (bottom) RF. In order to illustrate the benefits of a joint inversion, here the magnitude of noise added to the receiver function is larger than in sections 3.2 and 3.1.

[75] As expected, when SWD data are inverted alone, average velocities are well recovered in Figure 10c but discontinuities are not constrained at all in Figure 10d. Joint inversion results in Figures 10e and 10f show a dramatic improvement. By adding dispersion curves to RF data, the information on discontinuities has been clearly revealed. This might seem counter-intuitive since SWD data are poor at locating interfaces. However, as shown in Figure 7, RF data exhibit a trade-off between velocities and depth of interfaces. Therefore, given this correlation relation, by constraining velocities, SWD data also indirectly constrain

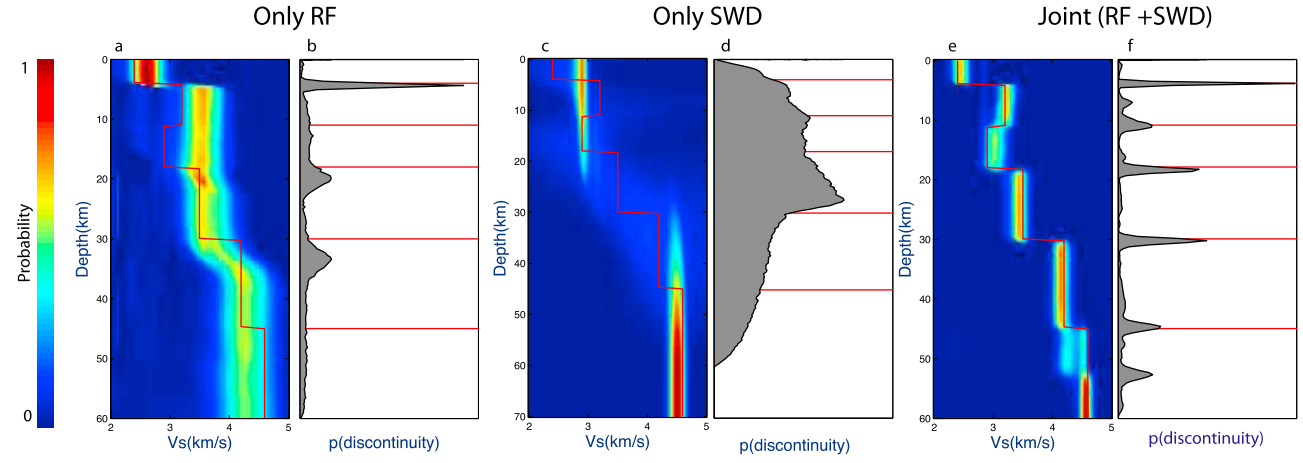

Figure 10. Separate and joint inversion results for the synthetic data in blue in Figure 9. (a and b) RF inversion. Receiver functions are sensitive to strong gradients in elastic properties (e.g. velocity discontinuities in the crust and upper mantle), but are not sensitive to absolute velocity structure. (c and d) SWD inversion. Contrarily to RFs, surface waves dispersion measurements are sensitive to absolute S-wave velocities but cannot constrain sharp gradients, and give poor results in locating interfaces. (e and f) Joint inversion. The information on seismic discontinuities is clearly revealed.

depth of discontinuities. Thus, the two data types are complementary.

[76] For both separate and joint inversions, the algorithm is able to recover the level of noise added to each data set, and hence to fit each data type up the required level. Posterior distributions on number of layers and noise parameters are not shown, since such plots would be similar to those in Figure 5. A number of experiments have been carried out by changing the number of data points and levels of noise in each data. In all cases, the algorithm allows inference on data errors and both data types are adequately fitted.

4. Inversion of Field Measurements

[77] To demonstrate how the algorithm fares on real data, we apply it to RF and ambient noise SWD data from the WOMBAT experiment [*Rawlinson and Kennett*, 2008], which is an extensive program of temporary seismic array deployments throughout southeast Australia. Each array consists of between 30 to 60 short period instruments that continuously record for between five to ten months. Over the last decade, a total of over 500 sites have been occupied resulting in a very large passive seismic data set that has been used for several studies [e.g., *Graeber et al.*, 2002; *Rawlinson et al.*, 2006; *Rawlinson and Urvoy*, 2006; *Clifford et al.*, 2007; *Rawlinson and Kennett*, 2008; *Arroucau et al.*, 2010; H. Tkalčić et al., Multi-step modeling of receiver-based seismic and ambient noise data from WOMBAT array: Crustal structure beneath southeast Australia, submitted to *Geophysical Journal International*, 2011]. Here we focus on SEAL3, which is one of the twelve temporary deployments occupying the eastern and central sub-province of the Lachlan Fold Belt located in New South Wales (Tkalčić et al., submitted manuscript, 2011). We show results of our 1D transdimensional joint inversion beneath a particular station.

4.1. Constructing Receiver Functions

[78] For each station of the SEAL3 array, Tkalčić et al. (submitted manuscript, 2011) constructed a RF waveform from a selection of events with $m_b \geq 5.5$, and with epicentral distances between 30° and 90° from the station, which ensured near-vertical incidence of P-waves. Furthermore, only events with back-azimuths confined between 0° and 90° (i.e. Tonga-Fiji) were used in this study. By choosing a narrow interval of ray-parameters and a narrow azimuthal range, possible Moho dip and anisotropy are neglected, which are only second-order effects in the context of deriving 1D models of the crust and upper mantle that are compatible with multiple geophysical data sets.

[79] For each event, all three components were cut for the same time window and rotated to radial and tangential. Then, radial RFs were calculated using the time domain iterative deconvolution procedure proposed by *Ligorria and Ammon* [1999]for a low pass Gaussian filter with parameter $a = 2.5$, which determines the width of the filter and hence the frequency content of the RF. In order to select RFs that are mutually coherent and could be stacked to determine an observed RF at each station, the cross-correlation matrix approach described by *Tkalčić et al.* [2011] and *Chen et al.* [2010] was used. More details and figures describing the preprocessing of RFs are given by *Tkalčić et al.* [2011]. In this study, we propose to invert for shear-wave structure beneath station S3A6 (−32.900S 148.909E), and hence the average RF across all selected events recorded at this station is used as the data vector in our algorithm (Figure 11). During the inversion, synthetic RFs predicted by proposed models are computed with a simple frequency domain deconvolution technique as detailed in section 2.2. Hence the observed and estimated data vectors in our inversion are computed with different deconvolution methods. There is a scale factor of two in amplitudes in the two different RF computation, and the observed RF has been normalized for comparison with estimated data.

4.2. Ambient Noise Surface Wave Dispersion

[80] SEAL3 deployment provided a large amount of high-quality continuous records for the duration of more than seven months. *Arroucau et al.* [2010] recently used the large volume of recorded noise, which comes from diffuse sources of seismicity such as oceanic or atmospheric disturbances, to produce SWD measurements. The cross-correlation of the ambient noise wavefield was computed on the vertical component of all simultaneously recording station pairs. The resulting time-averaged cross-correlograms exhibit a dispersed wavetrain, which can be interpreted as the Rayleigh wave component of the Green's function of the intervening medium between the two stations [*Shapiro and Campillo*, 2004]. Rayleigh wave group traveltimes were then determined from the cross-correlograms (see *Arroucau et al.* [2010] for details).

[81] The goal of this study is to invert for a 1D shear-wave velocity model from simultaneous modeling of RF and SWD. However, RFs are associated with discrete points in the geographical space, i.e. the location of single stations, while SWD measurements are path averaged data related to station pairs. In order to construct a dispersion curve associated with the location of a station, Tkalčić et al. (submitted manuscript, 2011) adopted the following strategy: Station pairs located within a radius of 150 km around the station of interest were selected and an average dispersion curve was calculated from all the group velocity measurements performed in that area. In order to insure reliable group velocity estimates, a minimum distance between station pairs equal to two wavelengths was required. Furthermore, the average velocities were only calculated if more than twenty observations were available for a given period. (Note that another way to produce point measurements for SWD is to carry out 2D tomographic inversions of travel-times at each period). The average SWD curve thus obtained at station S3A6 is shown in Figure 11, and is inverted simultaneously with the observed RF at this station in order to infer a 1D shear-wave model beneath the recording site.

4.3. Results

[82] Posterior inference was made using an ensemble of about 5.2×10^5 models. The reversible jump algorithm was implemented for 288 parallel cpu-cores sampling the model space simultaneously and independently. Each chain was run for 1.2×10^6 steps. The first 3×10^5 models were discarded as burn-in steps, only after which the sampling algorithm was judged to have converged. Then, every 500th model visited was selected for the ensemble.

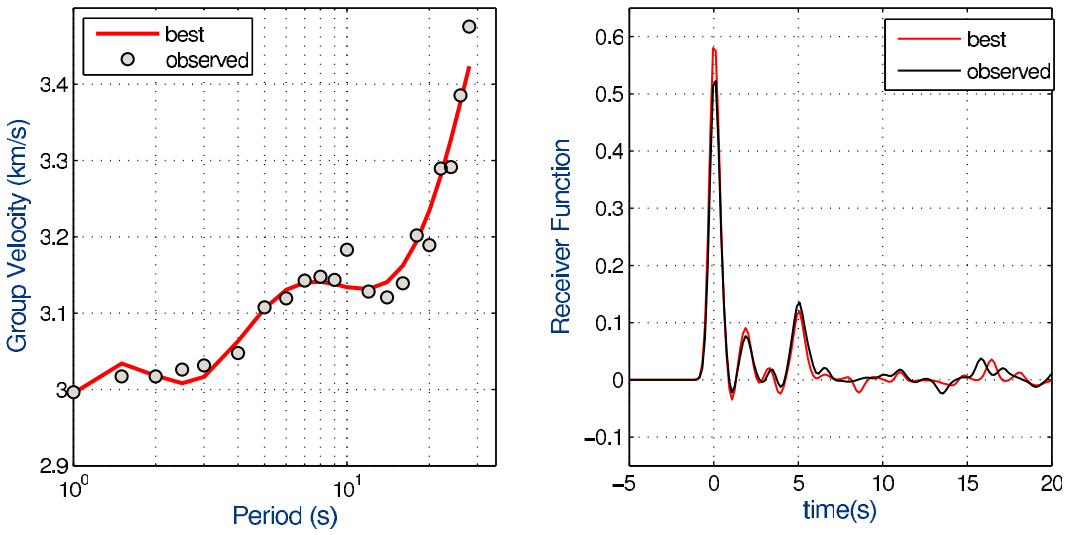

Figure 11. Observed data for station S3A6. (left) Fundamental model Rayleigh wave group velocity dispersion measurements. The red curve shows the dispersion curve obtained with the model that best fits these data in the ensemble solution. (right) Receiver function calculated with the method of *Tkalčić et al.* [2011]. The red curve shows the RF calculated with the model that best fits the observed RF in the ensemble solution.

[83] The data noise covariance matrix $\mathbf{C}_e$ in (12) was parameterized with 2 hyper-parameters σ^{RF} and σ^{SWD}, both treated as unknowns in the inversion, while correlation parameters r^{RF} and r^{SWD} were kept fixed to predefined values. As shown above, dispersion data were considered as independent measurements, and hence $r^{SWD} = 0$, although we acknowledge that measurements close in frequency could have correlated errors. Since P-phase waveforms were filtered during the RF construction process with a Gaussian filter (with $a = 2.5$), we used the Gaussian correlation function in (10) to model RF uncertainties. Hence, the correlation parameter r^{RF} was kept fixed to a value corresponding to a white noise that has been filtered with a Gaussian filter with parameter a. After some elementary calculus, this yields $r^{RF} = f_s/a$, where f_s is the sampling frequency in the RF (O. Gudmundsson, personal communication, 2010).

[84] Note that using the exponential correlation law in (9) to model RF uncertainties would erroneously assume high frequency components in the noise which have obviously been cut by the Gaussian filter (see Appendix D). Alternatively, we could have constructed RFs using exponential filters.

[85] As in our synthetic experiments, independent uniform prior distributions were used for each model parameter with bounds set to 2–50 for the number of layers, 2.5–4.5 km/s for S-wave velocity values, 0–60 km for depth of Voronoi nuclei (see Figure 1), 0–0.12 1/s for σ^{RF}, and 0–0.4 km/s for σ^{SWD}.

[86] The results obtained are shown in Figure 12. Almost all the computed S-wave velocity models are characterized by a very low velocity uppermost structure in the first kilometer of the crust. These low values are interpreted to be related to the presence at surface of either unconsolidated sediments or weathered exposed rocks. The upper-crust (0–15 km) shows complex structures and is characterized by the presence of velocity inversions. Indeed, Figure 12 (right) shows a large number of discontinuities in the first 15 km. The histogram of the location of transitions in the ensemble solution shows that the adaptive parameterization procedure automatically chose models with a number of thin shallow layers lying above thick deep layers. We suspect that this large number of inferred shallow discontinuities might be an indication of an inappropriate data noise parameterization. That is, by assuming the magnitude of data noise is constant along the time series, we give more weight to high amplitude signals (2 s and 5 s peaks in Figure 11), which are sensitive to shallow structures, and less weight to signals with a smaller amplitude, sensitive to deeper structures.

[87] The S-wave velocity in the second part of the crust (15–40 km) is relatively constant with a mean $Vs = 3.75$ km/s, although the inversion models also show some small velocity steps at these depths. It is difficult to identify a relatively sharp crust-mantle transition. The Moho is, rather, characterized as a gradient transition zone over a depth range of 40–45 km, which is consistent with the fact that S3A6 is located in the mountainous region like the Lachlan Fold Belt. At depths below 40 km, the posterior distribution on shear wave velocity is clearly bimodal with two modes at Vs = 4 km/s and Vs = 4.5 km/s. In this case, a typical feature is that the maximum model "jumps" from one mode of the distribution to another, resulting in a sharp discontinuity at 47 km. However, as can be seen in Figure 12 (right), this discontinuity is not required by the data, but rather results from the mode being an unstable measure of a bimodal distribution. Therefore, when interpreting results, the user needs to bear in mind that the information about Shear wave velocity is represented by a full probability distribution, and that the maximum model and average model are only properties of this distribution. The inferred seismic models are aligned with the results of other studies (Tkalčić et al., submitted manuscript, 2011). Although our analysis of the geological implications is rather limited, the aim here is to show that the ensemble solution produced by the transdimensional approach is in good agreement with the main geological features beneath the receiver.

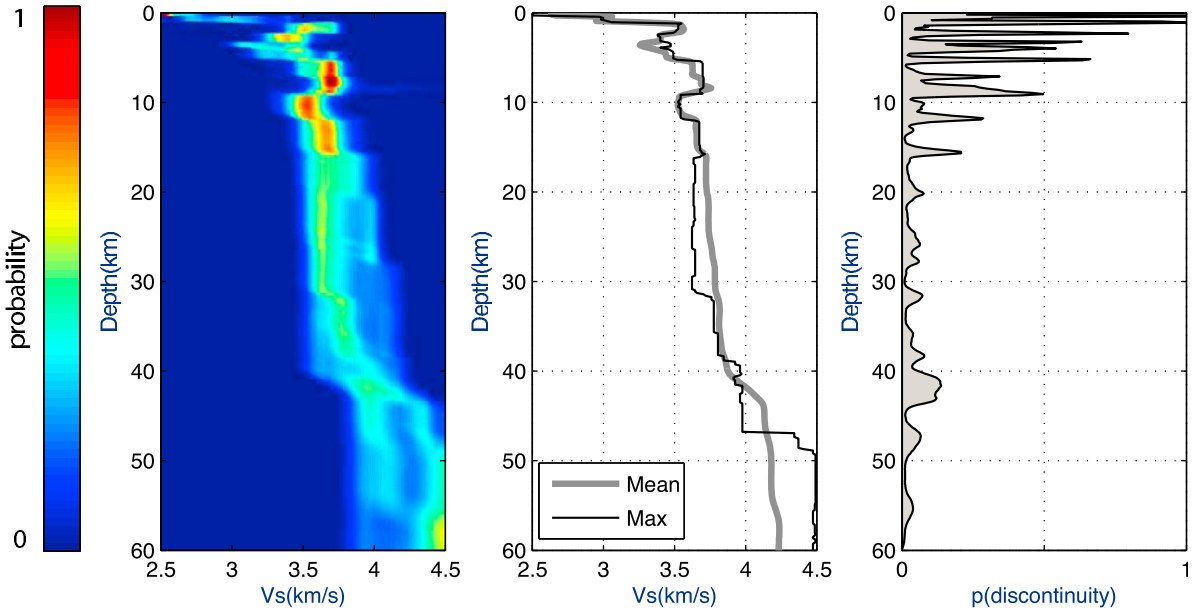

Figure 12. Joint inversion of field data for station S3A6. (left) The ensemble solution, which contains 5.2×10^5 models, is fully represented with a color density plot. This gives an estimation of the Posterior probability distribution for Vs at each depth. (middle) Thick grey line shows the posterior mean model (or average solution) constructed by taking the average Vs at each depth across the ensemble solution. The thin black line shows the maximum of marginal posterior model (or maximum solution) which follows the maximum of the distribution on Vs with depth. (right) Posterior probability for the position of discontinuities constructed from the histogram of transition depths in the ensemble solution.

[88] In order to emphasize the importance of having two independent geophysical data sets, we compare in Figure 13 results obtained after inverting the two data sets separately and jointly. In this way the joint solution in Figure 13 is equal to Figure 12 (left), but since SWD data are measured for periods ranging from 1 to 28 s, they only constrain shallow structure, and hence we only show results down to 25 km depth.

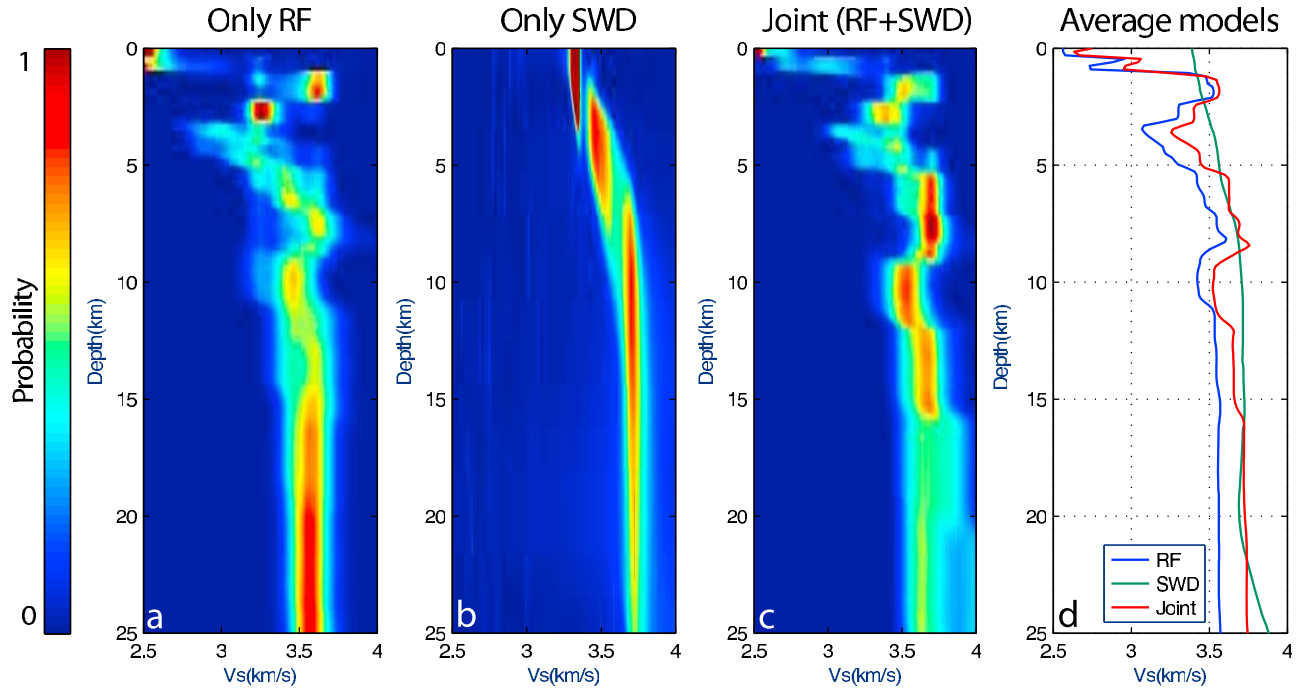

Figure 13. (a–c) Posterior probability distribution for Vs at each depth. Results are shown for individual and joint inversion of data sets in Figure 11. (d) The average solution model for the three inversions carried out.

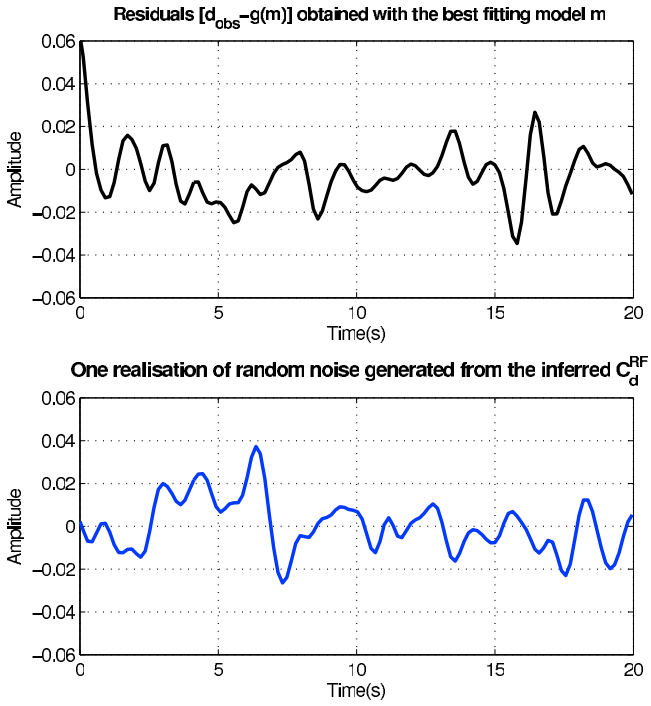

Figure 14. Comparison between residuals and estimated noise for RF data in the portion following the P-pulse for the Joint inversion. (top) Residual waveform (observed-predicted) for the best fitting model. (bottom) Example of a random noise realization generated from the expected hyperparameter σ^{RF} and the fixed parameter r^{RF}. From visual inspection, the two signals present the same variance and smoothness, which indicates that the magnitude of data noise has been estimated correctly, and that a reasonable input r value has been chosen.

[89] As is well known, RFs are poor at constraining absolute velocities. The predictions made by RF alone significantly underestimate the S-wave velocities, as revealed by joint inversion (see the average solution models in Figure 13d). Figure 13b shows the constraints given by SWD alone. The interpretation in this case is limited to the absolute velocity, without any indication where the discontinuities in elastic properties are. When adding SWD to RF (Figure 13c), the velocity profiles tend to change in terms of absolute velocity to accommodate the dispersion data, but without significant alterations in their overall shape as a function of depth, although we have seen in synthetic experiments that even this is possible due to Vs/depth tradeoff (see Figure 10).

[90] Similarly to synthetic examples in Figure 10, here the inferred uncertainties on the 1D velocity model, i.e. the width of the marginal posterior at each depth, is reduced when adding SWD into the problem, especially in the depth range 5–15 km. This quantitatively shows the advantage of simultaneously inverting different data sets. However, note that at shallow depths (2–4 km), the joint inversion seems to increase model uncertainties. This might be due to inconsistency between data sets at these depths, which is accounted for in the inversion as data noise, i.e. inability of a given model to fit the data. Indeed, the volume of Earth sampled by RFs (sensitive to structure directly localized under the station) and SWD data (sensitive to averages over large volumes under and around the station) is quite different.

[91] Since we are using a Gaussian correlation law to model $\mathbf{C}_e^{RF}$, the parameter r^{RF} needs to be fixed at the outset to a predefined value (see Appendix D). To validate this choice, we compare the residual waveform (observed - predicted) for the model that best fits the RF in the ensemble solution, to a realization of a random noise generated from the inferred expected $\mathbf{C}_e^{RF}$. If the choice of r^{RF} is adequate, and if σ^{RF} has been correctly inferred from the data, the noise realization and residuals should have similar properties (i.e. variance and smoothness). This is because the data residuals can be considered a realization of the data errors, i.e. the data noise is defined as the component of the measurements that cannot be explained by $g(\mathbf{m})$. From visual inspection, one can see in Figure 14 that residuals and estimated noise for RF data are similar. Although this is a qualitative test, note that posterior error validation can also be carried out by applying quantitative tests to residuals resulting from one or an ensemble of models [see *Dettmer et al.*, 2007, 2008, 2009, 2010].

5. Conclusion and Future Work

[92] Teleseismic receiver function analysis is now a well-established seismological technique, and a large number of schemes have been implemented in the last 30 years to infer seismic structure beneath broadband stations. In the last 15 years, receiver functions have been inverted together with surface wave dispersion curves and a number of joint inversion algorithms have been proposed. However, a recurring problem is the definition of the misfit function one tries to minimize and particularly the role of the data noise in this function. Here we have presented a novel joint inversion method where a Bayesian formulation is employed to produce a posterior probability distribution, where each model parameter can be described with a full probability density function. While the variance of the posterior can be used to assess uncertainty on model parameters, the posterior covariance directly quantifies the trade-offs (i.e. the correlation) between parameters. Hence the posterior can be examined from several point of views to infer different properties of the model (e.g. depth of transitions, mean Vs value at one depth, number of layers, etc).

[93] The parameterization of the Earth is adaptive and information is extracted from an ensemble of models with variable number of layers. But beyond the transdimensional character of the inversion, an original feature is that little needs to be assumed about the data noise covariance matrix. This matrix plays a crucial role in a Bayesian problem, since it directly determines the form of the posterior. The covariance of data errors represents the magnitude and correlation of data noise, which in the case of RFs and SWD can be difficult to quantify. In the case of a joint inversion, this matrix directly determines the level of information brought by each data set into the solution. Our philosophy is to let the data infer its own degree of uncertainty by treating the magnitude and correlation of noise as unknowns in the problem.

[94] There is a relative freedom in the design of 1D solution profiles needed for interpretation, e.g. average solution model, maximum solution model, best fitting model, etc.

Furthermore, instead of using the whole ensemble of collected models representing the posterior, one can first compute the expected value of hyper-parameters (expected number of cells or expected data noise) and construct a solution profile by only accounting for models that take these hyper-parameters values. This is known in the statistical literature as Empirical Bayes. Alternatively, instead of using expected values of hyper-parameters, one can take as well the mode of the marginal distribution on hyper-parameters.

[95] A potential criticism of our methodology is the computational cost. If the Earth is defined by too many parameters, the number of models needed to sample the posterior distribution becomes large. And since the predicted data have to be computed each time a model is proposed, our algorithm may become computationally prohibitive. Here it is necessary to recognize that, even parallelized and optimized, our method is between 1 and 3 order of magnitude slower than standard linearized inversions. Also, we have here focused on the mathematical problem and illustrated the algorithm in simple situations where a number of approximations have been made. We only inverted for S-wave velocity structure while considering Vp/Vs ratio constant throughout the velocity model. An obvious improvement of the algorithm would be, at increased computational cost, to also consider Vp/Vs ratio in each layer. In addition, layers have been assumed homogeneous and horizontal and it would be possible to treat anisotropy, slope of discontinuities, and lateral variations as unknowns in the problem. These improvements could be achieved by using densely spaced arrays, by including earthquake waveforms from a wide range of back-azimuth, and using more sophisticated forward solvers.

[96] The approach presented in this paper is a general joint inversion strategy, and it has a wide range of possible applications in geosciences (given that the model space is not too large). The method provides a solution model that can exhibit at the same time low gradients and sharp discontinuities. In this sense it appears to have a considerable potential in Earth sciences, given the dual continuous/discrete nature of the Earth. Geophysical inverse problems that seem appropriate here include resistivity surveying with vertical electric sounding or electromagnetic surveys (EM) [*Lowrie*, 1997], inversion of frequency-domain airborne electromagnetic (AEM) data [e.g., *Brodie and Sambridge*, 2006, 2009], potential fields studies [*Luo*, 2010], or seismic cross-hole tomography [*Nicollin et al.*, 2008]. From a statistical inversion point of view, the main difference between different geophysical inverse applications lies in the description of the forward problem. Since the method is based on a direct parameter search algorithm, any forward problem can be conveniently inverted. Hence the ideas here are relevant to other geophysical inverse problems, where depth/time profiles or indeed any 1-D functions are sought.

Appendix A: The Prior

[97] Since we have independent parameters of different physical dimension, in this work we only consider priors that are separable and hence can be written as a product of independent 1D priors on each variable. In some cases one might want to introduce joint priors on a subset of the variables, for example by making the prior variance of velocity in each layer dependent on the depth of the layer. In principle this could be done with additional calculations, but the algorithm would then be more difficult to implement and more computationally expensive. Here, the prior probability distributions is separated into three terms:

$$p(\mathbf{m}) = p(\mathbf{c}, \mathbf{v} \mid k)p(k)p(\mathbf{h}), \tag{A1}$$

where $p(k)$ is the prior on the number of layers, and $p(\mathbf{h})$ is the prior on noise hyper-parameters.

[98] We choose for $p(k)$ a uniform distribution over the interval $I = \{k \epsilon N \mid k_{\min} < k \leq k_{\max}\}$. Hence,

$$p(k) = \begin{cases} 1/(\Delta k) & \text{if} \quad k \in I \\ 0 & \text{otherwise,} \end{cases} \tag{A2}$$

where $\Delta k = (k_{\max} - k_{\min})$. Given a number of cells k, the prior probability distributions for the model parameters are independent from each other, and so can be written in separable form:

$$p(\mathbf{c}, \mathbf{v} \mid k) = p(\mathbf{c} \mid k)p(\mathbf{v} \mid k). \tag{A3}$$

Even though in the prior the parameterization variables $\mathbf{c}$ are independent of the velocity variables $\mathbf{v}$, this will not be the case once the data are introduced, and hence we expect significant correlation in the posterior distribution as shown in Figure 7.

[99] For velocity, the prior for the velocity v_i in layer i is specified by a constant value over a defined interval $J = \{v \epsilon \Re | V_{\min} < v < V_{\max}\}$. Hence we have

$$p(v_i \mid k) = \begin{cases} 1/(\Delta v) & \text{if} \quad v_i \in J \\ 0 & \text{otherwise,} \end{cases} \tag{A4}$$

where $\Delta v = (V_{\max} - V_{\min})$. Since the velocity in each layer is considered independent,

$$p(\mathbf{v} \mid k) = \prod_{i=1}^{k} p(v_i \mid k). \tag{A5}$$

As in the work of *Bodin and Sambridge* [2009], for mathematical convenience, let us for the moment assume that the Voronoi nuclei c_i can only take place on an underlying grid of finite points defined by N possible depths. For k Voronoi nuclei, there are $\frac{N!}{k!(N-k)!}$ possible configurations on the N possible depths of the underlying grid. We give equal probability to each of these configurations. Hence,

$$p(\mathbf{c} \mid k) = \left[\frac{N!}{k!(N-k)!}\right]^{-1}. \tag{A6}$$

Given a set of hyper-parameters $\mathbf{h}$, the prior for each hyper-parameter h_j is specified by a uniform distribution over a defined interval $H_j = \{h \epsilon \Re \mid h^j_{\min} < h < h^j_{\max}\}$. Hence we have

$$p(h_j) = \begin{cases} 1/(\Delta^j h) & \text{if} \quad h_j \in H_j \\ 0 & \text{otherwise,} \end{cases} \tag{A7}$$

where $\Delta^j h = (h^j_{\max} - h^j_{\min})$. Since we consider each hyper-parameter independent in the prior,

$$p(\mathbf{h}) = \prod_{j=1}^{m} p(h_j), \tag{A8}$$

where m is the number of hyper-parameters defining the data noise covariance matrix $\mathbf{C}_e$. Therefore, after substituting (A5), (A6), and (A8) into (A1), the full prior probability density function can be expressed as

$$p(\mathbf{m}) = \frac{k!(N-k)!}{\Delta k N! (\Delta v)^k \prod_{j=1}^{m} [\Delta^j h]}, \tag{A9}$$

given that all parameters in $\mathbf{m}$ fall into the range defined by their respective prior distribution. If at least one parameter falls outside the defined boundaries, the full prior becomes null.

Appendix B: Proposal Distributions

[100] Having randomly initialized the model parameters by drawing values from the prior distribution of each parameter, the algorithm proceeds iteratively. At each iteration of the chain, we propose a new model by drawing from a probability distribution $q(\mathbf{m}'|\mathbf{m})$ such that the new proposed model $\mathbf{m}'$ is conditional only on the current model $\mathbf{m}$. At each iteration of the reversible jump algorithm, one type of move is uniformly randomly selected from the five following possibilities:

[101] 1. Change the velocity in one layer. Randomly select a layer i from a uniform distribution over [1, k] and randomly propose a new value v_i' using a Gaussian probability distribution centered at the current value v_i:

$$q_{v1}(v_i' \mid v_i) = \frac{1}{\theta_1\sqrt{2\pi}} \exp\left\{ -\frac{(v_i' - v_i)^2}{2\theta_1^2} \right\}. \tag{B1}$$

The variance θ_1^2 of the Gaussian function is a parameter to be chosen. Hence we have

$$v_i' = v_i + u, \tag{B2}$$

where u is a random deviate from a normal distribution $N(0,\theta_1)$. All the other model parameters are kept constant, and hence this proposal does not involve a change in dimension.

[102] 2. Birth: create a new layer. Add a new Voronoi center with the position c_{k+1}' found by choosing uniformly randomly a point from the underlying grid that is not already occupied. There are $(N-k)$ discrete points available. Then, a new velocity value v_{k+1}' needs to be assigned to the new layer. This is drawn from a Gaussian proposal probability density with the same form as (B1):

$$q_{v2}(v_{k+1}' \mid v_i) = \frac{1}{\theta_2\sqrt{2\pi}} \exp\left\{ -\frac{(v_{k+1}' - v_i)^2}{2\theta_2^2} \right\}, \tag{B3}$$

where v_i is the current velocity value at the depth c_{k+1}' where the birth takes place. The variance θ_2^2 of the Gaussian function is a parameter to be chosen.

[103] 3. Death: Remove at random one layer by drawing a number from a uniform distribution over the range [1, k]. The response values of the neighboring cells remain unchanged.

[104] 4. Move: Randomly pick one layer (from a uniform distribution) and randomly change the position of its nucleus according to

$$q_c(c_i' \mid c_i) = \frac{1}{\theta_3\sqrt{2\pi}} \exp\left\{ -\frac{(c_i' - c_i)^2}{2\theta_3^2} \right\}. \tag{B4}$$

[105] 5. Change the estimated data noise. Randomly select a hyper-parameter j from a uniform distribution over the range [1, m]. Propose a new value h_j' using

$$q_{h'}(h_j' \mid h_j) = \frac{1}{\theta_{h'}\sqrt{2\pi}} \exp\left\{ -\frac{(h_j' - h_j)^2}{2\theta_{h'}^2} \right\}. \tag{B5}$$

[106] Note that the proposed model can fall outside the range defined by the uniform prior distribution. In this case, the prior (and hence the posterior) of the proposed model is null, and the model is rejected. In this way the proposal distributions are seen as true Gaussians rather than truncated versions.

[107] The standard deviations (θ_1, θ_2, θ_3, $\theta_{h'}$) of the Gaussian proposal functions are parameters to be fixed by the user. As shown by *MacKay* [2003], the magnitude of perturbations does not affect the solution but rather the sampling efficiency of the algorithm. Thus the width of proposal distributions are tuned by trial-and-error in order to have an acceptance rate as close to 44% for each type of perturbation [*Rosenthal*, 2000].

Appendix C: The Acceptance Probability

[108] Once a proposed model has been drawn from the distribution $q(\mathbf{m}'|\mathbf{m})$, the new model is then accepted with a probability $\alpha(\mathbf{m}'|\mathbf{m})$, i.e. a uniform random deviate, r, is generated between 0 and 1. If $r \leq \alpha$, the move is accepted, the current model $\mathbf{m}$ is replaced with $\mathbf{m}'$ and the chain moves to the next step. If $r > \alpha$, the move is rejected and the current model is retained for the next step of the chain where the process is repeated. The acceptance probability, $\alpha(\mathbf{m}'|\mathbf{m})$, is the key to ensuring that the samples will be generated according to the target density $p(\mathbf{m}|\mathbf{d}_{obs})$. It can be shown [*Green*, 1995, 2003] that the chain of sampled models will converge to the transdimensional posterior distribution, $p(\mathbf{m}|\mathbf{d}_{obs})$, if

$$\alpha(\mathbf{m}' \mid \mathbf{m}) = \min\left[1, \frac{p(\mathbf{m}')}{p(\mathbf{m})} \cdot \frac{p(\mathbf{d}_{obs} \mid \mathbf{m}')}{p(\mathbf{d}_{obs} \mid \mathbf{m})} \cdot \frac{q(\mathbf{m} \mid \mathbf{m}')}{q(\mathbf{m}' \mid \mathbf{m})} \cdot |\mathbf{J}|\right], \tag{C1}$$

where the matrix $\mathbf{J}$ is the Jacobian of the transformation from $\mathbf{m}$ to $\mathbf{m}'$ and is needed to account for the scale changes involved when the transformation involves a jump between dimensions [*Green*, 2003].The expression for $\alpha(\mathbf{m}'|\mathbf{m})$ involves the ratio of the posterior distribution evaluated at the proposed model, $\mathbf{m}'$ to the current model $\mathbf{m}$ multiplied by the ratio of the proposal distribution for the reverse step, $q(\mathbf{m} \mid \mathbf{m}')$, to the forward step, $q(\mathbf{m}' \mid \mathbf{m})$.

C1. Proposal Ratios

[109] The proposal ratio of forward and reverse moves needs to be calculated so that the acceptance probability in (C1) can be calculated in each case. For the proposal types that do not involve a change of dimension the distributions are symmetrical. That is, the probability to go from $\mathbf{m}$ to $\mathbf{m}'$ is equal to the probability to go from $\mathbf{m}'$ to $\mathbf{m}$. Hence

$$\begin{aligned} q_c(c_i' \mid c_i) &= q_c(c_i \mid c_i') \\ q_{h'}(h_j' \mid h_j) &= q_{h'}(h_j \mid h_j') \\ q_v(v_i' \mid v_i) &= q_v(v_i \mid v_i') \end{aligned} \tag{C2}$$

and in all three cases the proposal ratio equals one.

$$\frac{q(\mathbf{m} \mid \mathbf{m}')}{q(\mathbf{m}' \mid \mathbf{m})} = 1. \tag{C3}$$

[110] For a birth step, the algorithm jumps between a model $\mathbf{m}$ with k layers to a model $\mathbf{m}'$ with $(k + 1)$ layers. Since the new nucleus c'_{k+1} is generated independently from the new velocity value v'_{k+1} then proposal distribution can be separated and we write

$$\frac{q(\mathbf{m} \mid \mathbf{m}')}{q(\mathbf{m}' \mid \mathbf{m})} = \frac{q(\mathbf{c} \mid \mathbf{m}')}{q(\mathbf{c}' \mid \mathbf{m})} \cdot \frac{q(\mathbf{v} \mid \mathbf{m}')}{q(\mathbf{v}' \mid \mathbf{m})}. \tag{C4}$$

Specifically we have the probability of a birth at position c'_{k+1} which is given by

$$q(\mathbf{c}' \mid \mathbf{m}) = 1/(N - k), \tag{C5}$$

the probability of generating the new velocity value v'_{k+1} is given by

$$q(v' \mid \mathbf{m}) = q_{v2}(v'_{k+1} \mid v_i), \tag{C6}$$

the probability of deleting the cell at position c'_{k+1} (reverse step)

$$q(\mathbf{c} \mid \mathbf{m}') = 1/(k + 1), \tag{C7}$$

and the probability of removing a velocity when cell is deleted (reverse step)

$$q(\mathbf{v} \mid \mathbf{m}') = 1. \tag{C8}$$

Substituting these expressions in (C4) we obtain

$$\left(\frac{q(\mathbf{m} \mid \mathbf{m}')}{q(\mathbf{m}' \mid \mathbf{m})}\right)_{birth} = \frac{(N - k)}{(k + 1) q_{v2}(v'_{k+1} \mid v_i)}. \tag{C9}$$

[111] For the death of a randomly chosen nucleus, we move from k to $(k-1)$ cells. Suppose that nucleus c_i, with velocity v_i, is removed. In this case, a similar reasoning to the birth case above leads us to a proposal ratio (reverse to forward) of

$$\left(\frac{q(\mathbf{m} \mid \mathbf{m}')}{q(\mathbf{m}' \mid \mathbf{m})}\right)_{death} = \frac{k q_{v2}(v'_j \mid v_i)}{(N - k + 1)}, \tag{C10}$$

where v'_j is the velocity at depth c_i in the new structure, $\mathbf{c}'$, after removal of the ith layer.

C2. The Jacobian

[112] The Jacobian term "normalizes" the difference in volume between two spaces of different dimension. In our case, the Jacobian only needs to be calculated when there is a jump between two models of different dimensions, i.e. when a birth or death is proposed [*Green*, 1995]. If the current and proposed model have the same dimension, the Jacobian term is 1, and can be ignored.

[113] For a birth step, the bijective transformation used to go from $\mathbf{m}$ to $\mathbf{m}'$ can be written as

$$\mathbf{m} = (\mathbf{c}, \mathbf{v}, u_c, u_v) \longleftrightarrow (\mathbf{c}, \mathbf{v}, c'_{k+1}, v'_{k+1}) = \mathbf{m}'. \tag{C11}$$

The random variable u_c used to propose a new nucleus c'_{k+1} is drawn from a discrete distribution defined on the integers $[0, 1, \ldots, N - k]$. The random number u_v is drawn from the Gaussian distribution centered at 0 and the velocity assigned to the new cell is given by

$$v'_{k+1} = v_i + u_v, \tag{C12}$$

where v_i is the current velocity value where the birth takes place.

[114] Note that the model space is divided into a discrete space (nuclei position) and a continuous space (velocities). u_c is a discrete variable used for the transformation between discrete spaces and u_v is a continuous variable used for the transformation between continuous spaces. *Denison et al.* [2002] showed that the Jacobian term is always unity for discrete transformations. Therefore, the Jacobian term only accounts for the change in variables from

$$(\mathbf{v}, u_v) \longleftrightarrow (\mathbf{v}, v'_{k+1}) = \mathbf{v}'. \tag{C13}$$

Hence, here the Jacobian term is the determinant of the matrix of all first-order partial derivatives of the vector $\mathbf{v}'$ with respect to $(\mathbf{v},u_v)$, and we have

$$|\mathbf{J}|_{birth} = \left|\frac{\delta(\mathbf{v}, v_{k+1})}{\delta(\mathbf{v}, u_v)}\right| = \left|\frac{\delta(v_i, v_{k+1})}{\delta(v_i, u_v)}\right| = \begin{vmatrix} 1 & 0 \\ 1 & 1 \end{vmatrix} = 1. \tag{C14}$$

[115] So it turns out that for this style of birth proposal the Jacobian is also unity. Since the Jacobian for a death move is $|\mathbf{J}|_{death} = |\mathbf{J}^{-1}|_{birth}$, this is also equal to one. Conveniently, the Jacobian is unity for each case and can be ignored.

C3. The Acceptance Term

[116] We now substitute expressions for each proposal ratio into (C1) to get final expressions for the acceptance probability in each of the 5 possible moves described earlier. For the moves that do not include a change in dimension, we have seen that the proposal ratio becomes unity. Hence the acceptance term is simply given by the ratio of the posteriors

$$\alpha(\mathbf{m}' \mid \mathbf{m}) = \min\left[1, \frac{p(\mathbf{m}')}{p(\mathbf{m})} \cdot \frac{p(\mathbf{d}_{obs} \mid \mathbf{m}')}{p(\mathbf{d}_{obs} \mid \mathbf{m})}\right]. \tag{C15}$$

Since the dimension of the model does not change, according to (A9), the prior ratio is either null or unity. If one of the proposed parameter falls outside the bounds defined by the prior, the prior ratio is null and $\alpha(\mathbf{m}',\mathbf{m}) = 0$. Otherwise,

$$\alpha(\mathbf{m}', \mathbf{m}) = \min\left[1, \frac{p(\mathbf{d}_{obs} \mid \mathbf{m}')}{p(\mathbf{d}_{obs} \mid \mathbf{m})}\right]. \tag{C16}$$

For changes in velocity value and nuclei positions, we have

$$\alpha(\mathbf{m}', \mathbf{m}) = \min\left[1, \exp\left\{-\frac{\Phi(\mathbf{m}') - \Phi(\mathbf{m})}{2}\right\}\right]. \tag{C17}$$

When perturbing the data noise hyper-parameters $\mathbf{h}$ that define the data noise covariance matrix $\mathbf{C}_e$, the normalizing constant in the likelihood is changed and the ratio of determinants needs to be taken into account (see Appendix D).

$$\alpha(\mathbf{m}', \mathbf{m}) = \min\left[1, \frac{|\mathbf{C}_e|}{|\mathbf{C}'_e|} \exp\left\{-\frac{\Phi(\mathbf{m}') - \Phi(\mathbf{m})}{2}\right\}\right]. \tag{C18}$$

Note that $\Phi(\mathbf{m}')$ and $\Phi(\mathbf{m})$ also incorporate $\mathbf{C}'_e$ and $\mathbf{C}_e$, respectively.

[117] For a birth step, according to (A9), and provided the perturbed parameter falls into the bounds of the prior, the prior ratio takes the form

$$\left(\frac{p(\mathbf{m}')}{p(\mathbf{m})}\right)_{birth} = \frac{k+1}{(N-k)\Delta v}. \tag{C19}$$

After substituting (5), (C9), and (C19) into (C1), the acceptance term for the birth step reduces to

$$\alpha(\mathbf{m}', \mathbf{m}) = \min\left[1, \frac{\theta_2\sqrt{2\pi}}{\Delta v}.\exp\left\{\frac{(v'_{k+1} - v_i)^2}{2\theta_2^2} - \frac{\Phi(\mathbf{m}') - \Phi(\mathbf{m})}{2}\right\}\right], \tag{C20}$$

where i indicates the layer in the current tessellation $\mathbf{c}$ that contains the depth c'_{k+1} where the birth takes place. Again, if the perturbed parameters fall outside the bounds of the prior, then $\alpha(\mathbf{m}',\mathbf{m}) = 0$, and the move is rejected. For the birth step then we see the acceptance probability is a balance between the proposal probability (which encourages velocities to change) and the difference in data misfit which penalizes velocities if they change so much that they degrade fit to data.

[118] For the death step, the prior ratio in (C19) must be inverted. After substituting this with (5) and (C10) into (C1), and after simplification we get the acceptance probability

$$\alpha(\mathbf{m}', \mathbf{m}) = \min\left[1, \frac{\Delta v}{\theta_2\sqrt{2\pi}}.\exp\left\{-\frac{(v'_j - v_i)^2}{2\theta_2^2} - \frac{(\Phi(\mathbf{m}') - \Phi(\mathbf{m}))}{2}\right\}\right], \tag{C21}$$

where i indicates the layer that we remove from the current tessellation $\mathbf{c}$ and j indicates the cell in the proposed tessellation $\mathbf{c}'$ that contains the deleted point c_i. Unsurprisingly the death acceptance probability has a similar form to that of the birth, with proposal and data terms opposing each other.

[119] We see from these expressions that the variable N, i.e. the number of candidate positions for the nuclei, vanishes. This means that there is no need to use an actual discrete grid in generating nuclei positions. In fact it was only ever a mathematical convenience which ensures that the acceptance expressions have the correct analytic form. In practice we are at liberty to generate the nuclei using a continuous distribution over the region of the model (which is tantamount to $N \to \infty$).

Appendix D: Parameterizing $\mathbf{C}_e$

D1. First Type of Noise Parameterization

[120] The correlation function in (8) is assumed to decay exponentially and is thus given by $c_i = r^i$, where $r = c_1$ is a constant number between 0 and 1 which describes the correlation between two adjacent samples in the time series. This correlation function is plotted in Figure D1a for different values of r, and realizations of noise with such a correlation are shown in Figures D1c, D1e, and D1g. In this case, the data noise covariance matrix in (7) writes:

$$\mathbf{C}_e = \sigma^2 \begin{bmatrix} 1 & r & r^2 & \dots & r^{n-1} \\ r & 1 & r & \dots & r^{n-2} \\ r^2 & r & 1 & \dots & r^{n-3} \\ & & & \vdots & \\ r^{n-1} & r^{n-2} & r^{n-3} & \dots & 1 \end{bmatrix}, \tag{D1}$$

where n is the number of data points.

[121] Hence, with this type of noise parameterization, the two hyper-parameters $\mathbf{h} = [\sigma, r]$ describing the noise covariance can be given a wide uniform prior probability distribution and posterior inference can be done to infer the magnitude and correlation of data noise. The two noise parameters are perturbed along the transdimensional Markov chain and each time a new value is proposed, $\mathbf{C}_e^{-1}$ and $|\mathbf{C}_e|$ will be perturbed accordingly to compute the likelihood value of the proposed model in (5).

[122] It can be easily shown with linear algebra that the inverse of $\mathbf{C}_e$ is a symmetric tridiagonal matrix:

$$\mathbf{C}_e^{-1} = \frac{1}{\sigma^2(1-r^2)} \begin{bmatrix} 1 & -r & 0 & \dots & 0 & 0 \\ -r & 1+r^2 & -r & \dots & 0 & 0 \\ 0 & -r & 1+r^2 & \dots & 0 & 0 \\ & & & \vdots & & \\ 0 & 0 & 0 & \dots & 1+r^2 & -r \\ 0 & 0 & 0 & \dots & -r & 1 \end{bmatrix}. \tag{D2}$$

See *Malinverno and Briggs* [2004] for a detailed demonstration (note that when $r = 1$ all the elements of $\mathbf{C}_e$ are one and the inverse does not exist). The inverse data noise covariance matrix requires storage that is proportional to n, while computing the Mahalanobis distance in (4) only requires order n operations. The likelihood in (5) also needs the determinant of $\mathbf{C}_e$. As shown by *Malinverno and Briggs* [2004], an expression of this determinant can be obtained by writing the tridiagonal inverse covariance matrix $\mathbf{C}_e^{-1} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is a lower triangular matrix whose determinant is the product of its diagonal elements. The final result for the determinant of the data noise covariance matrix is

$$|\mathbf{C}_e| = \sigma^{2n}\left(1-r^2\right)^{n-1}. \tag{D3}$$

[123] With this form of correlation, the data noise covariance matrix is well-conditioned and there are stable analytical solutions for $\mathbf{C}_e^{-1}$ and $|\mathbf{C}_e|$. However, as shown below, this is usually not the case if we use other forms of correlation function. Here the main advantage is that each time we want to perturb r or σ along the random walk, we directly perturb the determinant and inverse without having to numerically compute them from a perturbed $\mathbf{C}_e$, which would be too computationally expensive.

D2. Second Type of Noise Parameterization

[124] The data noise covariance matrix can be also parameterized with a Gaussian correlation law $c_i = r^{(i^2)}$. Here r can be written as $r = e^{-1/2\rho^2}$ where ρ^2 is the variance of a

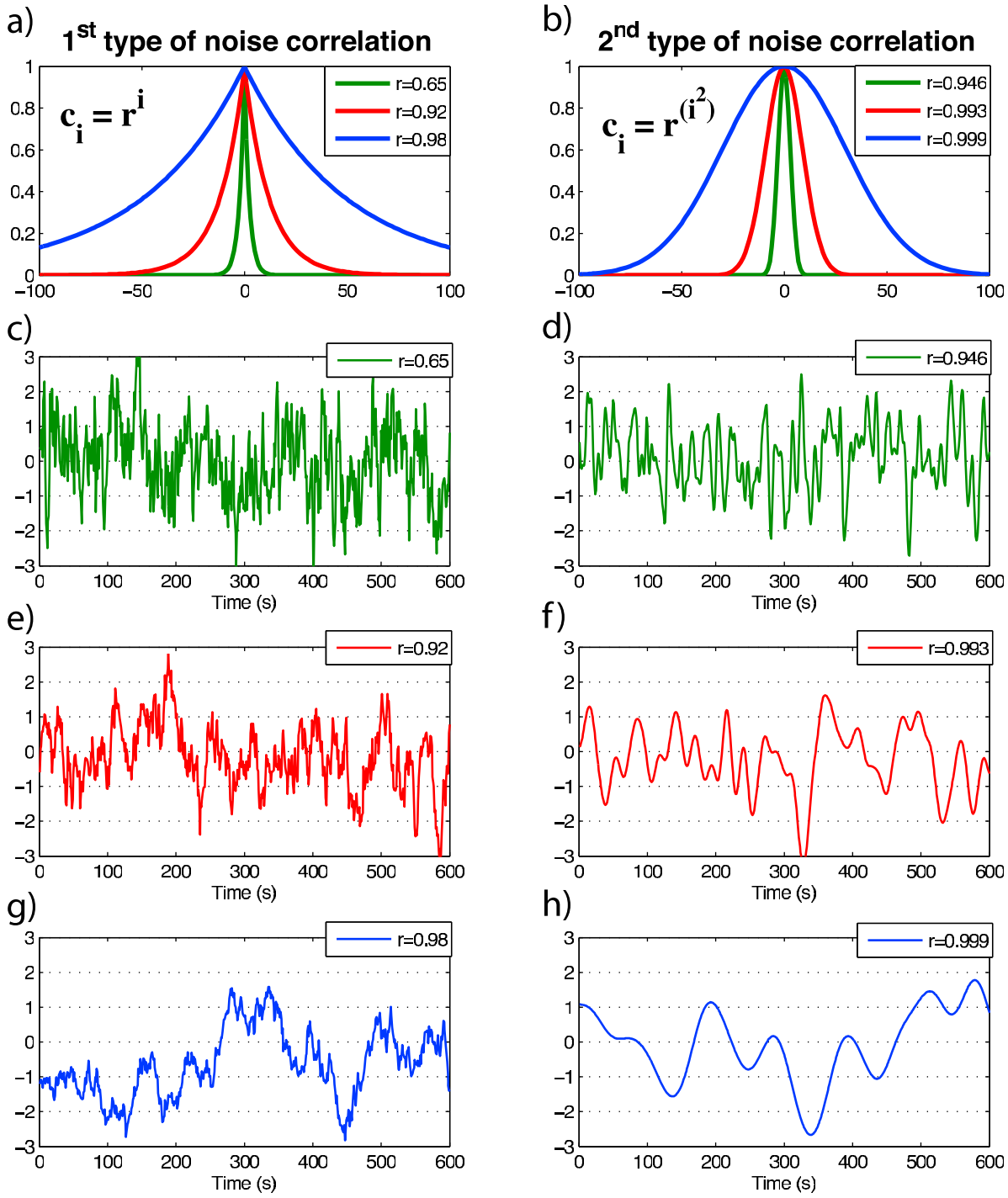

Figure D1. (a and b) The two types of correlation functions c_i for different values of r. These symmetric functions represent the correlation between a point in the time series and its neighbors. In order to compare the two forms of noise assumed in this study, we plot different realizations of noise for different values of r and a fixed standard deviation $\sigma = 1$. (c, e, and g) Different realizations with the first type of correlation (exponential decay), whereas (d, f, and h) noise vectors are generated with a Gaussian correlation. The second type of noise in Figures D1b, D1d, D1f, and D1h seem to be closer to what is observed on RF before the first arrival, however this way of parameterizing the noise turns out to be more difficult to implement for an Hierarchical Bayes inversion.

Gaussian correlation function, which is shown in Figure D1b. Realizations of such a noise for different values of r are shown in Figures D1d, D1f, and D1h. Note that a white noise which has been convolved with a Gaussian filter will have exactly the same structure as our second type of noise.

[125] Although this form of correlation clearly appears more relevant for our problem, it turns out that the associated

data noise covariance matrix in (7) is highly ill-conditioned, and hence there are no stable analytical formulation for its inverse and determinant. Therefore $\mathbf{C}_e^{-1}$ and $|\mathbf{C}_e|$ have to be numerically computed with SVD decomposition and removal of a large number of small eigenvalues that destabilize the process. Unfortunately an SVD decomposition of a $n \times n$ matrix is computationally expensive and cannot be carried out each time $\mathbf{C}_e$ is perturbed along the random walk. As a result, the correlation r need to be fixed and cannot be treated as an unknown in the inversion. However, the magnitude of data noise σ^2 can be perturbed without having to re-invert each time $\mathbf{C}_e$. This is because we have

$$\mathbf{C}_e^{-1} = \left(\sigma^2 \mathbf{R}\right)^{-1} = \frac{1}{\sigma^2}\mathbf{R}^{-1} \tag{D4}$$

$$|\mathbf{C}_e| = |\sigma^2 \mathbf{R}| = |\sigma^2 \mathbf{I}_d| \times |\mathbf{R}| = \sigma^{2n}|\mathbf{R}|\,. \tag{D5}$$

In this way $\mathbf{R}^{-1}$ is computed once at the beginning and remains fixed along the Markov chain. The variance of data noise σ^2 can be treated as an unknown since each time a new value is proposed, $\mathbf{C}_e^{-1}$ and $|\mathbf{C}_e|$ can be computed from (D4) and (D5) without having to redo any SVD decomposition.

[126] **Acknowledgments.** This research was supported under Australian Research Council Discovery projects funding scheme (project DP110102098). This project was also supported by French-Australian Science and Technology travel grant (FR090051) under the International Science Linkages program from the Department of Innovation, Industry, Science and Research. Calculations were performed on the Terrawulf II cluster, a computational facility supported through AuScope. Auscope Ltd is funded under the National Collaborative Research Infrastructure Strategy (NCRIS), an Australian Commonwealth Government Programme. Computer software implementing the algorithms described in this paper are available from the authors.

References

Ammon, C. (1992), A comparison of deconvolution techniques, *Rep. UCID-ID-111667*, Lawrence Livermore Natl. Lab., Livermore, Calif.

Ammon, C., G. Randall, and G. Zandt (1990), On the nonuniqueness of receiver function inversions, *J. Geophys. Res.*, *95*, 15,303–15,318.

Arroucau, P., N. Rawlinson, and M. Sambridge (2010), New insight into Cainozoic sedimentary basins and Palaeozoic suture zones in southeast Australia from ambient noise surface wave tomography, *Geophys. Res. Lett.*, *37*, L07303, doi:10.1029/2009GL041974.

Bannister, S., J. Yu, B. Leitner, and B. L. N. Kennett (2003), Variations in crustal structure across the transition from West to East Antarctica, Southern Victoria Land, *Geophys. J. Int.*, *155*(3), 870–880.

Bayes, T. (1763), An essay towards solving a problem in the doctrine of chances, *Philos. Trans. R. Soc. London*, *53*, 370–418. [Reprinted in *Biometrika*, *45*, 295–315, 1958.]

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

Box, G., and G. Tiao (1973), *Bayesian Inference in Statistical Inference*, Wiley-Intersci., Hoboken, N. J.

Brodie, R., and M. Sambridge (2006), A holistic approach to inversion of frequency-domain airborne EM data, *Geophysics*, *71*, G301–G312.

Brodie, R., and M. Sambridge (2009), Holistic inversion of frequency-domain airborne electromagnetic data with minimal prior information, *Explor. Geophys.*, *40*(1), 8–16.

Campillo, M., and A. Paul (2003), Long-range correlations in the diffuse seismic coda, *Science*, *299*(5606), 547–549.

Chang, S., C. Baag, and C. Langston (2004), Joint analysis of teleseismic receiver functions and surface wave dispersion using the genetic algorithm, *Bull. Seismol. Soc. Am.*, *94*(2), 691–704.

Charvin, K., K. Gallagher, G. Hampson, and R. Labourdette (2009a), A Bayesian approach to inverse modelling of stratigraphy, part 1: Method, *Basin Res.*, *21*(1), 5–25.

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

Chen, Y., F. Niu, R. Liu, Z. Huang, H. Tkalčić, L. Sun, and W. Chan (2010), Crustal structure beneath China from receiver function analysis, *J. Geophys. Res.*, *115*, B03307, doi:10.1029/2009JB006386.

Clayton, R., and R. Wiggins (1976), Source shape estimation and deconvolution of teleseismic bodywaves, *Geophys. J. R. Astron. Soc.*, *47*(1), 151–177.

Clifford, P., S. Greenhalgh, G. Houseman, and F. Graeber (2007), 3-D seismic tomography of the Adelaide fold belt, *Geophys. J. Int.*, *172*(1), 167–186.

Clitheroe, G., O. Gudmundsson, and B. Kennett (2000), The crustal thickness of Australia, *J. Geophys. Res.*, *105*, 13,697–13,713.

Denison, D., N. Adams, C. Holmes, and D. Hand (2002), Bayesian partition modelling, *Comput. Stat. Data Anal.*, *38*(4), 475–485.

Dettmer, J., S. Dosso, and C. Holland (2007), Uncertainty estimation in seismo-acoustic reflection travel time inversion, *J. Acoust. Soc. Am.*, *122*, 161–176.

Dettmer, J., S. Dosso, and C. Holland (2008), Joint time/frequency-domain inversion of reflection data for seabed geoacoustic profiles and uncertainties, *J. Acoust. Soc. Am.*, *123*, 1306–1317.

Dettmer, J., C. Holland, and S. Dosso (2009), Analyzing lateral seabed variability with Bayesian inference of seabed reflection data, *J. Acoust. Soc. Am.*, *126*, 56–69.

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

Di Bona, M., et al. (1998), Variance estimate in frequency-domain deconvolution for teleseismic receiver function computation, *Geophys. J. Int.*, *134*(2), 634–646.

Du, Z., and G. Foulger (1999), The crustal structure beneath the northwest fjords, Iceland, from receiver functions and surface waves, *Geophys. J. Int.*, *139*(2), 419–432.

Duijndam, A. (1988a), Bayesian estimation in seismic inversion. Part I: Principles, *Geophys. Prospect.*, *36*(8), 878–898.

Duijndam, A. (1988b), Bayesian estimation in seismic inversion. Part II: Uncertainty analysis, *Geophys. Prospect.*, *36*, 899–918.

Frederiksen, A., H. Folsom, and G. Zandt (2003), Neighbourhood inversion of teleseismic Ps conversions for anisotropy and layer dip, *Geophys. J. Int.*, *155*(1), 200–212.

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 Bayesian Transdimensional changepoint models, *Earth Planet. Sci. Lett.*, *311*, 182–194.

Gelman, A., J. Carlin, H. Stern, and D. Rubin (2004), *Bayesian Data Analysis*, CRC Press, Boca Raton, Fla.

Geyer, C., and J. Møller (1994), Simulation procedures and likelihood inference for spatial point processes, *Scand. J. Stat.*, *21*(4), 359–373.

Gouveia, W., and J. Scales (1998), Bayesian seismic waveform inversion-Parameter estimation and uncertainty analysis, *J. Geophys. Res.*, *103*, 2759–2780.

Graeber, F., G. Houseman, and S. Greenhalgh (2002), Regional teleseismic tomography of the western Lachlan Orogen and the Newer Volcanic Province, southeast Australia, *Geophys. J. Int.*, *149*(2), 249–266.

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.

Haskell, N. (1953), The dispersion of surface waves on multilayered media, *Bull. Seismol. Soc. Am.*, *43*(1), 17–34.

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

Hetényi, G., and Z. Bus (2007), Shear wave velocity and crustal thickness in the Pannonian Basin from receiver function inversions at four permanent stations in Hungary, *J. Seismol.*, *11*(4), 405–414.

Hopcroft, P., K. Gallagher, and C. Pain (2007), Inference of past climate from borehole temperature data using Bayesian Reversible Jump Markov chain Monte Carlo, *Geophys. J. Int.*, *171*(3), 1430–1439.

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.

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.

Juliá, J., C. Ammon, R. Herrmann, and A. Correig (2000), Joint inversion of receiver function and surface wave dispersion observations, *Geophys. J. Int.*, *143*(1), 99–112.

Juliá, J., C. Ammon, and R. Herrmann (2003), Lithospheric structure of the Arabian Shield from the joint inversion of receiver functions and surface-wave group velocities, *Tectonophysics*, *371*(1–4), 1–21.

Kind, R., G. Kosarev, and N. Petersen (1995), Receiver functions at the stations of the German Regional Seismic Network (GRSN), *Geophys. J. Int.*, *121*(1), 191–202.

Kosarev, G., N. Petersen, L. Vinnik, and S. Roecker (1993), Receiver functions for the Tien Shan analog broadband network: Contrasts in the evolution of structures across the Talasso-Fergana fault, *J. Geophys. Res.*, *98*, 4437–4448.

Langston, C. (1979), Structure under Mount Rainier, Washington, inferred from teleseismic body waves, *J. Geophys. Res.*, *84*, 4749–4762.

Lawrence, J., and D. Wiens (2004), Combined receiver-function and surface wave phase-velocity inversion using a niching genetic algorithm: Application to Patagonia, *Bull. Seismol. Soc. Am.*, *94*(3), 977–987.

Levin, V., and J. Park (1997), P-SH conversions in a flat-layered medium with anisotropy of arbitrary orientation, *Geophys. J. Int.*, *131*(2), 253–266.

Ligorria, J., and C. Ammon (1999), Iterative deconvolution and receiver-function estimation, *Bull. Seismol. Soc. Am.*, *89*(5), 1395–1400.

Lombardi, D. (2007), Alpine crustal and upper-mantle structure from receiver functions, A PhD thesis, ETH Zurich, Zurich, Switzerland.

Lowrie, W. (1997), *Fundamentals of Geophysics*, Cambridge Univ. Press, New York.

Lucente, F., N. Piana Agostinetti, M. Moro, G. Selvaggi, and M. Di Bona (2005), Possible fault plane in a seismic gap area of the southern Apennines (Italy) revealed by receiver function analysis, *J. Geophys. Res.*, *110*, B04307, doi:10.1029/2004JB003187.

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, New York.

Mahalanobis, P. (1936), On the generalised distance in statistics, *Proc. Natl. Acad. Sci. India*, *12*, 49–55.

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, doi:10.1190/1.1778243.

Malinverno, A., and R. Parker (2006), Two ways to quantify uncertainty in geophysical inverse problems, *Geophysics*, *71*, W15, doi:10.1190/1.2194516.

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.

Moorkamp, M., A. Jones, and S. Fishwick (2010), Joint inversion of receiver functions, surface wave dispersion, and magnetotelluric data, *J. Geophys. Res.*, *115*, B04318, doi:10.1029/2009JB006369.

Mosegaard, K., and A. Tarantola (1995), Monte Carlo sampling of solutions to inverse problems, *J. Geophys. Res.*, *100*, 12–431.

Nicholson, T., M. Bostock, and J. Cassidy (2005), New constraints on subduction zone structure in northern Cascadia, *Geophys. J. Int.*, *161*(3), 849–859.

Nicollin, F., D. Gibert, P. Bossart, C. Nussbaum, and C. Guervilly (2008), Seismic tomography of the excavation damaged zone of the Gallery 04 in the Mont Terri Rock Laboratory, *Geophys. J. Int.*, *172*(1), 226–239.

Owens, T., G. Zandt, and S. Taylor (1984), Seismic evidence for an ancient rift beneath the Cumberland Plateau, Tennessee: A detailed analysis of broadband teleseismic P waveforms, *J. Geophy. Res.*, *89*, 7783–7795.

Özalaybey, S., M. Savage, A. Sheehan, J. Louie, and J. Brune (1997), Shear-wave velocity structure in the northern Basin and Range province from the combined analysis of receiver functions and surface waves, *Bull. Seismol. Soc. Am.*, *87*(1), 183–199.

Phinney, R. (1964), Structure of the Earth's crust from spectral behavior of long-period body waves, *J. Geophys. Res.*, *69*, 2997–3017.

Piana Agostinetti, N., and C. Chiarabba (2008), Seismic structure beneath Mt Vesuvius from receiver function analysis and local earthquakes tomography: evidences for location and geometry of the magma chamber, *Geophys. J. Int.*, *175*(3), 1298–1308.

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

Piana Agostinetti, N., F. Lucente, G. Selvaggi, and M. Di Bona (2002), Crustal structure and Moho geometry beneath the Northern Apennines (Italy), *Geophys. Res. Lett.*, *29*(20), 1999, doi:10.1029/2002GL015109.

Rawlinson, N., and B. Kennett (2008), Teleseismic tomography of the upper mantle beneath the southern Lachlan Orogen, Australia, *Phys. Earth Planet. Inter.*, *167*(1–2), 84–97.

Rawlinson, N., and M. Urvoy (2006), Simultaneous inversion of active and passive source data sets for 3-D seismic structure with application to Tasmania, *Geophys. Res. Lett.*, *33*, L24313, doi:10.1029/2006GL028105.

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

Reading, A., B. Kennett, and M. Dentith (2003), Seismic structure of the Yilgarn Craton, Western Australia, *Aust. J. Earth Sci.*, *50*(3), 427–438.

Reading, A., T. Bodin, M. Sambridge, S. Howe, and M. Roach (2010), Down the borehole but outside the box: Innovative approaches to wireline log data interpretation, paper presented at 21st International Geophysics Conference and Exhibition, Aust. Soc. of Explor. Geophys., Sydney, Australia.

Rosenthal, J. (2000), Parallel computing and Monte Carlo algorithms, *Far East J. Theor. Stat.*, *4*(2), 207–236.

Saito, M. (1988), DISPER80: A subroutine package for the calculation of seismic normal-mode solutions, in *Seismological Algorithms: Computational Methods and Computer Programs*, pp. 293–319, Academic, London.

Salah, M., S. Chang, and J. Fonseca (2011), Crustal structure beneath the lower tagus valley, southwestern iberia using joint analysis of teleseismic receiver functions and surface-wave dispersion, *Geophys. J. Int.*, *184*, 919–933.

Sambridge, M. (1999a), Geophysical inversion with a neighbourhood algorithm I. Searching a parameter space, *Geophys. J. Int.*, *138*(2), 479–494.

Sambridge, M. (1999b), Geophysical inversion with a neighbourhood algorithm II. Appraising the ensemble, *Geophys. J. Int.*, *138*(3), 727–746.

Scales, J., and R. Snieder (1997), To Bayes or not to Bayes, *Geophysics*, *62*(4), 1045–1046.

Scales, J., and R. Snieder (1998), What is noise, *Geophysics*, *63*(4), 1122–1124.

Shapiro, N., and M. Campillo (2004), Emergence of broadband Rayleigh waves from correlations of the ambient seismic noise, *Geophys. Res. Lett.*, *31*, L07614, doi:10.1029/2004GL019491.

Shibutani, T., M. Sambridge, and B. Kennett (1996), Genetic algorithm inversion for receiver functions with application to the crust and uppermost mantle structure beneath eastern Australia, *Geophys. Res. Lett.*, *23*, 1829–1832.

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

Smith, A. (1991), Bayesian computational methods, *Philos. Trans. R. Soc. A*, *337*(1647), 369–386.

Stehly, L., B. Fry, M. Campillo, N. Shapiro, J. Guilbert, L. Boschi, and D. Giardini (2009), Tomography of the Alpine region from observations of seismic ambient noise, *Geophys. J. Int.*, *178*(1), 338–350.

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*(3), 150–170.

Thomson, W. (1950), Transmission of elastic waves through a stratified solid medium, *J. Appl. Phys.*, *21*, 89–93.

Tkalčić, H., M. Pasyanos, A. Rodgers, R. Gok, W. Walter, and A. Al-Amri (2006), A multistep approach for joint modeling of surface wave dispersion and teleseismic receiver functions: Implications for lithospheric structure of the Arabian Peninsula, *J. Geophys. Res.*, *111*, B11311, doi:10.1029/2005JB004130.

Tkalčić, H., Y. Chen, R. Liu, Z. Huang, L. Sun, and W. Chan (2011), Multi-step modelling of teleseismic receiver functions combined with constraints from seismic tomography: Crustal structure beneath southeast China, *Geophys. J. Int.*, *187*, 303–326.

Tokam, A., C. Tabod, A. Nyblade, J. Julià, D. Wiens, and M. Pasyanos (2010), Structure of the crust beneath Cameroon, West Africa, from the joint inversion of Rayleigh wave group velocities and receiver functions, *Geophys. J. Int.*, *183*(2), 1061–1076.

Vinnik, L., C. Reigber, I. Aleshin, G. Kosarev, M. Kaban, S. Oreshin, and S. Roecker (2004), Receiver function tomography of the central Tien Shan, *Earth Planet. Sci. Lett.*, *225*(1–2), 131–146.

Vinnik, L., I. Aleshin, M. Kaban, S. Kiselev, G. Kosarev, S. Oreshin, and C. Reigber (2006), Crust and mantle of the Tien Shan from data of the receiver function tomography, *Izv. Phys. Solid Earth*, *42*(8), 639–651.

Yoo, H., R. Herrmann, K. Cho, and K. Lee (2007), Imaging the three-dimensional crust of the Korean Peninsula by joint inversion of surface-wave dispersion and teleseismic receiver functions, *Bull. Seismol. Soc. Am.*, *97*(3), 1002–1011.

Zhao, L., M. Sen, P. Stoffa, and C. Frohlich (1996), Application of very fast simulated annealing to the determination of the crustal structure beneath Tibet, *Geophys. J. Int.*, *125*(2), 355–370.

P. Arroucau, Environmental, Earth and Geospatial Sciences, North Carolina Central University, Durham, NC 27707, USA. (parroucau@NCCU.EDU)

T. Bodin, N. Rawlinson, M. Sambridge, and H. Tkalčić, Research School of Earth Sciences, Australian National University, Bldg. 61, Canberra, ACT 0200, Australia. (thomas.bodin@anu.edu.au; nicholas.rawlinson@anu.edu.au; Malcolm.Sambridge@anu.edu.au)

K. Gallagher, Géosciences Rennes, Université de Rennes 1, Rennes F-35042, France. (kerry.gallagher@univ-rennes1.fr)