Geophysical Journal International

Geophys. J. Int. (2016) **206,** 328–344
Advance Access publication 2016 April 15
GJI Seismology

doi: 10.1093/gji/ggw149

Highly efficient Bayesian joint inversion for receiver-based data and its application to lithospheric structure beneath the southern Korean Peninsula

Seongryong Kim,[1] Jan Dettmer,[1,2] Junkee Rhie[3] and Hrvoje Tkalčić[1]

[1]*Research School of Earth Sciences, The Australian National University, Canberra, ACT* 2601*, Australia. E-mail: seongryong.kim@anu.edu.au*
[2]*School of Earth and Ocean Sciences, University of Victoria, Victoria BC* V8W3P6*, Canada*
[3]*School of Earth and Environmental Sciences, Seoul National University, Seoul,* 151-742*, Republic of Korea*

Accepted 2016 April 12. Received 2016 April 12; in original form 2015 November 17

SUMMARY

With the deployment of extensive seismic arrays, systematic and efficient parameter and uncertainty estimation is of increasing importance and can provide reliable, regional models for crustal and upper-mantle structure. We present an efficient Bayesian method for the joint inversion of surface-wave dispersion and receiver-function data that combines trans-dimensional (trans-D) model selection in an optimization phase with subsequent rigorous parameter uncertainty estimation. Parameter and uncertainty estimation depend strongly on the chosen parametrization such that meaningful regional comparison requires quantitative model selection that can be carried out efficiently at several sites. While significant progress has been made for model selection (e.g. trans-D inference) at individual sites, the lack of efficiency can prohibit application to large data volumes or cause questionable results due to lack of convergence. Studies that address large numbers of data sets have mostly ignored model selection in favour of more efficient/simple estimation techniques (i.e. focusing on uncertainty estimation but employing *ad-hoc* model choices). Our approach consists of a two-phase inversion that combines trans-D optimization to select the most probable parametrization with subsequent Bayesian sampling for uncertainty estimation given that parametrization. The trans-D optimization is implemented here by replacing the likelihood function with the Bayesian information criterion (BIC). The BIC provides constraints on model complexity that facilitate the search for an optimal parametrization. Parallel tempering (PT) is applied as an optimization algorithm. After optimization, the optimal model choice is identified by the minimum BIC value from all PT chains. Uncertainty estimation is then carried out in fixed dimension. Data errors are estimated as part of the inference problem by a combination of empirical and hierarchical estimation. Data covariance matrices are estimated from data residuals (the difference between prediction and observation) and periodically updated. In addition, a scaling factor for the covariance matrix magnitude is estimated as part of the inversion. The inversion is applied to both simulated and observed data that consist of phase- and group-velocity dispersion curves (Rayleigh wave), and receiver functions. The simulation results show that model complexity and important features are well estimated by the fixed dimensional posterior probability density. Observed data for stations in different tectonic regions of the southern Korean Peninsula are considered. The results are consistent with published results, but important features are better constrained than in previous regularized inversions and are more consistent across the stations. For example, resolution of crustal and Moho interfaces, and absolute values and gradients of velocities in lower crust and upper mantle are better constrained.

Key words: Inverse theory; Probability distributions; Surface waves and free oscillations; Computational seismology; Statistical seismology; Crustal structure.

1 INTRODUCTION

Surface wave dispersion (SWD) and teleseismic receiver function (RF) data have been widely used to constrain variability of *S*-wave velocity (V_S) with depth. Observations of SWD over a finite range of periods (dispersion curves) provide constraints on absolute V_S with smooth and continuous depth sensitivity (Takeuchi *et al.* 1964; Aki & Richards 2002). RF data are sensitive to velocity discontinuities which produce converted and reverberated phases due to a wave incident (near vertical) from below (Langston 1979; Ammon 1991). Joint analysis of SWD and RF allows combining their distinct sensitivities to estimate improved 1-D V_S structure (e.g. Özalaybey *et al.* 1997; Du & Foulger 1999; Tkalčić *et al.* 2006; Yoo *et al.* 2007).

The recent expansion of increasingly dense regional networks (e.g. USArray, F-net, China Array, Virtual European Broadband Seismograph Network) has provided opportunities to infer high-resolution structures. In particular, quasi-3-D structure can be estimated from a series of joint inversions of 1-D RF and SWD data (e.g. Shen *et al.* 2013a), which is generally possible when the lateral variability is small beneath each station. However, less attention has been given to the more challenging problem of rigorous uncertainty estimation, in particular due to the high computational cost associated with meaningful uncertainties. This work develops a joint inversion technique that provides rigorous model selection via trans-dimensional (trans-D) optimization and uncertainty estimation via Bayesian sampling that is sufficiently efficient for regional studies with dense regional networks.

The inversion of SWD and RF data for elastic properties is inherently ill-posed and ill-conditioned (Ammon *et al.* 1990; Lomax & Snieder 1995, 2012; Snieder 1998). The non-uniqueness and stability of an inversion can be considered in terms of parameter uncertainties which quantify the range of parameter values that sufficiently fit the data given a likelihood function. However, in linearized approaches, widely applied for efficiency, the inverse problem is approximated (by linearization), often resulting in difficult-to-interpret uncertainty estimates (e.g. Özalaybey *et al.* 1997; Du & Foulger 1999; Julià *et al.* 2000). Such inversions can be strongly affected by subjective information (e.g. the initial model) and regularization (e.g. damping, smoothing) (Jackson & Matsu'ura 1985; Sambridge & Mosegaard 2002) which is often arbitrarily determined by *ad-hoc* methods and the impact of regularization parameters on results is not quantified (Scales & Snieder 2000; Mosegaard & Sambridge 2002; Sambridge & Mosegaard 2002). For instance, inversion uncertainties from Bootstrap resampling tests (e.g. Efron & Tibshirani 1991) tend to be underestimated due to regularization (Sambridge *et al.* 2006a).

Nonlinear approaches have been applied to SWD and RF to address these issues (e.g. Lomax & Snieder 1995; Chang *et al.* 2004; Lawrence & Wiens 2004). For example, optimization methods (e.g. simulated annealing and genetic algorithms) obtain a global, optimal parameter vector which does not depend on initial parameter value choice. Note that global optimization also requires some form of model selection or regularization to avoid over/under fitting of data. Regularization is commonly applied when the problem is overparametrized (by many thin layers of fixed thickness) but causes an approximation of the inverse problem that can make results difficult to interpret. Model selection, the process of identifying a parametrization that is consistent with data information, is often ignored when simple parametrizations are applied (a few layers of unknown thickness based on *ad-hoc* decisions) to avoid the associated computational cost and conceptual challenges of quantitative model selection, or to be simply based on prior information. However, this can lead to underparametrization issues (erroneous results, biases) that are difficult to detect.

Parameter uncertainties due to non-uniqueness cannot be fully estimated by optimization and depend strongly on model choice and ensemble inference based on arbitrary criteria is insufficient (Douma *et al.* 1996; Sambridge 1999; Sambridge & Mosegaard 2002; Lomax & Snieder 2012). Bayesian inference provides rigorous uncertainty estimates by quantifying the posterior probability density (PPD) of model parameters. The PPD is estimated by numerical, multidimensional integration via nonlinear sampling methods (e.g. Sambridge & Mosegaard 2002; Tarantola 2005) that produce an ensemble of parameter-vector values that is statically representative of the data errors. The most common and efficient sampling methods for high-dimensional problems are based on Markov chain Monte Carlo (MCMC) integration (Brooks *et al.* 2003). The PPD is constrained by data information (the likelihood function) and prior probabilities (independent information) (Jackson & Matsu'ura 1985; Duijndam 1988; Scales & Snieder 1997; MacKay 2003; Tarantola 2005) and parameter uncertainties are obtained by marginalization to obtain quantities of interest (means, marginal profiles, variances and covariances). Bayesian methods have been increasingly adopted for inversions using SWD and RF data (e.g. Piana Agostinetti & Malinverno 2010; Bodin *et al.* 2012; Shen *et al.* 2013b; Dettmer *et al.* 2015).

The choice of specific model parametrization (e.g. the number of layers) strongly affects parameter uncertainties. The resolution of earth structure by observed data is not straightforward to quantify because data are generally band-limited and sampling is incomplete, resulting in sensitivity that varies as a function of space. For example, for layered parametrizations, overparametrization with many layers can result in spurious structure and overfitting in some parts of the model while other regions may be appropriately parametrized. Therefore, uncertainty estimates can also vary as a function of space and may be overestimated in overparametrized parts of the model and underestimated in underparametrized regions. Quantitative model selection can be applied to address these issues in a Bayesian framework where the normalizing constant in Bayes' theorem (evidence) quantifies the data support for a parametrization. In particular, model selection is applied to identify parametrizations that are consistent with the information in the observed data (MacKay 2003). Since full evidence computation can be prohibitively expensive, asymptotic point estimates, such as the Bayesian information criterion (BIC; Schwarz 1978), constitute a practical approach based on a Gaussian approximation of the posterior around the main posterior mode. Limitations arise due to the fact that the Gaussian approximation around the maximum *a posteriori* (MAP) model may not be representative of the whole posterior. However, in many cases the BIC has been documented to provide meaningful parsimonious estimates.

More complete uncertainty estimation can be achieved by integrating model selection and sampling via trans-D models (Malinverno 2002; Bodin *et al.* 2012; Dettmer *et al.* 2012): Model specification is relaxed from a single model to a group of models for which an algorithm can be formulated that jumps between the various parameter spaces while maintaining Markov chain properties. The uncertainty estimates then include the effects due to the ambiguity of competing models. However, this requires fine-tuning to ensure proper dimensional changes and to minimize additional cost in computations. For example, the popular birth–death algorithm (Malinverno 2002) is strongly affected by the way new states of the parameter vector are proposed, which is problem dependent

and efficient ways are often not known or possible (Dosso *et al.* 2014).

The uncertainties from Bayesian inversions are fundamentally linked to the data errors via the likelihood function, which is a measure of the fit of the prediction to the observation given an assumed statistical distribution of data errors (Tarantola 2005). Data errors are typically a combination of theory (e.g. due to model limitations, assumptions in data processing) and measurement (due to the observation process) errors, which are difficult to separate. While measurement errors may be estimated from ensembles of data or pre-event noise, theory error is model dependent and can be significant. Since it is impossible to obtain data errors independently, they are often approximated by residual errors (Dosso & Wilmut 2006). In that case, the distribution parameters (e.g. variance and covariance) must be estimated either to form the residuals directly (empirical Bayesian estimation) or to be treated as unknown (hierarchical estimation) (e.g. Gouveia & Scales 1998; Sambridge 1999; Malinverno & Briggs 2004; Bodin *et al.* 2012; Dettmer & Dosso 2012).

In this paper, we present a procedure for Bayesian joint inversion of SWD and RF data that applies efficient model selection and uncertainty estimation while avoiding some of the high computational cost associated with trans-D models. The BIC is applied as the quantitative criterion to select the most probable parametrization and is applied within a trans-D optimization algorithm (e.g. Brooks *et al.* 2003) to carry out the selection in a computationally efficient and automated manner. In particular, the BIC replaces the likelihood function in the trans-D optimization phase of the inversion. In classic model selection with the BIC (e.g. Molnar *et al.* 2010), multiple fixed-dimensional (fixed-D) optimizations are performed for the various models under investigation. Once maximum likelihood (ML) parameters are determined for each model, the BIC is computed and the smallest BIC value identifies the optimal model. This approach is straightforward to implement and to monitor the selection process, and several complexities of trans-D schemes are not required (e.g. birth-death). However, in the case where the models to examine are little-known *a priori*, this can be rather tedious and requires additional practitioner interaction, such as searching unnecessary models and trial-and-error to find proper ranges of models. In our approach, the BIC constrains the variability of the complexity parameter (e.g. the number of layers) more than would be the case in trans-D sampling. Therefore, we avoid extensive exploration of the tails of the distribution of that parameter; an aspect of trans-D sampling that is known to be inefficient. Finally, the optimization does not require detailed balance and resampling can be periodically applied to increase efficiency. Once the optimal parametrization is obtained, it is held fixed and efficient sampling is carried out for uncertainty estimation. Although a limitation of our approach is that uncertainty estimates do not include the effect of limited knowledge about the parameterization, we show that meaningful uncertainty estimates are obtained by our method.

Both optimization and sampling are implemented with parallel tempering (PT, Geyer 1991; Dettmer & Dosso 2012; Sambridge 2013) to provide good scaling of algorithm performance with the number of processors on computer clusters. The PT algorithm employs many Markov chains that concurrently sample a sequence of distributions of which at least one distribution is identical to the PPD. The other distributions are increasingly tempered (relaxed by raising the likelihood to a power of <1) and aid sampling efficiency. The tempering facilitates free movement of the Markov chain causing tempered chains to have higher acceptance rates of proposed model vectors resulting in wide exploration of the parameter space. Information exchange between chains is implemented by a Metropolis–Hastings (MH) criterion for Markov chain pairs which can improve efficiency dramatically (Dosso *et al.* 2012). As opposed to most previous work applying PT in geophysical problems, we apply PT as an efficient and inherently parallel optimization and an alternative to sequential techniques (e.g. simulated annealing). The combination of the PT, trans-D optimization steps with the BIC, and parallel computer hardware result in an efficient optimization.

Data errors are addressed with a combination of empirical and hierarchical estimation based on residual errors. Note that errors in RF data are often strongly correlated and non-stationary between data points. Similarly, SWD data often exhibit non-stationary errors as a function of period. Hence, simple parametric covariance models (with few parameters) may not be sufficient to describe the error statistics. To quantify potentially complex structure in the data covariance matrix without requiring large numbers of parameters, we empirically estimate covariance matrices from the autocovariance functions of data residuals with non-stationary weighting (Dettmer *et al.* 2007). In addition, these non-Toeplitz matrices are scaled by unknown hierarchical parameters during sampling. This approach accounts for complex error statistics without introducing complicated error parameters that may be difficult to justify. In the optimization phase, only the magnitude of noise (standard deviation) is accounted throughout the ML technique (Dosso & Wilmut 2006), since well-converged data residuals may not be available in the initial stage. Regarding possible trade-offs between the model selection and the error estimation, however, our results show that this simplified approach is generally effective.

The method is studied for both simulated and observed SWD and RF data for three stations on the southern Korean Peninsula, where each station represents a specific tectonic region. SWD data are obtained from ambient noise analysis (e.g. Bensen *et al.* 2007) using continuous data of the entire array in the southern Korean Peninsula. In all cases, joint inversions are performed for RF together with phase velocity (PV) and group velocity (GV) we invert them jointly to provide better constraints on velocity structure due to the different depth sensitivities (Rodi *et al.* 1975), and due to the different measurement processes that result in independent noise on the two data types (e.g. Martinez *et al.* 2000; Shapiro & Ritzwoller 2002). The inversion results provide better constraints about the tectonic evolution of the southern Korean Peninsula. It is clearly evident that stations in the southeastern part of the Korean Peninsula show relatively high V_S in the lower crust and low V_S in upper mantle, which can be interpreted as evidence for magma underplating and a hot upper mantle along the southeastern edge of the Peninsula.

2 THEORY AND METHOD

This section provides a brief overview of Bayesian methodology and then describes in detail the inversion method applied in this study.

2.1 Bayesian formulation

Let $\mathbf{d}$ be a vector of random variables with N observed data which contains information about the earth and $\mathcal{M}$ denote a group of models (specifying parametrization, theory for predictions and error statistics). Let $\mathbf{m}$ be a vector of M random variables containing all model parameters. Bayes' rule defines the PPD $p(\mathbf{m}|\mathbf{d}, \mathcal{M})$ as

$$p(\mathbf{m}|\mathbf{d}, \mathcal{M}) = \frac{p(\mathbf{d}|\mathbf{m}, \mathcal{M})p(\mathbf{m}|\mathcal{M})}{p(\mathbf{d}|\mathcal{M})}, \quad (1)$$

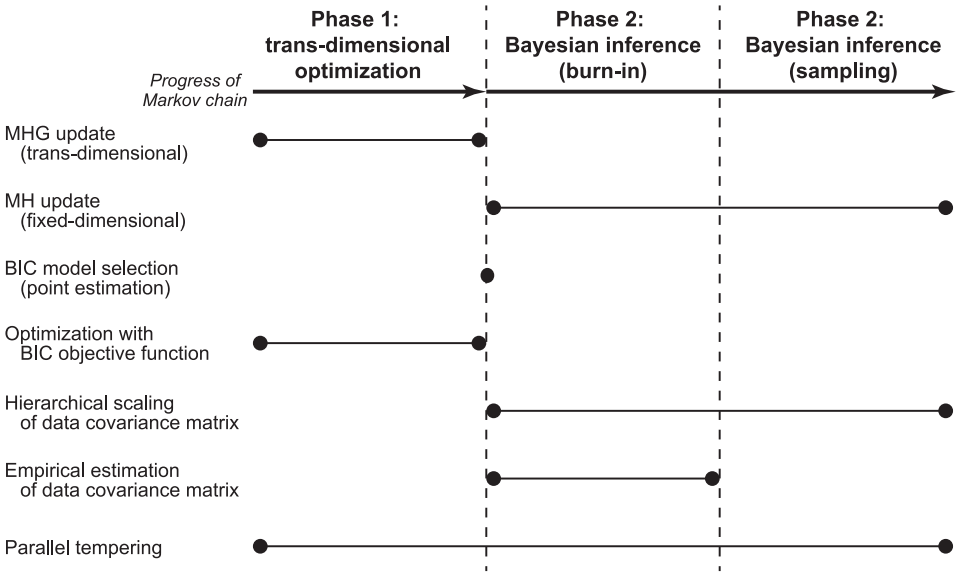

Figure 1. Schematic diagram of the inversion process. The names of applied algorithms are presented with their working periods in the Markov chains. MHG and MH indicate Metropolis–Hastings–Green (see eq. 5) and Metropolis–Hastings algorithms, respectively. Bayesian Information Criterion (BIC) is presented in eq. (7) for the model selection and in eq. (9) for the objective function.

which quantifies the state of information about the parameters given data and prior $p(\mathbf{m}|\mathcal{M})$. Data information is represented by $p(\mathbf{d}|\mathbf{m}, \mathcal{M})$, which is the distribution of data errors but for observed (fixed) data is interpreted as the likelihood function $L(\mathbf{x})$, where $\mathbf{x} = (\mathbf{m}, \mathcal{M})$. The normalizing constant in eq. (1) is referred to as Bayesian evidence and plays a key role in model selection.

For uncertainty estimation, the likelihood function must be specified based on an assumption about the residual-error statistics. We assume Gaussian-distributed residuals ($\mathbf{r}_i = \mathbf{d}_i - \mathbf{d}_i(\mathbf{x})$, the difference between prediction and observation) which leads to

$$L(\mathbf{x}) = \prod_{i=1}^{S} \frac{1}{\sqrt{(2\pi)^{N_i}\,|\mathbf{C}_{\mathbf{d}i}|}} exp\left\{-\frac{1}{2}\mathbf{r}_i{}^T \mathbf{C}_{\mathbf{d}i}{}^{-1}\mathbf{r}_i\right\}, \tag{2}$$

where i indexes S data sets $\mathbf{d}_i$ including SWD and RF. N_i are the number of data for the ith set and $\mathbf{C}_{\mathbf{d}i}$ the data covariance matrices. Note that the likelihood function concept is general and can be applied to any combination of data sets (PV, GV and RF). Importantly, the appropriate weight for various data sets is given by errors on the data and does not require subjective specification. Prior distributions are chosen to be bounded and uniform over a wide range of parameter values. For the parameter that indexes model complexity, a bounded Poisson distribution is applied (see Supporting Information Fig. S1 and Green 1995)

$$p(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \tag{3}$$

where λ is a scale parameter.

The evidence in eq. (1) represents the probability of the data resulting from model $\mathcal{M}$ and can be interpreted as a likelihood of the model parametrization, which is the foundation of Bayesian model selection (Sambridge *et al.* 2006b). Because the evidence normalizes the PPD, it is an integral over the parameter space and given by

$$p(\mathbf{d}|\mathcal{M}) = \int p(\mathbf{d}|\mathbf{m}, \mathcal{M})p(\mathbf{m}|\mathcal{M}). \tag{4}$$

Direct estimation of evidence is numerically challenging and costly for high-dimensional, nonlinear problems. However, for a given model, the un-normalized PPD is sufficient for parameter estimation. While important, evidence computations it is often ignored due to the challenges it poses.

Metropolis–Hastings–Green (MHG) sampling (Metropolis *et al.* 1953; Hastings 1970; Green 1995) was developed to facilitate posterior sampling for trans-D models. For the models considered here, a simple case of MHG sampling is sufficient to allow jumps between parameter vectors that support different numbers of layers: We apply birth-death moves (Geyer & Møller 1994) to either add or remove a layer in the stratified earth model. While posterior sampling requires a likelihood function to be applied, for optimization, the likelihood function is replaced by a model selection criterion (Brooks *et al.* 2003), in this case the BIC (the function b in eq. 9). The birth-death update is carried out by proposing one of three essential updates, where a layer is either added, removed, or perturbed. These updates change the current state $\mathbf{x}$ to state $\mathbf{x}'$ based on a proposal distribution q and the new state is accepted with probability

$$\alpha_{\mathrm{MHG}} = \min\left[1, \frac{p(\mathbf{x}')}{p(\mathbf{x})}\left\{\frac{b(\mathbf{x}')}{b(\mathbf{x})}\right\}^{\beta}\frac{q(\mathbf{x}|\mathbf{x}')}{q(\mathbf{x}'|\mathbf{x})}\,|\mathbf{J}|\right], \tag{5}$$

where β is an annealing parameter for the PT algorithm (described later) and $|\mathbf{J}|$ is the determinant of the Jacobian matrix of the transformation from $\mathbf{x}$ to $\mathbf{x}'$. We implement dimension changes such that $|\mathbf{J}| = 1$ (e.g. Malinverno 2002).

A birth is carried out by sampling a new interface position from a uniform proposal which splits an existing layer into two new ones. Then, the V_S of a randomly selected new layer above or below the interface is perturbed by a Gaussian proposal probability which is centred on the previous value of that layer. A death is based on selecting a random interface and deleting it. The merged layer has a V_S which is randomly selected from the values of the previous two layers.

Parallel tempering (Geyer 1991; Dettmer & Dosso 2012) is applied to improve efficiency not only for the optimization but also for the subsequent sampling processes (Fig. 1). For PT, several Markov chains concurrently sample a sequence of increasingly tempered/relaxed distributions with at least one distribution identical to

the PPD. The tempered distributions are increasingly relaxed by raising the likelihood to a power of <1. The acceptance probability in eq. (5) includes a tempering parameter β with value between 0 and 1 which de-emphasizes the likelihood function for small values. As a result, low-β chains explore the parameter space more widely, while high-β chains are more probable to explore local high-likelihood modes. Our PT scheme considers only $\beta \leq 1$, which is the typical approach in sampling. In the optimization phase, including chains of $\beta > 1$ may improve convergence to the best model. However, in the applications considered here, it is important to improve algorithm search efficiency in the trans-D space which is predominantly improved by chains $\beta < 1$ (Dettmer & Dosso 2012). The PT algorithm is particularly suited for parallel computers where many processors update the Markov chains concurrently. Information exchange between chains is implemented by exchange moves which apply the MH criterion for chain pairs (Geyer 1991)

$$\alpha_{\mathrm{PT}} = \min\left[1, \left\{\frac{b(\mathbf{x}_l)}{b(\mathbf{x}_n)}\right\}^{\beta_n - \beta_l}\right], \tag{6}$$

where $\mathbf{x}_l$ and $\mathbf{x}_n$ are the states of two randomly selected chains l and m, respectively. Note that the prior and proposal ratios in eq. (5) cancel since exchange moves are proposed for randomly selected chain pairs and only the likelihood is tempered. Exchange moves are carried out at the end of every iteration, since the process requires minimum computational effort and communication between chains is of low latency.

2.2 Parametrization and data prediction

We assume horizontally stratified isotropic, homogeneous layers over a semi-infinite half-space, where $\mathbf{x} = (k, \mathbf{m}_k)$ indicates a set of parameters for $k-1$ interfaces. The parameter vector $\mathbf{m}_k$ includes the elastic properties of $k-1$ layers and the half-space, the interface positions $\mathbf{z}$ and the noise parameters ($\mathbf{m} = (\mathbf{z}, \mathbf{v}, \kappa, \mathbf{w})$). The interface positions (depths) are between $z = 0$ and $z = z_{\max}$. The V_S of k layers and the half-space is given by $\mathbf{v}$. While P-wave velocities (V_P) and density can affect SWD and RF data, sensitivity is much weaker than to V_S (Takeuchi *et al.* 1964; Tanimoto 1991; Julià *et al.* 2000) and we sample the bulk V_P/V_S ratio κ (independent of layering). This approach avoids biases due to the choice of a significantly wrong V_P/V_S value when little prior knowledge is available but also avoids over parametrization. The density of each layer is given by an empirical relationship (Brocher 2005). The vector $\mathbf{w}$ contains the scaling parameters of the data covariance matrices for S data sets. To predict RF data for a set of parameters, we apply the reflection matrix method (Randall 1989) and water-level deconvolution (Langston 1979) for RF data. To predict SWD data, a normal mode solutions is applied (Saito 1988).

2.3 Trans-D optimization and model selection (Phase 1)

Prior to Bayesian sampling, an optimal parametrization is determined via trans-D optimization (Brooks *et al.* 2003), see Fig. 1. In this phase, the evidence in eq. (4) is asymptotically approximated by the BIC which makes the assumption of a Gaussian posterior around the ML parameter vector $\hat{\mathbf{x}}$. The BIC is defined as

$$\mathrm{BIC} = -2\log\{L(\hat{\mathbf{x}})\} + M\log N. \tag{7}$$

The likelihood term in eq. (7) quantifies the fit to the data while the second term penalizes complexity. The most common application of the BIC is to carry out optimization for multiple parametrizations and then identify the most probable one based on the minimum BIC value. Here, the BIC is applied as the objective function in an optimization that allows jumps between parametrizations based on the reversible jump algorithm (Green 1995). Since no prior information about the data-error statistics is available, we initially assume uncorrelated errors of unknown standard deviation σ_i (i.e. $\mathbf{C}_{\mathbf{d}i} = \sigma_i^2\mathbf{I}$). In this case, the (un-normalized) likelihood function is given by (Dosso & Wilmut 2006)

$$L(\mathbf{x}) \propto \prod_{i=1}^{S} \exp\left\{-\frac{N_i}{2}\log\left(\mathbf{r}_i^T\mathbf{r}_i\right)\right\}, \tag{8}$$

and the objective function for the optimization is given by the BIC

$$b(\mathbf{x}) \propto \exp(2\log\{L(\mathbf{x})\} - M\log N). \tag{9}$$

Note that $\mathbf{x}$ does not contain the ML parameters but rather the current state of the optimization.

We combine trans-D sampling, optimization with a BIC objective function, and PT to provide a computationally efficient and practical model selection. The trans-D optimization operates on a range of models (varying in the number of layers) and adapts the search space based on the particular data. In addition, the choice of selection criteria (here the BIC) for the objective function provides constraints on model complexity. This improves efficiency greatly by eliminating models with relatively low probability but does require the prior choice of a reasonable selection criterion.

At the end of the optimization phase, the PT algorithm provides a number of parameter vectors from which the one with minimum BIC value is chosen to conclude the trans-D optimization (Fig. 1). Note that the algorithm maintains a memory about the lowest BIC value throughout the history of the algorithm so that this selection identifies the optimal model over the full Phase-1 history.

In MCMC sampling, some iterations (burn-in) are often discarded until chains obtain equilibrium with regard to the posterior. We note that burn-in for non-interacting parallel simulations can be highly inefficient. Here, the interacting aspect of the algorithm provides efficient burn-in during the optimization phase.

2.4 Posterior sampling (Phase 2)

To estimate rigorous parameter uncertainties, the data error parameters in eq. (2) must be quantified based on the residual-error $\mathbf{r}$ since we do not have direct access to data errors. In particular, theory errors often dominate strongly and affect the PPD estimate. While, in principle, hierarchical methods (e.g. Bodin *et al.* 2012) can estimate general structure in the data covariance matrix, such general models can strongly trade off with the ability to resolve earth structure. In addition, a problematic aspect of basing error estimates on residuals is the residual dependence on the particular set of parameters used to compute them. To avoid strong dependence of results on this initial choice, we estimate data covariance matrices $\mathbf{C}_{\mathbf{d}}$ in eq. (2) from residual-errors based on the optimization result but allow hierarchical scaling during sampling so that the magnitude of covariance matrices is treated as unknown. With this approach, we can account for strongly correlated and non-stationary errors without requiring elaborate parametrizations of data errors.

Data covariance matrices are estimated here using an empirical approach (Dettmer *et al.* 2007) before the sampling phase (Fig. 1). First, data residuals are scaled by the standard deviations of the residuals which are smoothed by a moving Gaussian window centred on the corresponding point. The width of the Gaussian window is a control parameter that is chosen to provide numerical

Table 1. True model parameters to generate synthetic data.

Layer	Depth (km)	V_S (km s^{-1})
1	2	2.2
2	9	3.2
3	17	3.0
4	26	3.4
5	35	4.8
6	50	4.6
7	half-space	4.8

stability of the estimate (the method is not sensitive to this choice). Then, Toeplitz autocorrelation matrices are calculated for the scaled residuals. Finally, the covariance matrices are scaled by the non-stationary standard deviations to give the final non-Toeplitz data covariance matrices. This process is performed several times to update the matrices iteratively (Fig. 1). In our procedure, a short period of iterations is carried out following the trans-D optimization where data covariance matrices are iteratively updated. In these examples, the process is performed five times.

The hierarchical scaling is applied during sampling (Phase 2) and based on the empirical matrices $\mathbf{C}_{\mathbf{e}i}$ which are scaled by factors $\mathbf{w}$. The scaling factors have uniform priors centred on zero and the scaling is given by

$$\mathbf{C}_{\mathbf{d}i} = e^{2w_i}\mathbf{C}_{\mathbf{e}i}. \tag{10}$$

Substituting eq. (10) into eq. (2) gives the likelihood function with hierarchical scaling

$$L(\mathbf{m}) \propto \prod_{i=1}^{S} \exp\left(-N_i w_i - \frac{1}{2e^{2w_i}}\mathbf{r}_i{}^T\mathbf{C}_{\mathbf{e}i}{}^{-1}\mathbf{r}_i\right). \tag{11}$$

The PPD sampling is carried out for the optimal parametrization that was determined by trans-D optimization and the empirical covariance matrix estimates. In this case the MHG acceptance (eq. 5) simplifies since the parametrization does not change. Note that a proper likelihood function must be applied here instead of the BIC objective function. Parallel tempering is applied for sampling efficiency. Since prior distributions are considered to be uniform, the acceptance probability of the update reduces to the tempered likelihood ratio.

3 SIMULATION RESULTS

This section applies the algorithm to simulated data which are produced for a set of parameters consisting of six layers over a half-space (Table 1). The maximum depth of the trans-D model during optimization is fixed to 70 km and the V_P/V_S ratio is 1.73 for all layers. A simulated data set consists of a RF waveform and fundamental-mode Rayleigh-wave SWD curves for GV and PV (Figs 2a–c). To generate correlated random noise, realistic data covariance matrices are applied for each data type (Figs 2d–f). We use the relationship $\mathbf{C}_{\mathbf{s}ij} = \sigma_i\sigma_j r^{\sqrt{(i-j)^2}}$, where σ_i and σ_j are assigned the standard deviations of the ith and jth data points, respectively, and $r < 1$ is a scaling controlling the correlation lengths. To simulate realistic errors, we further assume that error magnitudes depend on data magnitudes. Hence, the standard deviation of the ith data point is defined as $\sigma_i = \sigma_b + a|d_i|$, where σ_b is the standard deviation of the background noise and a is a magnitude scaling.

For the SWD data, we use 0.05 km s^{-1}, 0.01 and 0.80 for σ_b, a and r, respectively. The values to generate RF data are 0.03, 0.10 and 0.90. Correlated noise (Figs 2a–c) is obtained by multiplying a vector of Gaussian random numbers by the Cholesky decomposition of the $\mathbf{C}_\mathbf{s}$. The synthetic SWD curves consist of 48 data points over periods of 3–50 s with uniform 1 s sampling (Figs 2a and b). The Gaussian filter width for the simulated RF is 2.5, and the RF has 216 data points with a sampling rate of 6.25 Hz (Fig. 2c).

Uniform priors are applied with bounds of [2.0,5.5] km s^{-1}, [0.0, 70.0] km, and [1.6,2.0], for V_S, layer boundaries, and V_P/V_S ratio, respectively. The prior for scaling factors of RF, PV and GV data covariance matrices is $[-3, 3]$. During the trans-D optimization, the number of layers ranges from 2-30. We use $\lambda = 6.0$ (eq. 3) for the Poisson prior on k (Supporting Information Fig. S1).

Twelve processors are used and each one simulates from a single Markov chain. A total of 200 000 iterations are computed, including 50 000 iterations for trans-D optimization, 50 000 for burn-in and 100 000 for sampling. Here we empirically determine the numbers of iterations by visually monitoring the progress of chains for stationary distributions of data root-mean-square errors and the likelihood function. The overall quality of samples and the suitability of assumptions can be checked *a posteriori* by the residual analysis (Appendix A). Note that we carefully choose the number of iterations based on many trials in all inversions of this work. However, the proper length of chains may vary according to data and prior ranges. Future goals include more rigorous automated checks for the convergence of chains. Each processor is assigned a β value from an exponential sequence from $\sim$0.001 to 1. Samples are collected only for $\beta = 1.0$ such that the final PPD ensemble has 100 000 samples. The trans-D optimization provides rigorous and efficient model selection without practitioner interaction.

The optimum number of layers is selected by taking the lowest BIC value throughout the optimization phase (Fig. 3). The inferred number of layers is 7, which is consistent with the true value. It is noted that the maximum values of the likelihood increase with the number of layers until models with 8 layers, which is greater than the selected dimension. This can be useful to empirically check the convergence of the model selection process. The following inferences are all based on this (fixed) number of interfaces. The inversion requires a total of less than one hour of computer time, which is at least an order of magnitude less cost than our trans-D sampling algorithms require without chain interactions. Inferences from the high-dimensional PPD are presented in several ways (see Appendix B). Fig. 4 shows V_S profile marginals and interface probabilities. These V_S results exhibit clear velocity transitions for each layer interface. Since the V_S uncertainty estimates include the true model throughout, we conclude that V_S is resolved well by the data despite the correlated noise. The 1-D marginal of bulk V_P/V_S shows a single sharp peak centred at the true value (1.73). Interface probability as a function of depth (Fig. 4c) shows six clear peaks close to the true values. The MAP, posterior mean-, and marginal mean-models (defined in Appendix B) are presented in Fig. 4(d) and all are close to each other. In addition, we carry out a comparison with a full trans-D uncertainty estimation approach (see Supporting Information Fig. S2), which is that the trans-D algorithm is applied throughout the entire inversion process. Based on our test, sampling duration should be at least four times longer to obtain a converged PPD. While the trans-D convergence is much more challenging, the PPD does not differ substantially from the results obtained with our method (Fig. 4).

The range of data predictions produced by the PPD and the fit for the MAP model are shown in Fig. 5. Despite the significant correlated noise on the data, the range of PPD data predictions and the MAP prediction are reasonably close to the raw synthetic data without noise. The estimated error statistics and scaling

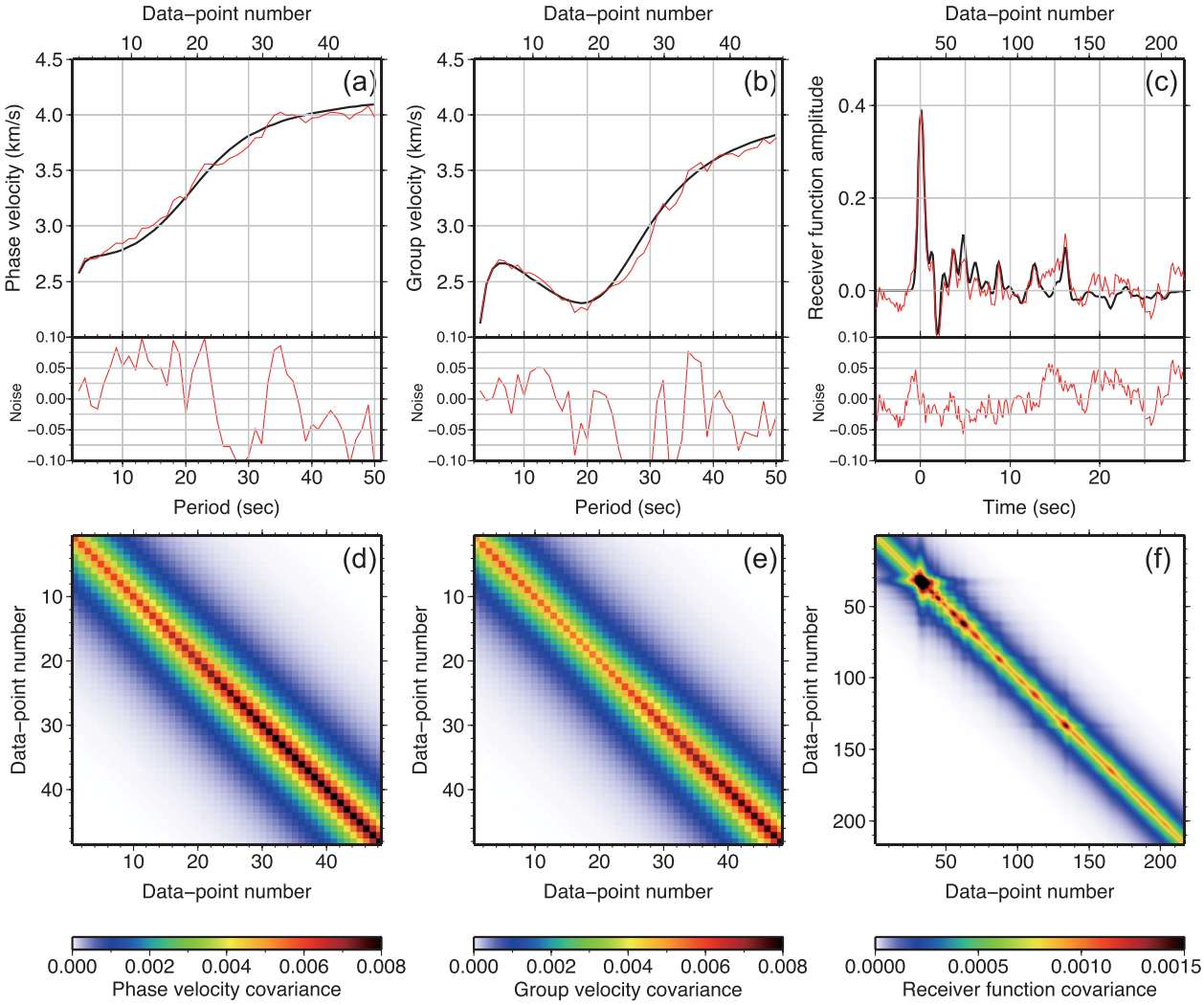

Figure 2. Synthetic data produced using the assumed true model for (a) phase velocity, (b) group velocity and (c) receiver function. Black and red lines show noise-free and noisy data, respectively. Insets at the bottom of the figures show applied Gaussian random numbers to generate the noises in data. Data covariance matrix shown in panels (d)–(f) is used to generate each of the noisy data in panels (a)–(c).

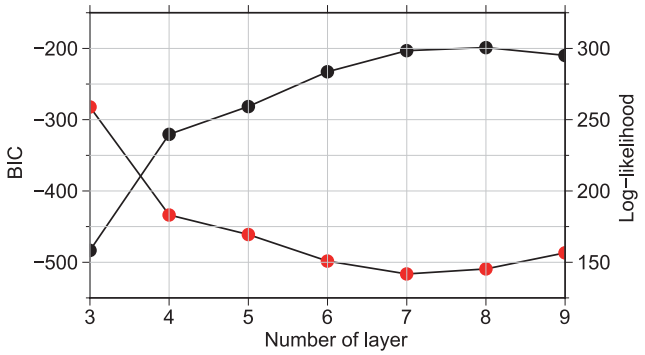

Figure 3. Estimated maximum values of logarithmic likelihood (black) and corresponding values of the BIC function (red) as a function of the number of layers in the trans-D optimization phase.

factors are compared to true values in Fig. 6. The estimated covariance matrices generally match the true ones in terms of magnitude and shape. The marginals of scaling factors (Figs 6g–i) are not centred on zero and are quite broad, indicating that the effects of theory errors are not fully resolved by the empirical data covariance approximation.

Fig. 7 shows the posterior normalized model covariance (correlation) matrix which illustrates dependence between parameters. For instance, V_S values of adjacent layers are often negatively correlated (e.g. V_2 and V_1). However, correlation strength decreases with separation (e.g. V_1 and V_3). Positive correlations are observed between layer interface positions and the appropriate V_S values. Neighbouring interface positions are also often positively correlated. The scaling factors show less correlation with other parameters, while the V_S are negatively correlated with V_P/V_S.

Joint marginal distributions (Fig. 8) provide additional insights into correlations and uncertainties, such as multi-modality and how similar the posterior is to a multivariate Gaussian. For example, negative correlation exists between V_S of neighbouring layers (Fig. 8a), and positive correlation between V_S and z (Fig. 8b), which is consistent with Fig. 7. In addition, weak negative correlation is observed between various interface positions (Fig. 8c) and between V_P/V_S (k) and V_S (Fig. 8d). The correlation between V_S and interface depths appears to be the clearest. Such velocity-depth trade-off is a common inter-relationship in many geophysical problems using layered

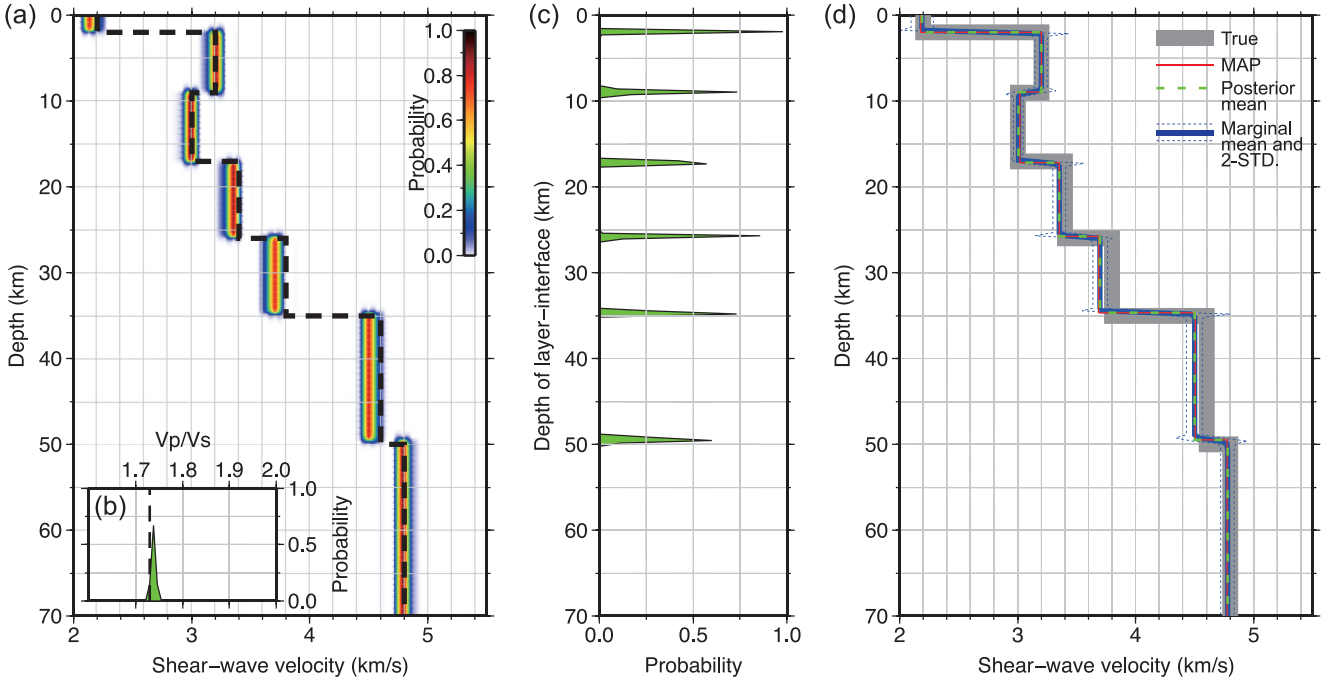

Figure 4. Synthetic inversion results: (a) posterior profile marginal densities (colour scale) and true model (dashed) for V_S, (b) 1-D marginal and the true value (dashed) for V_P/V_S, (c) interface probability and (d) individual models from the PPD. In panel (d), the uncertainties of the marginal mean model are presented with ± 2 standard deviations of the V_S marginal at corresponding depth bins.

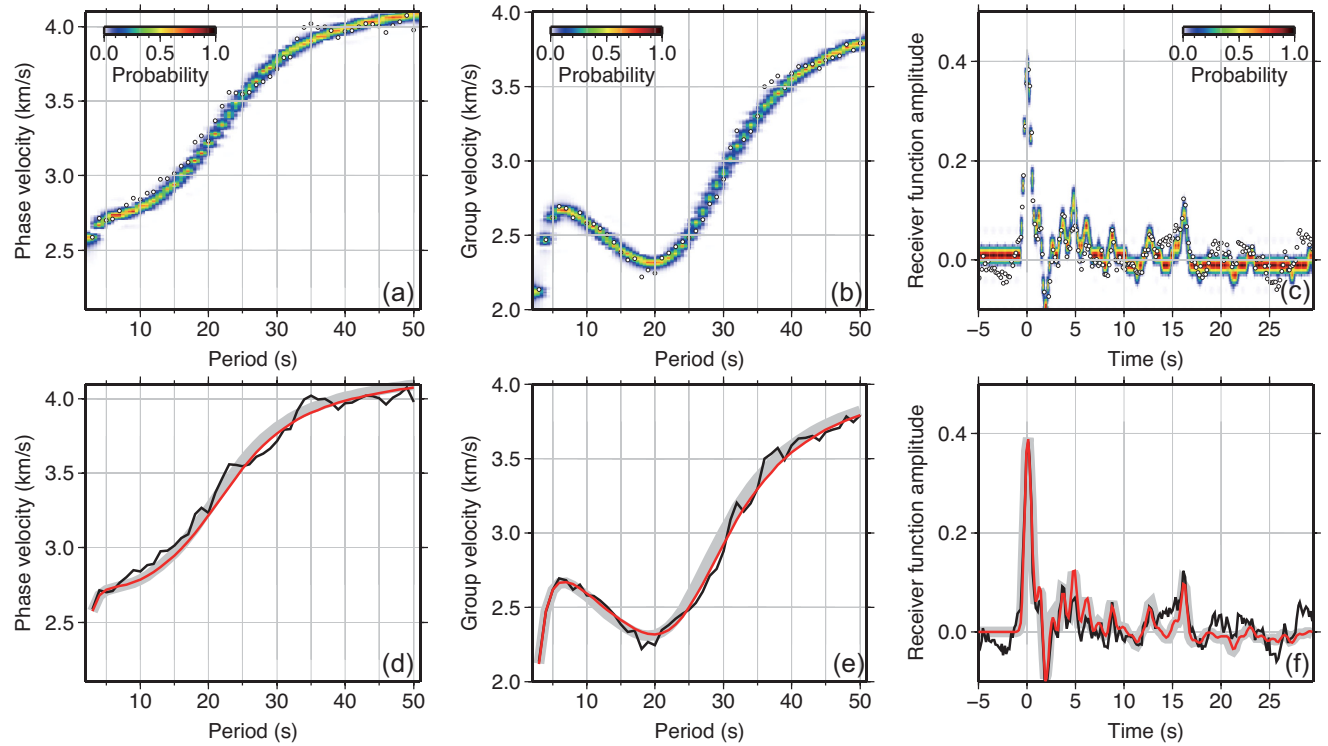

Figure 5. Range of data predictions for the PPD compared to synthetic observed data (circles): (a) PV, (b) GV, and (c) RF. The predictions for the MAP model (red) are compared to the synthetic data (with and without correlated errors, black and thick grey, respectively) for (d) PV, (e) GV, and (f) RF.

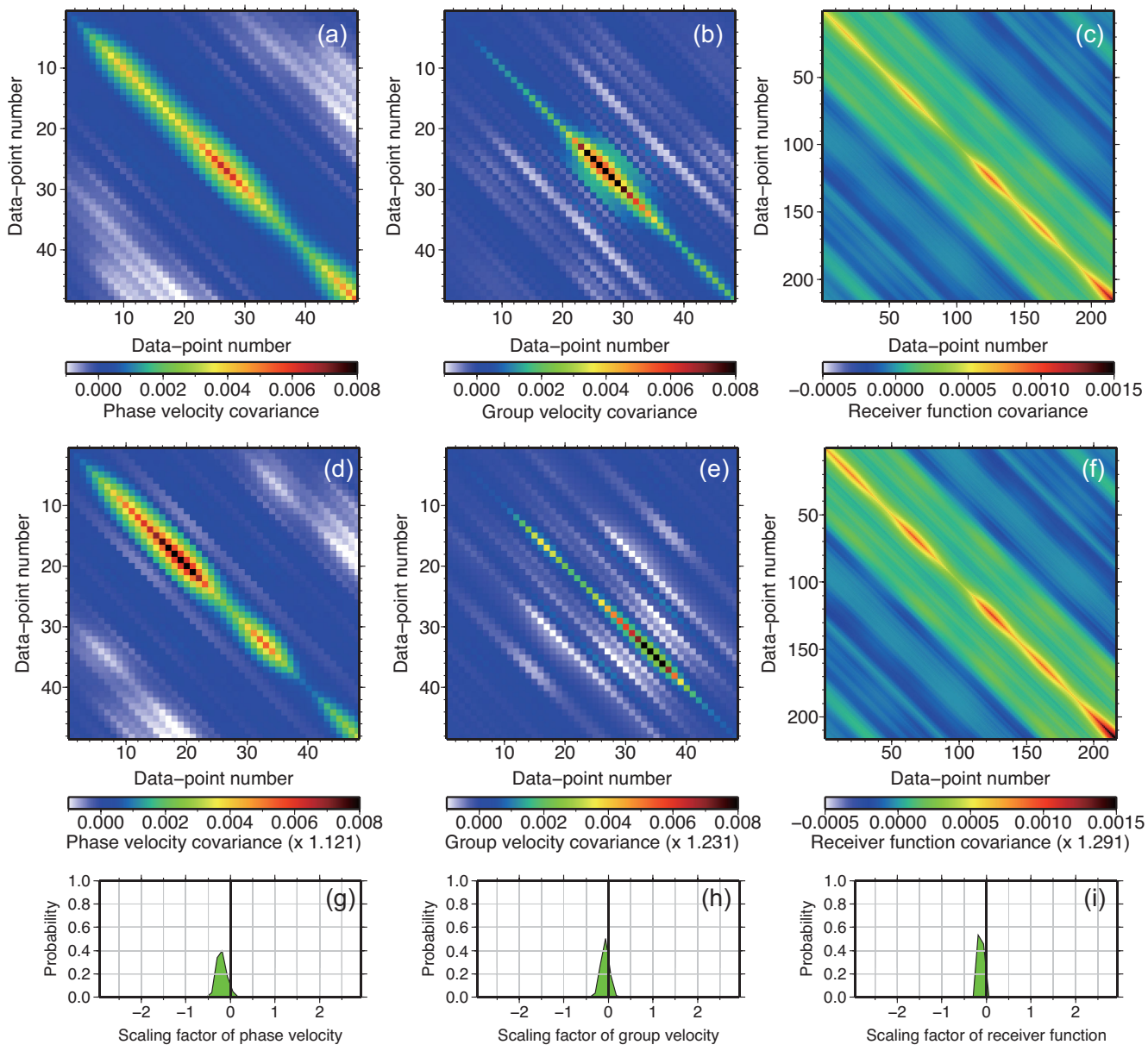

Figure 6. True and estimated empirical and hierarchical errors for PV (left), GV (middle) and RF (right) data: (a–c) True non-stationary data covariance matrices (see Section 2.4), (d–f) inversion estimates, and (g–i) 1-D marginals PPDs of scaling factors (eq. 10). The MAP and true values are shown as red and black lines, respectively. The estimated matrices are multiplied by the MAP values of the scaling factors as indicated by numbers in scale bars.

structures since much information arises from the traveltime of a wave in a layer which trades off between velocity and thickness.

4 APPLICATION TO DATA FROM THE SOUTHERN KOREAN PENINSULA

4.1 Data processing

This section applies the method to several stations on the southern Korean Peninsula. Data are considered for the stations SEO, TJN and GKP1 (Fig. 9) in terms of PV and GV SWD curves, and RF waveforms. The RF are obtained from teleseismic earthquakes in the distance range of 20°–100° with magnitude (Mb) greater than 5.5, and a Gaussian filter width of 2.5. Only RFs with >90 per cent fitness during the deconvolution process are selected for the analysis (Ligorría & Ammon 1999). We apply a statistical ensemble approach to select only similar RF for the stack (Tkalčić *et al.* 2011). The method is based on computing the correlation of pairs of RFs and only those with normalized cross-correlation coefficients >0.9 are retained for stacking.

The SWD curves are obtained from ambient noise tomography for the southern Korean Peninsula. The process of estimating SWD from ambient noise cross-correlation and surface wave tomography is now well established (e.g. Yao *et al.* 2008; Bensen *et al.* 2009; Kim *et al.* 2012) and only briefly described here: (1) Continuous recordings from 2005 to 2009 on 32 broad-band stations (Fig. 9) are subdivided into 1-d-length windows with 50 per cent overlap and the windowed data are normalized in time and spectral domains. (2) Between pairs of stations, cross-correlation functions are calculated for all available data to form empirical Green's functions

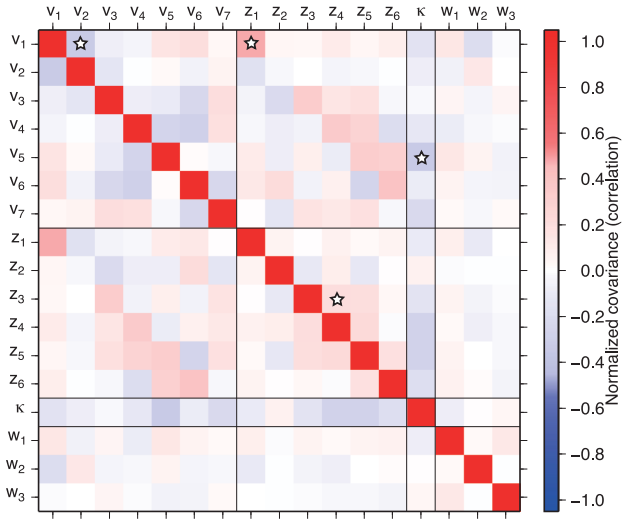

Figure 7. Posterior model covariance (correlation) matrix in the synthetic experiment. Subscript numbers indicate layers from top to bottom and w_1, w_2, and w_3 are the scaling factors for PV, GV, and RF, respectively. Stars indicate parameter pairs for which joint marginals are presented in Fig. 8.

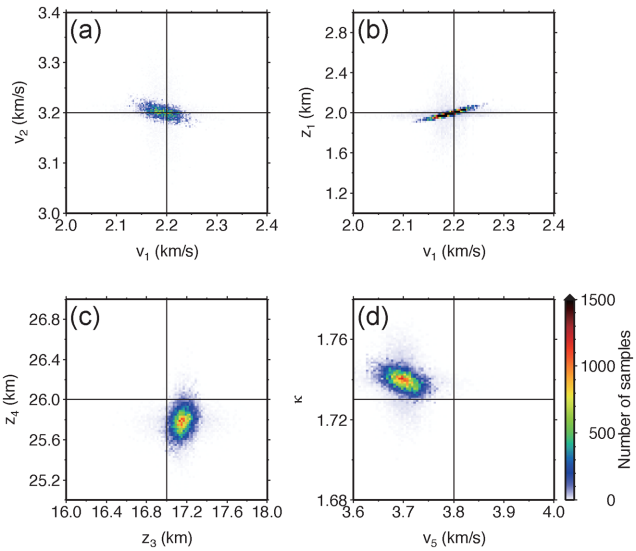

Figure 8. Examples of joint marginal distributions with true value (solid black).

(EGF). (3) The frequency-time analysis is performed with quality-control procedures to estimate PV and GV from the ambient noise EGFs (e.g. Bensen *et al.* 2007; Lin *et al.* 2008). (4) Surface wave tomography (e.g. Kim *et al.* 2012; Saygin & Kennett 2012) is carried out to provide maps showing horizontal variations of PV and GV for wave-periods from 3–30 s. (5) SWD curves at locations of individual stations are composed from the maps by linear interpolation.

4.2 Inversion results and uncertainties

Inversions are carried out with similar settings to those applied in the simulations. Since SWD data are estimates from ambient noise, the number of data-points is limited to 28 and the period range to 3–30 s. Similar to the inversion results in Section 3, computer times for the results at each station are less than one hour.

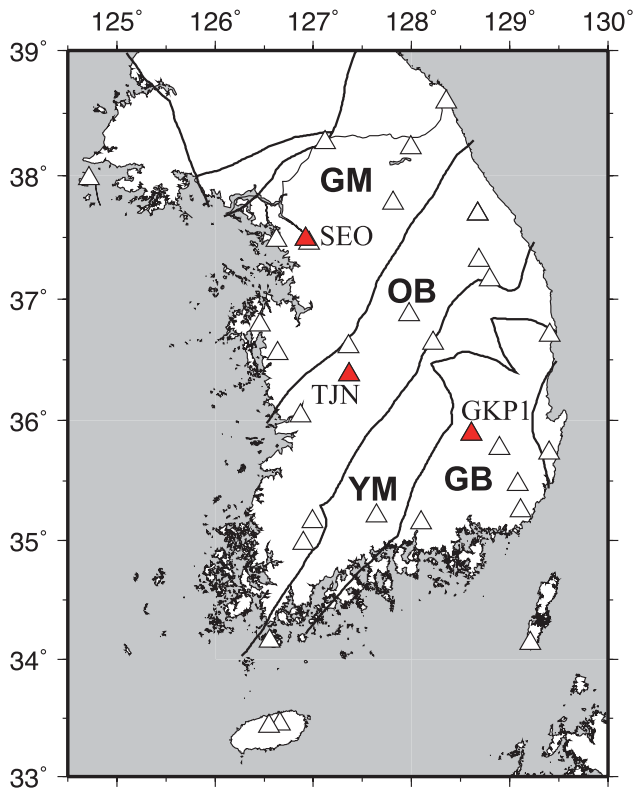

Figure 9. Map of stations on the southern Korean Peninsula used for this study. Joint inversions are carried out for three stations (red triangles), while all the stations (triangles) are used to estimate GV and PV SWD data. Major tectonic regions are also shown (solid lines): Gyeounggi Massif (GM), Okcheon Belt (OB), Yeongnam Massif (YM), and Gyeongsang Basin (GB).

Fig. 10 presents marginals for V_S, interface probability, and V_P/V_S ratio. The trans-D model selection identified significantly different parametrization complexity for the various stations. The optimal numbers of layers above the half-space are 6, 9, and 11 for SEO, TJN and GKP1, respectively. Therefore, it would not be desirable to make an *ad-hoc* decision, assuming the same number of layers for each station. In addition, note that visual inspection of the data does not provide sufficient insight to decide on an optimal parametrization. While the three stations are located in different tectonic regions, V_S structure appears to be similar and the depths of velocity discontinuities are clearly resolved by the inversions. The uppermost layers at all sites show low V_S values ($<$3.0 km s^{-1}) for depths of $<$1 km. The crustal structure is generally simple and ranging from 3.4 to 3.8 km s^{-1}. Stations SEO and TJN have a clear upper crustal discontinuity at $\sim$5–10 km depth. This part of the crust is more complex at GKP1. In addition, a weaker but clear discontinuity is observed at mid-crustal depths ($\sim$20 km) in TJN and GKP1. The V_S estimates are substantially different between sites in the lower crust and uppermost mantle layers. Station GKP1 have higher V_S (3.8–4.0 km s^{-1}) in the lower crust, while other stations have relatively uniform mid-to-lower crustal V_S (3.5–3.8 km s^{-1}). In the uppermost mantle, relatively low V_S (4.2–4.4 km s^{-1}) is estimated at GKP1 compared to higher values at other stations (4.4–4.6 km s^{-1}). Therefore, the Moho velocity contrasts vary significantly. The Moho is shallowest at SEO ($\sim$29 km) and $>$30 km for all other stations. Notably, the Moho is not resolved as a sharp discontinuity at GKP1. Rather, interface probability suggests that the data resolve the discontinuity over an up to 10-km thick region.

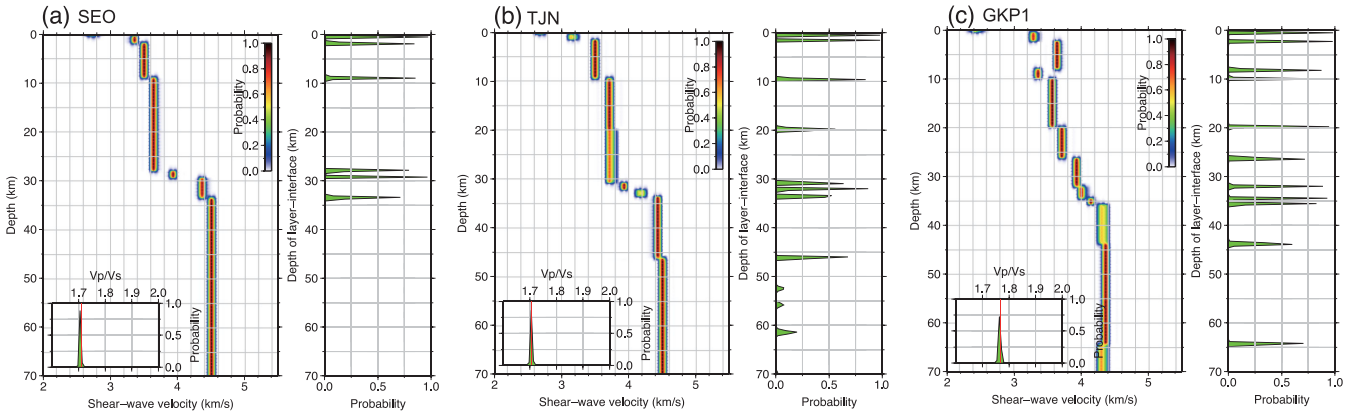

Figure 10. Inversion results in terms of marginal profiles, interface probability, and V_P/V_S 1-D marginals for three stations.

Where the discontinuity is least sharp, the inversion infers a gradual transition from crust to mantle.

The V_P/V_S are well resolved (Fig. 10) and probabilities are significantly higher for SEO and TJN. However, note that these low uncertainties likely do not suggest high sensitivity to *in-situ* V_P. Rather it is commonly observed that extremely simple parametrizations such as selected here (V_P/V_S is not a function of depth) produce low uncertainties. The V_P/V_S values in individual layers may well be significantly outside these narrow uncertainty estimates, since the parametrization only constrain the *average* value. Therefore, interpretation of these values should be conservative.

The range of data predictions produced by the PPD samples and observed data are shown in Fig. 11. The data appear to be fit well, and the range of predictions generally include the observations. The GV data fit at station GKP1 shows a discrepancy at periods shorter than 8 s, while PV and RF data are fit well. The poorer fit for GV data suggests a possible error in measurement processes, resulting in inconsistent information between the various data types.

The correlation matrices in Fig. 12 show strong correlations between many parameters. The correlations appear to be much stronger than in the synthetic inversion which may be explained by the higher model complexity with many thin layers and also by potentially more complicated data errors which are not fully accounted for by the empirical-hierarchical model. For example, z in the sixth to eighth layers at TJN are strongly correlated and correspond to Moho depths and upper mantle.

The joint marginals show examples of several strong positive correlations for layers around the Moho depth (Figs 13b and c). It is evident that both strong trade-offs and multi-modal distributions exist, emphasizing the importance to apply nonlinear inversion for this problem. Note that some strong correlations may be caused by layered parametrizations to resolve (Moho) gradient structure. At station SEO (Fig. 13a), negative correlation is shown between v_1 and z_2, suggesting a trade-off to compensate for slower velocity of the surface layer.

4.3 Comparison with previous studies and tectonic interpretation

This section compares our results in terms of posterior marginal-mean profiles to previous studies, including an average model for the entire southern Korean Peninsula and regional models for its tectonic regions (Kim *et al.* 2011), and results from joint inversion with genetic algorithms (GA, Chang *et al.* 2004; Chang & Baag 2005). While the average and regional models were obtained from broad-band regional waveform grid-search, the GA results are for individual stations from joint RF and PV data inversion. The southern Korean Peninsula is composed of two Precambrian massifs, the Yeongnam Massif (YM) and the Gyeounggi Massif (GM). The Palaeozoic and the Mesozoic orogenic events between the YM and GM resulted in the NE–SW trending Okcheon Belt (OB), formed by accretion, folding and shearing. From the early Cretaceous the Gyeongsang Basin (GB) formed by subduction of oceanic plates, which involved the development of a pull-apart basin mainly in the southeastern part and the east of the peninsula (Chough *et al.* 2000). The three stations are located in the three different tectonic regions (Fig. 9) GM, OB and GB.

Comparison of our results with the average and regional models in Fig. 14 shows clear similarities in terms of absolute velocities and crustal boundaries. The crustal interfaces estimated in our work at depths between 5 and 12 km (Fig. 14) agree reasonably well with the 7.5-km discontinuity in the average model. We further note that these interface depths as well as the velocities of the adjacent layers also largely agree with the regional models. However, our results contain better resolved features and are more straightforward to interpret. For example, our results clearly resolve shallow low-velocity layers and gradients of velocity discontinuities, which are difficult to resolve by using grid-search methods with a small number of layers and no quantitative model selection.

The comparison of our results to the GA results shows significant advantages of our method. While similarity is expected due to the data observations being similar in both cases, the differences highlight the significantly different inversion approaches. The GA results show clear effects that are due to overparametrization and the required global regularization (smoothest model) (Chang *et al.* 2004). As a result of overparametrization and regularization, the GA models exhibit more gradual changes as a function of depth and also some unconstrained structure where the velocity estimate jumps back and forth from layer to layer (e.g. lower crust at TJN and GKP1). The high velocity estimates in the lower crust in the GA models are likely also due to the smoothness constraint which prevents strong velocity discontinuities. At the same time, several low velocity layers (e.g. 5–10 km at SEO) and velocity jumps near sharp discontinuities (e.g. around Moho depth in SEO) appear to be unrealistic in the GA results. The problem that smoothest model regularization produce both too smooth and too rough results is common and rooted in the *global* nature of the regularization. In practice, such features could be interpreted as actual structure. However, after overcompensating near the Moho, the GA results do revert

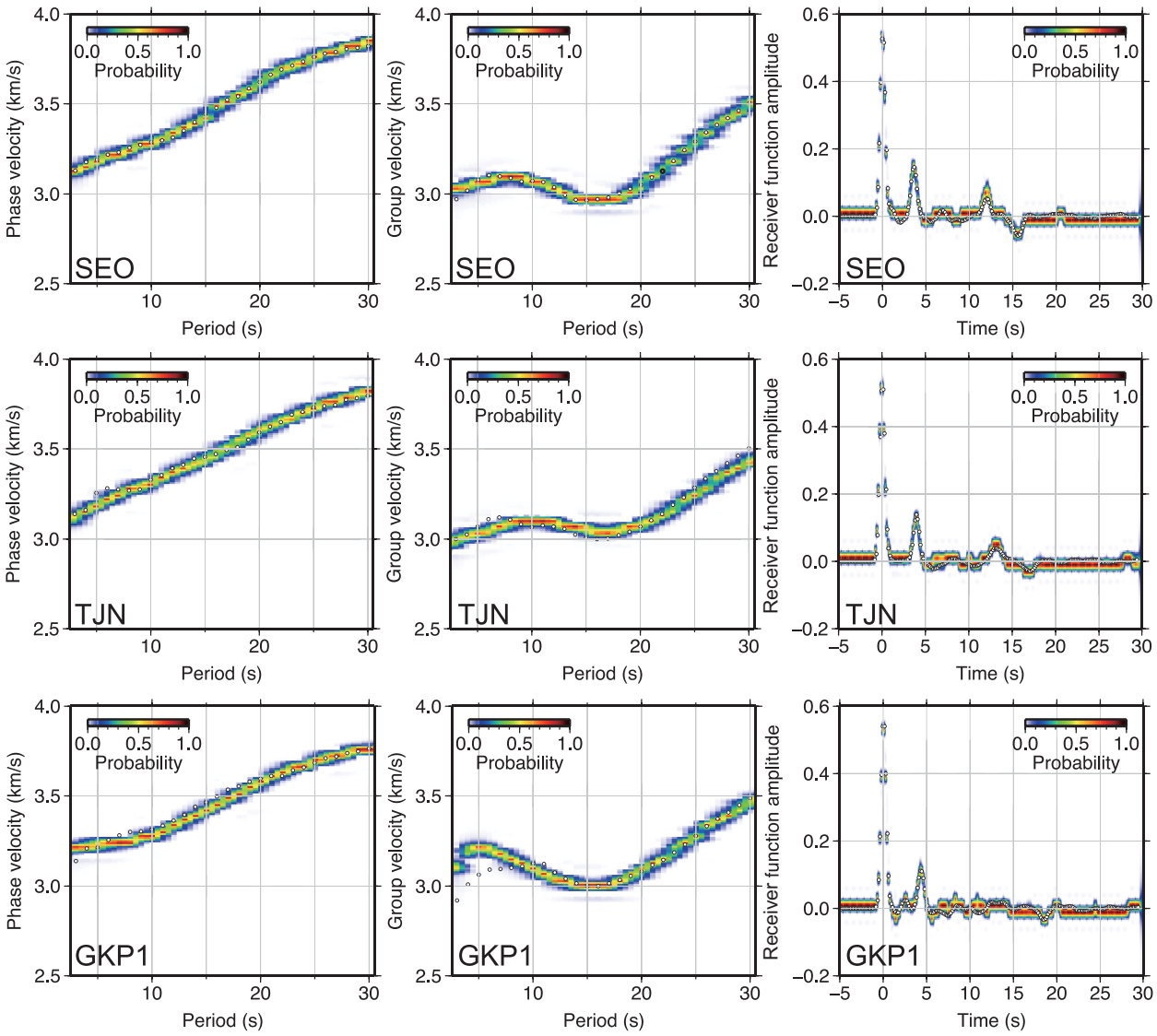

Figure 11. Range of data predictions for the PPD and observed data (white circles).

to estimates similar to our results. The work presented here applies Bayesian model selection to identify an *optimal* (parsimonious) parametrization such that the inversion method does not require any form of regularization and resulting models are more straightforward to interpret. Importantly, this quantitative model selection results in rigorous uncertainty estimation, although we note that optimal models are not as general as trans-D models (our approach fixes the parametrization after trans-D optimization for efficiency reasons which can result in lower uncertainty estimates).

Although the number of observation is limited in this study, the crustal boundaries have clear velocity contrasts (∼0.2 km s^{-1}) that are consistent between stations. These discontinuities may suggest boundaries in rock composition (Chang & Baag 2005) that are observable at multiple stations. However, the boundary depths are shallower than previously detected mid-crustal (∼15–20 km) discontinuities (e.g. Cho *et al.* 2006; He & Hong 2010). These deeper discontinuities may be present in our model as weak velocity increases near 20-km depth at stations TJN and GKP1 (Figs 10b and d). Since the crustal discontinuities are observed in all tectonic regions, we interpret that they are developed during or after the formation of the Korean Peninsula by a common geological processes in a compressional regime, such as an amphibolite-granulite transitions which are abundant in the central Korean Peninsula (e.g. Lee *et al.* 2000). For the GB region, the boundary is overlain by a low-velocity layer, an upper crust (∼3–8 km) with relatively (when compared to the other stations) higher velocity (∼3.6 km s^{-1}), and a deeper shallow crustal boundary (∼4 km). Therefore, the GB has a thicker sedimentary structure and a relatively high V_S in the upper crust. However, the slower layer around 9-km depth is difficult to interpret because no evidence about layering of partial melting or slower materials is reported at this depth. The thicker sedimentary structure or dipping of the sediment-upper crust boundary could affect RF waveform data in this region. The marginal-mean models at SEO and TJN (Fig. 14) exhibit a relatively simple, slower (by ∼0.2 km s^{-1}) lower crust and sharp Moho boundaries compared to GKP1.

Our results estimate the depths of the Moho discontinuity between 30 and 34 km, which is similar to the variation in regional

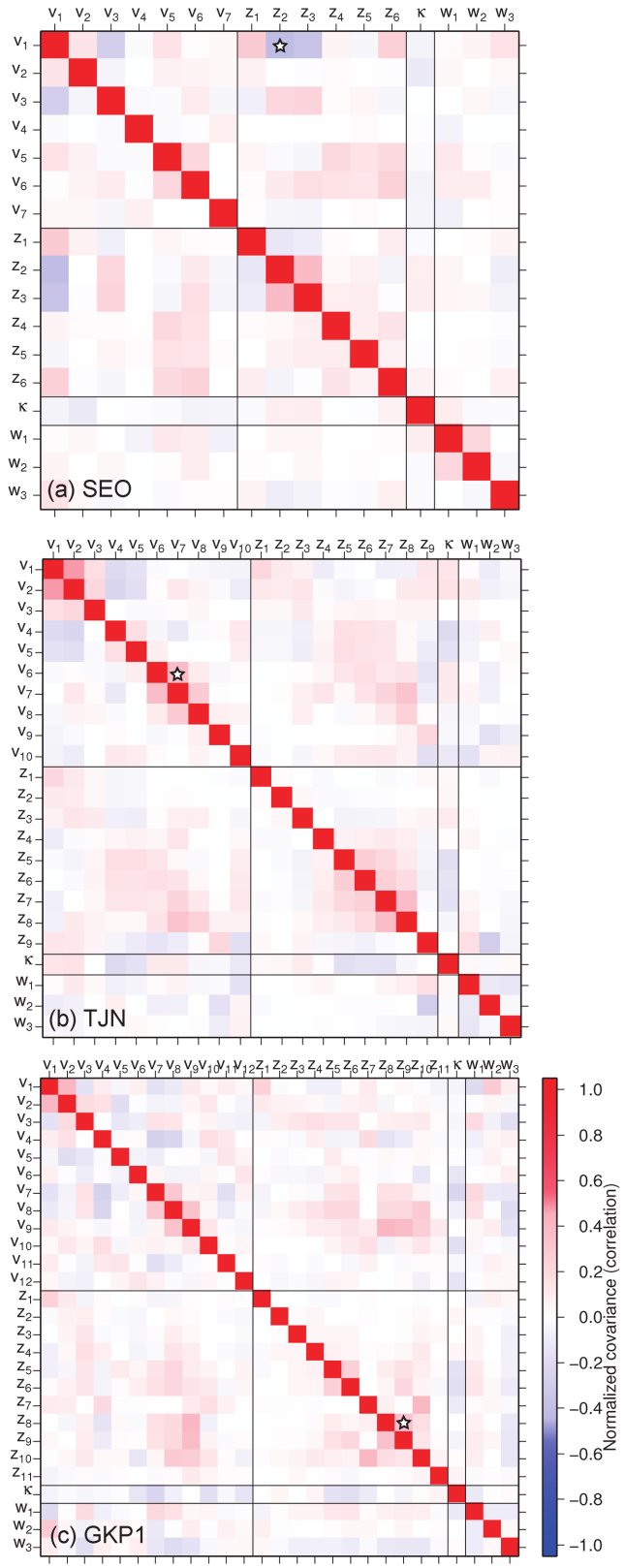

Figure 12. Estimated model covariance matrices, where stars indicate parameter pairs for which joint marginals are presented in Fig. 13.

models and agrees with previous studies using RF data (Chang & Baag 2007; Yoo *et al.* 2007) where the centre of the YM and OB regions, and the west of the GB region have thicker crust. However, the Moho structure exhibits some form of gradient at each station (sharpest transition at SEO, smoothest transition at GKP1). This illustrates an advantage of applying quantitative model selection in our study: The parametrization is adapted to the information content of the data and permits more complex models that result in more elaborate information about V_S where required by the data.

The velocities below the Moho are similar compared to the average and regional models at SEO and TJN. However, GKP1 show substantial differences (>0.2 km s^{-1}) that are much larger than our uncertainty estimates and suggest a regional feature. In addition, the southeastern station (GKP1) have lower crustal structures with significantly high V_S (>0.3 km s^{-1}) and a less clear Moho discontinuity compared to the stations in the west of the southern Korean Peninsula (SEO and TJN). These features in the southeastern part of the Korean Peninsula distinguish them from stations in other parts. Magmatic underplating has been suggested beneath the GB (Cho *et al.* 2004; Chang & Baag 2005), which can be associated with lower crustal modification and relatively hot upper mantle (e.g. Zheng *et al.* 2011).

5 CONCLUSIONS

This work developed an efficient and convenient method to carry out rigorous uncertainty estimation for receiver-side structure for stations in dense regional networks. The method is based on the probabilistic joint inversion of Rayleigh-wave phase-velocity and group-velocity, and RF data. The inversion operates in two phases, where Phase 1 carries out a trans-D optimization and quantitative model selection based on an automated application of the Bayesian information criterion. Phase 2 uses the optimal model identified in Phase 1 and carried out posterior sampling to estimate parameter uncertainties. The approach is much more efficient than full trans-D sampling but lacks some of the generality of trans-D uncertainty estimation. However, the model selection we apply is efficient, quantitative, and practical for potentially large numbers of stations which is a significant advance over the commonly applied *ad-hoc* choices for parametrizations.

Parallel tempering is applied as both an efficient parallel optimization algorithm and to improve sampling efficiency in posterior estimation. The application of a BIC-based objective function during optimization favours simple models which leads to faster convergence to an optimum parametrization (avoids computer time spent on exploring less probable models with excessive complexity). In the sampling phase, empirical data covariance estimates (based on data residuals) are updated by a hierarchical magnitude scaling to avoid biases due to the model assumption (assuming uncorrelated errors of unknown standard deviation) that is required to obtain the residuals.

The inversion is first applied to simulated data with realistic correlated and non-stationary errors. The inferred V_S profile marginals are close to the true model and exhibit reasonable uncertainty estimates. In addition, the range of data predictions for the PPD capture all data features well.

Most importantly, the method is applied to observed data from three stations on the southern Korean Peninsula. We obtain SWD data from ambient noise analysis using continuous broad-band recordings from 32 stations. These data are then jointly inverted

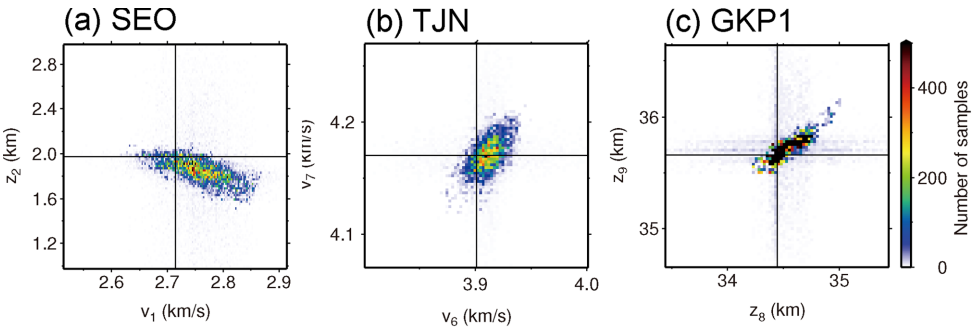

Figure 13. Examples of joint marginals for parameter pairs indicated by stars in Fig. 12. Map values are presented by vertical and horizontal lines.

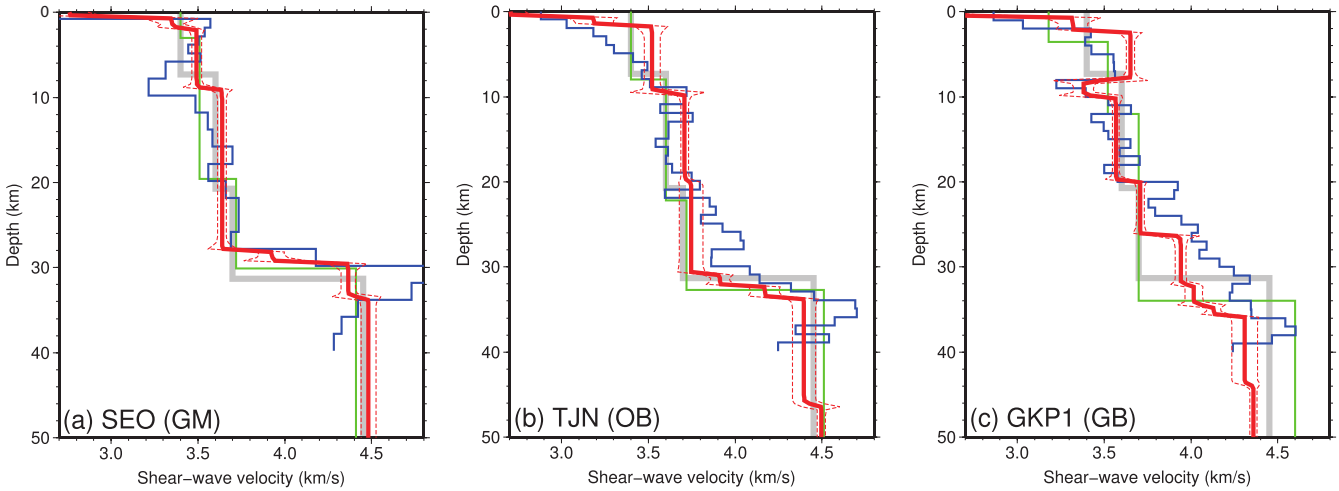

Figure 14. Comparison of marginal mean models (solid red) and ±2 standard deviation bounds (dashed red) from this study with previously published results. Thick grey lines are averages for the entire southern Korean Peninsula and green lines are average models of the tectonic regions (Kim *et al.* 2011). Models estimated by Genetic algorithms (Chang *et al.* 2004; Chang & Baag 2005) are in blue.

with RF data at the three stations to demonstrate the inversion method. The results are consistent with previously published results but are much more straightforward to interpret and lead to tectonical insights. When compared to results of regularized inversions, our results resolve features better. In particular, discontinuities are clearer and more consistent across the three stations and show less spurious velocity structure. From these results, we can support the theory of magma-underplating beneath the southeastern part of the Korean peninsula.

ACKNOWLEDGEMENTS

The authors wish to thank two anonymous reviewers for their constructive comments. This work was supported by the Contract Number FA9453-13-C-0268. JR was funded by the Korea Meteorological Administration and Development Program under Grant KMIPA2015-3052. Figures presented in this paper are formed using the Generic Mapping Tool. We used data from broad-band stations operated by the Korea Institute of Geoscience and Mineral Resources, the Korea Meteorological Administration, the IRIS (INCN), and the F-net (IZH). Inversions were performed on the Terrawulf III computational facility supported through the AuScope Australian Geophysical Observing System. AusScope is funded under the National Collaborative Research Infrastructure Strategy, and the Education Investment Fund (EIF3), both Australian Commonwealth Government Programs.

REFERENCES

Aki, K. & Richards, P.G., 2002. *Quantitative Seismology,* 2nd edn, University Science Books.

Ammon, C.J., 1991. The isolation of receiver effects from teleseismic P waveforms, *Bull. seism. Soc. Am.,* **81**(6), 2504–2510.

Ammon, C.J., Randall, G.E. & Zandt, G., 1990. On the nonuniqueness of receiver function inversions, *J. geophys. Res.,* **95**(B10), 15 303–15 318.

Bensen, G.D., Ritzwoller, M.H., Barmin, M.P., Levshin, A.L., Lin, F., Moschetti, M.P., Shapiro, N.M. & Yang, Y., 2007. Processing seismic ambient noise data to obtain reliable broad-band surface wave dispersion measurements, *Geophys. J. Int.,* **169**(3), 1239–1260.

Bensen, G.D., Ritzwoller, M.H. & Yang, Y., 2009. A 3-D shear velocity model of the crust and uppermost mantle beneath the United States from ambient seismic noise, *Geophys. J. Int.,* **177**(3), 1177–1196.

Bodin, T., Sambridge, M., Tkalčić, H., Arroucau, P., Gallagher, K. & Rawlinson, N., 2012. Transdimensional inversion of receiver functions and surface wave dispersion, *J. geophys. Res.,* **117**(B2), B02301, doi:10.1029/2011JB00856.

Brocher, T.M., 2005. Empirical relations between elastic wavespeeds and density in the Earth's crust, *Bull. seism. Soc. Am.,* **95**(6), 2081–2092.

Brooks, S.P., Friel, N. & King, R., 2003. Classical model selection via simulated annealing, *J. R. Stat. Soc. B,* **65**(2), 503–520.

Chang, S.J. & Baag, C.E., 2005. Crustal structure in southern Korea from joint analysis of teleseismic receiver functions and surface-wave dispersion, *Bull. seism. Soc. Am.,* **95**(4), 1516–1534.

Chang, S.-J. & Baag, C.E., 2007. Moho depth and crustal *Vp/Vs* variation in southern Korea from teleseismic receiver functions: implication for tectonic affinity between the Korean Peninsula and China, *Bull. seism. Soc. Am.,* **97**(5), 1621–1631.

Chang, S.J., Baag, C.E. & Langston, C.A., 2004. Analysis of teleseismic receiver functions and surface wave dispersion using the genetic algorithm, *Bull. seism. Soc. Am.,* **94**(2), 691–704.

Cho, H.-M., Kim, H.-J., Jou, H.-T., Hong, J.-K. & Baag, C.E., 2004. Transition from rifted continental to oceanic crust at the southeastern Korean margin in the East Sea (Japan Sea), *Geophys. Res. Lett.,* **31**(7), L07606, doi:10.1029/2003GL019107.

Cho, H.-M., Baag, C.E., Lee, J.M., Moon, W.M., Jung, H., Kim, K.Y. & Asudeh, I., 2006. Crustal velocity structure across the southern Korean Peninsula from seismic refraction survey, *Geophys. Res. Lett.,* **33**(6), L06307, doi:10.1029/2005GL025145.

Chough, S.K., Kwon, S.T., Ree, J.H. & Choi, D.K., 2000. Tectonic and sedimentary evolution of the Korean peninsula: a review and new view, *Earth-Sci. Rev.,* **52**(1–3), 175–235.

Dettmer, J. & Dosso, S.E., 2012. Trans-dimensional matched-field geoacoustic inversion with hierarchical error models and interacting Markov chains, *J. acoust. Soc. Am.,* **132**(4), 2239–2250.

Dettmer, J., Dosso, S.E. & Holland, C.W., 2007. Uncertainty estimation in seismo-acoustic reflection travel time inversion, *J. acoust. Soc. Am.,* **122**(1), 161–176.

Dettmer, J., Molnar, S., Steininger, G., Dosso, S.E. & Cassidy, J.F., 2012. Trans-dimensional inversion of microtremor array dispersion data with hierarchical autoregressive error models, *Geophys. J. Int.,* **188**(2), 719–734.

Dettmer, J., Dosso, S.E., Bodin, T., Stipčević, J. & Cummins, P.R., 2015. Direct-seismogram inversion for receiver-side structure with uncertain source–time functions, *Geophys. J. Int.,* **203**(2), 1373–1387.

Dosso, S.E. & Dettmer, J., 2011. Bayesian matched-field geoacoustic inversion, *Inverse Probl.,* **27,** 1–23.

Dosso, S.E. & Wilmut, M.J., 2006. Data uncertainty estimation in matched-field geoacoustic inversion, *IEEE J. Ocean. Eng.,* **31**(2), 470–479.

Dosso, S.E., Holland, C.W. & Sambridge, M., 2012. Parallel tempering for strongly nonlinear geoacoustic inversion, *J. acoust. Soc. Am.,* **132**(5), 3030–3040.

Dosso, S.E., Dettmer, J., Steininger, G. & Holland, C.W., 2014. Efficient trans-dimensional Bayesian inversion for geoacoustic profile estimation, *Inverse Probl.,* **30**(11), 114018, doi:10.1088/0266-5611/30/11/114018.

Douma, H., Snieder, R. & Lomax, A., 1996. Ensemble inference in terms of empirical orthogonal functions, *Geophys. J. Int.,* **127**(2), 363–378.

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

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

Efron, B. & Tibshirani, R., 1991. Statistical data analysis in the computer age, *Science,* **253**(5018), 390–395.

Geyer, C.J., 1991. Markov chain Monte Carlo maximum likelihood, in *Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface,* pp. 156–163, Interface Foundation of North America.

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

Gouveia, W.P. & Scales, J.A., 1998. Bayesian seismic waveform inversion: parameter estimation and uncertainty analysis, *J. geophys. Res.,* **103**(B2), 2759–2779.

Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika,* **82**(4), 711–732.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications, *Biometrika,* **57**(1), 97–109.

He, X. & Hong, T.-K., 2010. Evidence for strong ground motion by waves refracted from the Conrad discontinuity, *Bull. seism. Soc. Am.,* **100**(3), 1370–1374.

Jackson, D.D. & Matsu'ura, M., 1985. A Bayesian approach to nonlinear inversion, *J. geophys. Res.,* **90**(B1), 581–591.

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

Kim, S., Rhie, J. & Kim, G., 2011. Forward waveform modelling procedure for 1-D crustal velocity structure and its application to the southern Korean Peninsula, *Geophys. J. Int.,* **185**(1), 453–468.

Kim, S., Nyblade, A.A., Rhie, J., Baag, C.E. & Kang, T.S., 2012. Crustal *S*-wave velocity structure of the Main Ethiopian Rift from ambient noise tomography, *Geophys. J. Int.,* **191**(2), 865–878.

Langston, C.A., 1979. Structure under Mount Rainier, Washington, inferred from teleseismic body waves, *J. geophys. Res.,* **84**(B9), 4749–4762.

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

Lee, S.R., Cho, M., Yi, K. & Stern, R.A., 2000. Early Proterozoic granulites in central Korea: tectonic correlation with Chinese cratons, *J. Geol.,* **108**(6), 729–738.

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

Lin, F.-C., Moschetti, M.P. & Ritzwoller, M.H., 2008. Surface wave tomography of the western United States from ambient seismic noise: Rayleigh and Love wave phase velocity maps, *Geophys. J. Int.,* **173**(1), 281–298.

Lomax, A. & Snieder, R., 1995. The contrast in upper mantle shear-wave velocity between the east european platform and tectonic Europe obtained with genetic algorithm inversion of rayleigh-wave group dispersion, *Geophys. J. Int.,* **123**(1), 169–182.

Lomax, A. & Snieder, R., 2012. Finding sets of acceptable solutions with a genetic algorithm with application to surface wave group dispersion in Europe, *Geophys. Res. Lett.,* **21**(24), 2617–2620.

MacKay, D.J.C., 2003. *Information Theory, Inference & Learning Algorithms,* Cambridge Univ. Press.

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

Malinverno, A. & Briggs, V.A., 2004. Expanded uncertainty quantification in inverse problems: hierarchical Bayes and empirical Bayes, *Geophysics,* **69**(4), 1005–1016.

Martinez, M.D., Lana, X., Olarte, J., Badal, J. & Canas, J.A., 2000. Inversion of Rayleigh wave phase and group velocities by simulated annealing, *Phys. Earth. planet. Inter.,* **122**(1–2), 3–17.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E., 1953. Equation of state calculations by fast computing machines, *J. Chem. Phys.,* **21**(6), 1087–1092.

Molnar, S., Dosso, S.E. & Cassidy, J.F., 2010. Bayesian inversion of microtremor array dispersion data in southwestern British Columbia, *Geophys. J. Int.,* **183**(2), 923–940.

Mosegaard, K. & Sambridge, M., 2002. Monte Carlo analysis of inverse problems, *Inverse Probl.,* **18**(3), R29–R54.

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

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

Randall, G.E., 1989. Efficient calculation of differential seismograms for lithospheric receiver functions, *Geophys. J. Int.,* **99**(3), 469–481.

Rodi, W.L., Glover, P., Li, T. M.C. & Alexander, S.S., 1975. A fast, accurate method for computing group-velocity partial derivatives for Rayleigh and Love modes, *Bull. seism. Soc. Am.,* **65**(5), 1105–1114.

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, ed. Doornbos, J.D., Academic Press.

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

Sambridge, M., 2013. A Parallel Tempering algorithm for probabilistic sampling and multimodal optimization, *Geophys. J. Int.,* **196**(1), 357–374.

Sambridge, M. & Mosegaard, K., 2002. Monte Carlo methods in geophysical inverse problems, *Rev. Geophys.,* **40**(3), 3-1–3-29.

Sambridge, M., Beghein, C., Simons, F.J. & Snieder, R., 2006a. How do we understand and visualize uncertainty?, *Leading Edge,* **25**(5), 542–546.

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

Saygin, E. & Kennett, B.L.N., 2012. Crustal structure of Australia from ambient seismic noise tomography, *J. geophys. Res.,* **117**(B1), B01304, doi:10.1029/2011JB008403.

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

Scales, J.A. & Snieder, R., 2000. The anatomy of inverse problems, *Geophysics,* **65**(6), 1708–1710.

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

Shapiro, N.M. & Ritzwoller, M.H., 2002. Monte-Carlo inversion for a global shear-velocity model of the crust and upper mantle, *Geophys. J. Int.,* **151**(1), 88–105.

Shen, W., Ritzwoller, M.H. & Schulte-Pelkum, V., 2013a. A 3-D model of the crust and uppermost mantle beneath the Central and Western US by joint inversion of receiver functions and surface wave dispersion, *J. geophys. Res.,* **118**(1), 262–276.

Shen, W., Ritzwoller, M.H., Schulte-Pelkum, V. & Lin, F.C., 2013b. Joint inversion of surface wave dispersion and receiver functions: a Bayesian Monte-Carlo approach, *Geophys. J. Int.,* **192**(2), 807–836.

Snieder, R., 1998. The role of nonlinearity in inverse problems, *Inverse Probl.,* **14**(3), 387–404.

Takeuchi, H., Dorman, J. & Saito, M., 1964. Partial derivatives of surface wave phase velocity with respect to physical parameter changes within the Earth, *J. geophys. Res.,* **69**(16), 3429–3441.

Tanimoto, T., 1991. Waveform inversion for three-dimensional density and *S* wave structure, *J. geophys. Res.,* **96**(B5), 8167–8189.

Tarantola, A., 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation,* SIAM.

Tkalčić, H., Pasyanos, M.E., Rodgers, A.J., Goek, R., Walter, W.R. & Al-Amri, A., 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**(B11), doi:10.1029/2005JB004130.

Tkalčić, H., Chen, Y., Liu, R., Zhibin, H., Sun, L. & Chan, W., 2011. Multistep modelling of teleseismic receiver functions combined with constraints from seismic tomography: crustal structure beneath southeast China, *Geophys. J. Int.,* **187**(1), 303–326.

Yao, H., Beghein, C. & van der Hilst, R.D., 2008. Surface wave array tomography in SE Tibet from ambient seismic noise and two-station analysis – II. Crustal and upper-mantle structure, *Geophys. J. Int.,* **173**(1), 205–219.

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

Zheng, Y., Shen, W., Zhou, L., Yang, Y., Xie, Z. & Ritzwoller, M.H., 2011. Crust and uppermost mantle beneath the North China Craton, northeastern China, and the Sea of Japan from ambient noise tomography, *J. geophys. Res.,* **116**(B12), B12312, doi:10.1029/2011JB008637.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this paper:

Figure S1. Examples of the Poisson prior distribution for different values of the scale parameter (λ).

Figure S2. The same with Fig. 4 for synthetic inversion results, but from the inversion with the fully trans-dimensional sampling. Note that four times longer chains are used to sample this posterior distribution (http://gji.oxfordjournals.org/lookup/suppl/doi:10.1093/gji/ggw149/-/DC1).

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the paper.

APPENDIX A: RESIDUAL EXAMINATION

In Bayesian inversion, uncertainty estimates critically depend on the formulation of the likelihood function and, by extension, on reasonable estimates of the data errors. These are typically a combination of theory and measurement errors that cannot be

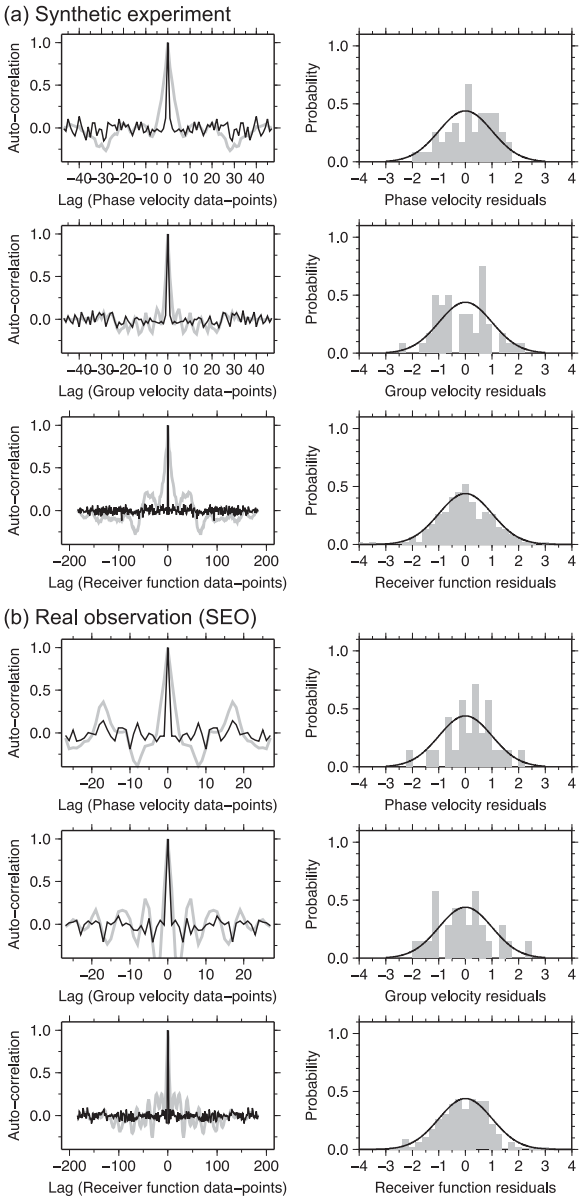

Figure A1. Examples of residual analysis in terms of (a) residual autocorrelation functions of raw residuals (grey) and standardized residuals (black), and (b) residual histograms of standardized residuals (grey) and standard normal distribution (black). Note that from −5 to 0 s of the RF waveforms excluded in the residual analysis.

separated. In this work, data covariance matrices are initially estimated from residual-error statistics (which include the effects of measurement and theory errors and non-Gaussianity). The initial estimates are updated by magnitude scaling parameters that are unknowns in the uncertainty estimation. This approach provides a relatively simple approach to capture complex covariances while avoiding large numbers of parameters to model such covariances. The validity of the covariance estimates can be examined by *a posteriori* residual analysis (Dettmer *et al.* 2007). Fig. A1 shows examples of the residual analysis for the simulation (Fig. A1a) in Section 3, and for the observations at station SEO (Fig. A1b) in Section 4. From the residuals for the MAP model, standardized residuals are obtained by multiplication with the inverse of the Cholesky decomposition of the scaled data covariance matrices (Dettmer *et al.* 2007). If the covariance estimate quantifies the error process reasonably well, standardized residuals are not correlated. This can be examined by considering the autocorrelation function (left panels in Fig. A1), which has a centre-peak width of only one point for uncorrelated residuals. In addition, the standardized residuals should be close to a zero-mean Gaussian distribution of unit standard deviation (right panels in Fig. A1), if the Gaussian assumptions is valid. Fig. A1 shows that the covariance estimates account for the most significant correlations and that residuals appear to be reasonably close to Gaussian, with no significant outliers.

APPENDIX B: UNCERTAINTY QUANTIFICATION

To quantify inversion uncertainty, the PPD should be analysed using statistical approaches including integrals of high-dimensional probabilities. The MH sampling procedure yields an ensemble of models that asymptotically represents the PPD. Therefore, inferring the ensemble provides statistical approaches to analyse the PPD without explicit high-dimensional integrations as (e.g. Dosso & Dettmer 2011),

$$I = \int f(\mathbf{m}) p(\mathbf{m}|\mathbf{d}) d\mathbf{m} \approx \frac{1}{Q} \sum_{i=1}^{Q} f(\mathbf{m}_i) \tag{B1}$$

where f is a specific function applied to examine the PPD and Q is number of samples drawn via the MH.

Individual models are useful that represents the PPD to interpret physical structures in many applications using geophysical data. The MAP model maximizes the posterior probability, or simply the likelihood function for the case of uniform priors:

$$\hat{\mathbf{m}} = argmax\,\{p(\mathbf{m}|\mathbf{d})\} = argmax\,\{L(\mathbf{m})\}\,. \tag{B2}$$

And, the posterior mean model $\bar{\mathbf{m}}$ is defined as,

$$\bar{\mathbf{m}} = \int \mathbf{m} p(\mathbf{m}|\mathbf{d}) d\mathbf{m}. \tag{B3}$$

Important properties of inversion uncertainties are deduced from inter-relationship or covariance between model parameters. The model covariance matrix $\mathbf{C_m}$ provides a quantitative value about co-relationship between each pair of parameters. Here we use the correlation matrix $\mathbf{R}$, which is from normalizing the covariance matrix by corresponding standard deviations. The model covariance matrix is defined as,

$$\mathbf{C_m} = \int (\mathbf{m} - \bar{\mathbf{m}})(\mathbf{m} - \bar{\mathbf{m}})^T p(\mathbf{m}|\mathbf{d}) d\mathbf{m}. \tag{B4}$$

Then, the elements of correlation matrix is obtained by $R_{ij} = C_{mij}/\sqrt{C_{mii} C_{mjj}}$. In the case that the PPD is potentially multimodal and non-Gaussian, the properties of the PPD are not fully explained by the correlation coefficients. The marginal PPDs show probability densities of specific parameters over the entire parameter space. For ith element of the model vector, the marginal PPD is defined by,

$$p(m_i|\mathbf{d}) = \int \delta(m_i - m_i') p(\mathbf{m}'|\mathbf{d}) d\mathbf{m}'. \tag{B5}$$

Regarding with the multi-dimensionality of the PPD, the 'joint' marginal PPD can be estimated by combinations of different parameters. Particularly in layered model parametrizations, we can quantify and visualize the PPD conveniently by using a profile of the marginal PPD of V_S, which is bin-stacked in a tabulated V_S-depth domain with a similar manner to eq. (B5). For an additional individual model deduced from the marginal PPD profile, a marginal mean model is estimated by taking average of V_S values from every depth-bins. The uncertainty of the marginal mean model can be obtained by standard deviation of the V_S distributions at each depth.