The Annals of Applied Statistics
2013, Vol. 7, No. 2, 1111–1138
DOI: 10.1214/12-AOAS623

A BAYESIAN LINEAR MODEL FOR THE HIGH-DIMENSIONAL INVERSE PROBLEM OF SEISMIC TOMOGRAPHY

BY RAN ZHANG[1], CLAUDIA CZADO AND KARIN SIGLOCH

Technische Universität München, Technische Universität München and Ludwig-Maximilians Universität München

We apply a linear Bayesian model to seismic tomography, a high-dimensional inverse problem in geophysics. The objective is to estimate the three-dimensional structure of the earth's interior from data measured at its surface. Since this typically involves estimating thousands of unknowns or more, it has always been treated as a linear(ized) optimization problem. Here we present a Bayesian hierarchical model to estimate the joint distribution of earth structural and earthquake source parameters. An ellipsoidal spatial prior allows to accommodate the layered nature of the earth's mantle. With our efficient algorithm we can sample the posterior distributions for large-scale linear inverse problems and provide precise uncertainty quantification in terms of parameter distributions and credible intervals given the data. We apply the method to a full-fledged tomography problem, an inversion for upper-mantle structure under western North America that involves more than 11,000 parameters. In studies on simulated and real data, we show that our approach retrieves the major structures of the earth's interior as well as classical least-squares minimization, while additionally providing uncertainty assessments.

1. Introduction. Seismic tomography is a geophysical imaging method that allows to estimate the three-dimensional structure of the earth's deep interior, using observations of seismic waves made at its surface. Seismic waves generated by moderate or large earthquakes travel through the entire planet, from crust to core, and can be recorded by seismometers anywhere on earth. They are by far the most highly resolving wave type available for exploring the interior at depths to which direct measurement methods will never penetrate (tens to thousands of kilometers). Seismic tomography takes the shape of a large, linear(ized) inverse problem, typically featuring thousands to millions of measurements and similar numbers of parameters to solve for.

To first order, the earth's interior is layered under the overwhelming influence of gravity. Its resulting, spherically symmetric structure had been robustly estimated by the 1980s [Dziewonski and Anderson (1981), Kennett and Engdahl (1991)] and is characterized by $O(10^2)$ parameters. Since then seismologists have been mainly

Received March 2012; revised September 2012.

[1]Supported by the Munich Center of Advanced Computing (MAC/IGSSE) and by the LRZ Supercomputing Center.

Key words and phrases. High-dimensional inverse problems, seismic tomography, Bayesian linear model, Markov chain Monte Carlo, spatial prior.

concerned with estimating lateral deviations from this spherically symmetric reference model [Nolet (2008)]. Though composed of solid rock, the earth's mantle is in constant motion (the mantle extends from roughly 30 km to 2900 km depth and is underlain by the fluid iron core). Rock masses are rising and sinking at velocities of a few centimeters per year, the manifestation of advective heat transfer: the hot interior slowly loses its heat into space. This creates slight lateral variations in material properties, on the order of a few percent, relative to the statically layered reference model. The goal of seismic tomography is to map these three-dimensional variations, which embody the dynamic nature of the planet's interior.

Beneath well-instrumented regions—such as our chosen example, the United States—seismic waves are capable of resolving mantle heterogeneity on scales of a few tens to a few hundreds of kilometers. Parameterizing the three-dimensional earth, or even just a small part of it, into blocks of that size results in the mentioned large number of unknowns, which mandate a linearization of the inverse problem. Fortunately this is workable, thanks to the rather weak lateral material deviations of only a few percent (larger differences cannot arise in the very mobile mantle).

Seismic tomography is almost always treated as an optimization problem. Most often a least squares approach is followed requiring general matrix inverses [Aki and Lee (1976), Crosson (1976), Montelli et al. (2004), Sigloch, McQuarrie and Nolet (2008)], while adjoint techniques are used when an explicit matrix formulation is computationally too expensive [Fichtner et al. (2009), Sieminski et al. (2007), Tromp, Tape and Liu (2005)]. While probabilistic seismic tomography using Markov chain Monte Carlo (MCMC) methods has been given considerable attention by the geophysical (seismological) community, these applications have been restricted to linear or nonlinear problems of much lower dimensionality assuming Gaussian errors [Mosegaard and Tarantola (1995), Mosegaard and Tarantola (2002), Sambridge and Mosegaard (2002)]. For example, Dębski (2010) compares the damped least-squares method (LSQR), a genetic algorithm and the Metropolis–Hastings (MH) algorithm in a low-dimensional linear tomography problem involving copper mining data. He finds that the MCMC sampling technique provides more robust estimates of velocity parameters compared to the other approaches. Bodin and Sambridge (2009) capture the uncertainty of the velocity parameters in a linear model by selecting the representation grid of the corresponding field, using a reversible jump MCMC (RJMCMC) approach. In Bodin et al. (2012a, 2012b) again RJMCMC algorithms are developed to solve certain transdimensional nonlinear tomography problems with Gaussian errors, assuming unknown variances. Khan, Zunino and Deschamps (2011) and Mosca et al. (2012) study seismic and thermo-chemical structures of the lower mantle and solve a corresponding low-dimensional nonlinear problem using a standard MCMC algorithm.

For exploring high-dimensional parameter space the MCMC sampling faces difficulties in evaluating the expensive nonlinear physical model while efficiently

traversing the high-dimensional parameter space. We approach linearized tomographic problems (physical forward model inexpensive to solve) in a Bayesian framework, for a fully dimensioned, continental-scale study that features ≈53,000 data points and ≈11,000 parameters. To our knowledge, this is by far the highest dimensional application of Monte Carlo sampling to a seismic tomographic problem so far. Assuming Gaussian distributions for the error and the prior, our MCMC sampling scheme allows for characterization of the posterior distribution of the parameters by incorporating flexible spatial priors using Gaussian Markov random field (GMRF). Spatial priors using GMRF arise in spatial statistics [Congdon (2003), Pettitt, Weir and Hart (2002), Rue and Held (2005)], where they are mainly used to model spatial correlation. In our geophysical context we apply a spatial prior to the parameters rather than to the error structure, since the parameters represent velocity anomalies in three-dimensional space. Thanks to the sparsity of the linearized physical forward matrix as well as the spatial prior sampling from the posterior density, a high-dimensional multivariate Gaussian can be achieved by a Cholesky decomposition technique from Wilkinson and Yeung (2002) or Rue and Held (2005). Their technique is improved by using a different permutation algorithm. To demonstrate the method, we estimate a three-dimensional model of mantle structure, that is, variations in seismic wave velocities, beneath the Unites States down to 800 km depth.

Our approach is also applicable to other kinds of travel time tomography, such as cross-borehole tomography or mining-induced seismic tomography [Dębski (2010)]. Other types of tomography, such as X-ray tomography in medical imaging, can also be recast as a linear matrix problem of large size with a very sparse forward matrix. However, the response is measured on pixel areas and, thus, the error structure is governed by a spatial Markov random field, while the regression parameters are modeled nonspatially using, for example, Laplace priors [Kolehmainen et al. (2007), Mohammad-Djafari (2012)]. Some other inverse problems such as image deconvolution and computed tomography [Bardsley (2012)], electromagnetic source problems deriving from electric and magnetic encephalography, cardiography [Hämäläinen and Ilmoniemi (1994), Kaipio and Somersalo (2007), Uutela, Häämäläinen and Somersalo (1999)] or convection-diffusion contamination transport problems [Flath et al. (2011)] can also be written as linear models. However, the physical forward matrix of those problems is dense in contrast to the situation we consider. For solutions to these problems, matrix-inversion or low-rank approximation to the posterior covariance matrix, as introduced in Flath et al. (2011), are applied to high-dimensional linear problems. In image reconstruction problems Bardsley (2012) demonstrates Gibbs sampling on (1D and 2D-) images using an intrinsic GMRF prior with the preconditioned conjugate gradient method in cases where efficient diagonalization or Cholesky decomposition of the posterior covariance matrix is not available. In other tomography problems, such as electrical capacitance tomography, electrical impedance tomography or optical absorbtion and scattering tomography, the physical forward model cannot

be linearized, so that the Bayesian treatment of those problems is limited to low dimensions [Kaipio and Somersalo (2007), Watzenig and Fox (2009)].

The remainder of this paper is organized as follows: Section 2 describes the geophysical forward model and the seismic travel time data. Section 3 discusses flexible specifications for the spatial prior of the three-dimensional velocity model and the Metropolis–Gibbs sampling algorithm for estimating its posterior distribution. Method performance under various model assumptions is examined in simulation studies in Section 4. Section 5 applies the method to real travel time data, which have previously been used in conventional tomography [Sigloch (2011), Sigloch, McQuarrie and Nolet (2008)], allowing for comparison. Section 6 discusses the advantages, limitations and possible extensions of our model.

2. Geophysical models and the data. Here we explain the physics and the data that enter seismic tomography and how they are formulated into a linear inverse problem, which will be treated by our Markov chain Monte Carlo method in subsequent sections.

2.1. *The linear inverse problem of seismic tomography.* Every larger earthquake generates seismic waves of sufficient strength to be recorded by seismic stations around the globe. Such seismograms are time series at discrete surface locations, that is, spatially sparse point samples of a continuous wavefield that exists everywhere inside the earth and at its surface. Figure 1 illustrates the spatial distribution of sources (large earthquakes, blue) and receivers (seismic broadband

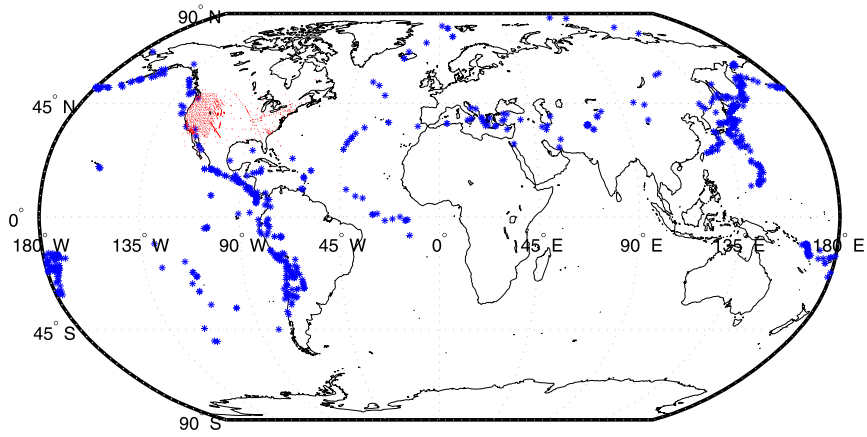

FIG. 1. *Distribution of the seismic wave sources* (*large earthquakes*, *blue*) *and receivers* (*seismic broadband stations*, *red*) *that generated our data. This is a regional tomography study that includes only data recorded in North America. In the mantle under this region*, *down to a few hundreds of kilometers depth*, *paths of incoming waves cross densely and from many directions*, *yielding good resolution for a three-dimensional imaging study.*

stations, red) that generated our data. Each datum y_i measures the difference between an observed arrival time y_i^{obs} of a seismic wave i and its predicted arrival time y_i^{pred}:

$$y_i = y_i^{\text{obs}} - y_i^{\text{pred}}.$$

y_i^{pred} is evaluated using the spherically symmetric reference model IASP91 by Kennett and Engdahl (1991). For the teleseismic P waves used in our application, this difference y_i would typically be on the order of one second, whereas y_i^{obs} and y_i^{pred} are on the order 600–1000 seconds. y_i can be explained by slightly decreasing the modeled velocity in certain sub-volumes of the mantle.

We adopt the parametrization and a subset of the data measured by Sigloch, McQuarrie and Nolet (2008). The earth is meshed as a sphere of irregular tetrahedra with 92,175 mesh nodes. At each mesh mode, the parameters of interest are the relative velocity variation of the mantle with respect to the reference velocity of spherically-symmetric model IASP91 [Kennett and Engdahl (1991)]. The parameter vector is denoted as $\boldsymbol{\beta} := (\beta(\mathbf{r}), \mathbf{r} \in M_{\text{Earth}}) \in \mathbb{R}^{92,175}$, where the set of mesh node M_{Earth} fills the entire interior of the earth. Since both travel time deviations y_i and the $\beta(\mathbf{r})$ are small, the wave equation may be linearized around the layered reference model:

$$y_i = \iiint_{\text{Earth}} x_i(\mathbf{r})\beta(\mathbf{r})\, d^3\mathbf{r}, \tag{1}$$

where $x_i(\mathbf{r}) \in \mathbb{R}$ represents the Fréchet sensitivity kernel of the ith wavepath, that is, the partial derivatives of the chosen misfit measure or data y_i with respect to the parameters $\beta(\mathbf{r})$. After numerical integration of kernel $x_i(\mathbf{r})$ onto the mesh, (1) takes the form

$$y_i = \sum_{\mathbf{r} \in M_{\text{Earth}}} x_i(\mathbf{r})\beta(\mathbf{r}) = \mathbf{x}_i'\boldsymbol{\beta}. \tag{2}$$

Geometrically speaking, row vector $\mathbf{x}_i'$ maps out the mantle subvolume that would influence the travel time y_i if some velocity anomaly $\beta(\mathbf{r})$ were located within it. This sensitivity region between an earthquake and a station essentially has ray-like character (Figure 2), though in physically more sophisticated approximations, the ray widens into a banana shape [Dahlen, Hung and Nolet (2000)]. Over the past decade, intense research effort has gone into the computability of sensitivity kernels under more and more realistic approximations [Dahlen, Hung and Nolet (2000), Nolet (2008), Tian et al. (2007), Tromp, Tape and Liu (2005)]. Since this issue is only tangential to our focus, we chose to keep the sensitivity calculations as simple as possible by modeling them as rays (the $\mathbf{x}_i'$ are computed only once and stored). We note that the dependence of x_i on $\boldsymbol{\beta}$ can be neglected, as is common practice. This is justified by two facts: (i) velocity anomalies $\boldsymbol{\beta}$ deviate from those of the (spherically symmetric) reference model by only a few percent, since

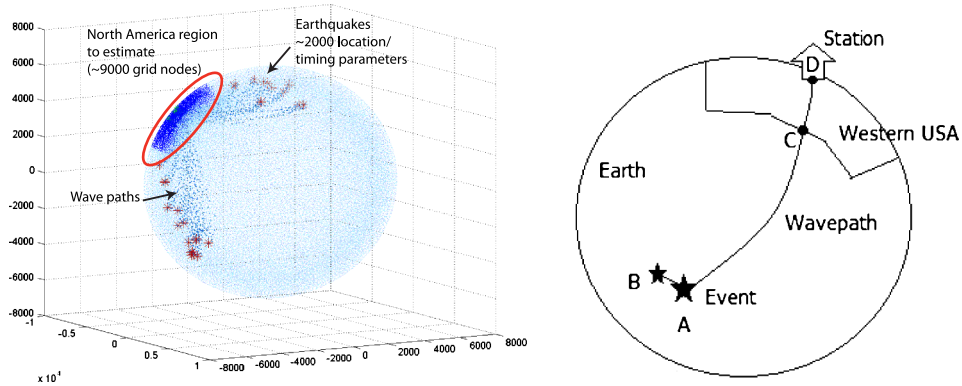

FIG. 2. *Physical setup and forward modeling of the seismic tomography problem. Left: parametrization of the spherical earth. Grid nodes are shown as blue dots. The goal is to estimate seismic velocity deviations* $\boldsymbol{\beta}$ *at* ∼9000 *grid nodes under North America, inside the subvolume marked by the red ellipse. Red stars mark a few of the earthquake sources shown in Figure 1. The densified point clouds, between the sources and a few stations in North America, map out the sensitivity kernels of the selected wave paths. Each sensitivity kernel fills one row of matrix* X*. Left: schematic illustration of the components of an individual wave path.*

the very mobile mantle does not support larger disequilibria, and (ii), even though the ray path in the true earth differs (slightly) from that in the reference model, this variation affects the travel time observable only to second order, according to Fermat's principle [and analogous arguments for true finite-frequency sensitivities, Dahlen, Hung and Nolet (2000), Mercerat and Nolet (2013), Nolet (2008)]. Whatever the exact modeling is, it is very sparse, since every ray or banana visits only a small subvolume of the entire mantle—this sparsity is important for the computational efficiency of the MCMC sampling.

Gathering all N observations, (2) can be rewritten as $\mathbf{y} = X\boldsymbol{\beta}$, where sparse matrix $X \in \mathbb{R}^{N\times d}$ contains in its rows the N sensitivity kernels. The left panel of Figure 2 illustrates the sensitivity kernels between one station and several earthquakes (i.e., several matrix rows). In practice, the problem never attains full rank, so that regularization must be added to remove the remaining nonuniqueness. The linear system $\mathbf{y} = X\boldsymbol{\beta}$ is usually solved by some sparse matrix solver—a popular choice is the Sparse Equations and Least Squares (LSQR) algorithm by Paige and Saunders (1982), which minimizes $\|X\boldsymbol{\beta} - \mathbf{y}\|^2 + \lambda^2\|\boldsymbol{\beta}\|^2$, where λ is a regularization parameter that removes the underdeterminacy in X [Dębski (2010), Montelli et al. (2004), Sigloch, McQuarrie and Nolet (2008), Tian, Sigloch and Nolet (2009)].

In summary, we have formulated the seismic tomography problem as it is overwhelmingly practiced by the geophysical community today. We use travel time differences y_i as the misfit criterion, that is, as input data to the inverse problem, and seek to estimate the three-dimensional distribution of seismic velocity deviations $\boldsymbol{\beta}$ that have caused these travel time anomalies. The sensitivity kernels $\mathbf{x}_i'$ are

modeled using ray theory, a high-frequency approximation to the full wave equation. In the conventional optimization approach, a regularization term is added, and the inverse problem is solved by minimizing the L2 norm misfit.

2.2. *Setup of our example problem.* Since all 92,175 velocity deviation parameters of the entire earth are currently not manageable for MCMC sampling, we regard as free parameters only 8977 of those parameters which are located beneath the western U.S., that is, between latitudes 20°N to 60°N, longitudes 90°W to 130°W, and 0–800 km depth. Tetrahedra nodes are spaced by 60–150 km. We denote this subset of velocity parameters as $\boldsymbol{\beta}_{\text{usa}}$.

Besides velocity parameters, we also consider the uncertainty in the location and the origin time of each earthquake source, which contribute to the travel time measurement. Government and research institutions routinely publish location estimates for every larger earthquake, but any event may easily be mistimed by a few seconds, and mislocated by ten or more kilometers (corresponding to a travel duration of 1 s or more). This is a problem, since the structural heterogeneities themselves only generate travel time delays on the order of a few seconds. Hence, the exact locations and timings of the earthquakes—or rather: their deviations from the published catalogue values—need to be treated as additional free parameters, to be estimated jointly with the structural parameters. These so-called "source corrections" are captured by three-dimensional shift corrections of the hypocenter ($\boldsymbol{\beta}_{\text{hyp}}$) and time corrections ($\boldsymbol{\beta}_{\text{time}}$) per earthquake.

Using the LSQR method, Sigloch, McQuarrie and Nolet (2008) jointly estimate all 92,175 parameters together with these "source corrections." Using those LSQR solutions, we have two modeling alternatives for the earth structural inversion with N travel delay time observations $\mathbf{y} \in \mathbb{R}^N$:

$$\textit{Model } 1: \quad \mathbf{y}_{\text{usa}} = X_{\text{usa}}\boldsymbol{\beta}_{\text{usa}} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}_N\left(\mathbf{0}, \frac{1}{\phi} I_N\right), \tag{3}$$

where $X_{\text{usa}} \in \mathbb{R}^{N\times 8977}$ denotes the ensemble of sensitivity kernels of the western USA. $\mathcal{N}_N(\boldsymbol{\mu}, \Sigma)$ denotes the N-dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance Σ, and the N-dimensional unity matrix is denoted by I_N. In model 1, we only estimate the velocity parameters $\boldsymbol{\beta}_{\text{usa}}$ and keep the part of the travel delay time for the corrections parameters (path AB in right panel of Figure 2) fixed at the LSQR solutions of $\boldsymbol{\beta}_{\text{hyp}}$ and $\boldsymbol{\beta}_{\text{time}}$ estimated by Sigloch, McQuarrie and Nolet (2008). The extended model with joint estimation of source corrections is given by

$$\begin{aligned} \textit{Model } 2: \quad & \mathbf{y}_{\text{cr}} = X_{\text{usa}}\boldsymbol{\beta}_{\text{usa}} + X_{\text{hyp}}\boldsymbol{\beta}_{\text{hyp}} + X_{\text{time}}\boldsymbol{\beta}_{\text{time}} + \boldsymbol{\varepsilon}, \\ & \boldsymbol{\varepsilon} \sim \mathcal{N}_N\left(\mathbf{0}, \frac{1}{\phi} I_N\right). \end{aligned} \tag{4}$$

Here we apply the travel delay time $\mathbf{y}_{\mathrm{cr}}$ assuming that the part of the travel time running through path AC is given. This given part of the travel times is again based on the LSQR solution estimated by Sigloch, McQuarrie and Nolet (2008).

The number of travel time data from source-receiver pairs is $N = 53{,}270$, collected from 760 stations and 529 events. The number of hypocenter correction parameters is 1587 (529 earthquakes $\times$ 3) and there are 529 time correction parameters. Sigloch (2008) found that in the uppermost mantle, between 0 km to 100 km depth, the velocity can deviate by more than $\pm 5\%$ from the spherically symmetric reference model. As depth increases, the mantle becomes more homogeneous and the velocity deviates less from the reference model.

3. Estimation method.

3.1. *Modeling the spatial structure of the velocity parameters.* In both models (3) and (4) we have the spatial parameter $\boldsymbol{\beta}_{\mathrm{usa}}$, which we denote generically as $\boldsymbol{\beta}$ in this section. In the Bayesian approach we need a proper prior distribution for this high-dimensional parameter vector $\boldsymbol{\beta}$. To account for their spatially correlated structure, we apply the conditional autoregressive model (CAR) and assume a Markov random field structure for $\boldsymbol{\beta}$. This assumption says that the conditional distribution of the local characteristics β_i, given all other parameters β_j, $j \neq i$, only depends on the neighbors, that is, $P(\beta_i \mid \boldsymbol{\beta}_{-i}) = P(\beta_i \mid \beta_j, j \sim i)$, where $\boldsymbol{\beta}_{-i} := (\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_d)'$ and "$\sim i$" denotes the set of neighbors of site i. The CAR model and its application have been investigated in many studies, such as Pettitt, Weir and Hart (2002) or Rue and Held (2005). Since the earth is heterogeneous and layered, lateral correlation length scales are larger than over depths, and so we propose an ellipsoidal neighborhood structure for the velocity parameters. Let $(x_j, y_j, z_j)' \in \mathbb{R}^3$ be the positions of the ith and the jth nodes in Cartesian coordinates. The jth node is a neighbor of node i if the ellipsoid equation is satisfied, that is, $(\frac{x_i - x_j}{D_x})^2 + (\frac{y_i - y_j}{D_y})^2 + (\frac{z_i - z_j}{D_z})^2 \leq 1$. To add a rotation of the ellipsoid to an arbitrary direction in the space, we could simply modify the vector $(x_i - x_j, y_i - y_j, z_i - z_j)'$ to $R(x_i - x_j, y_i - y_j, z_i - z_j)'$ with a rotation matrix $R = R_x R_y R_z$ for given rotation matrices in the x, y and z directions, respectively. The spherical neighborhood structure is a special case of the ellipsoidal structure with $D_x = D_y = D_z$. Let D be the maximum distance of D_x, D_y and D_z. For weighting the neighbors we adopt either the exponential $w_e(\cdot)$ or reciprocal weight functions $w_r(\cdot)$, that is,

$$w_e(d_{ij}) := \exp\left\{-\frac{3d_{ij}^2}{D^2}\right\} \quad \text{and} \quad w_r(d_{ij}) := \frac{D}{d_{ij}} - 1, \tag{5}$$

where d_{ij} is the Euclidean distance between node i and node j. The exponential weight function is bounded while the reciprocal weight function is unbounded. Those weighting functions have been studied by Pettitt, Weir and Hart (2002) or

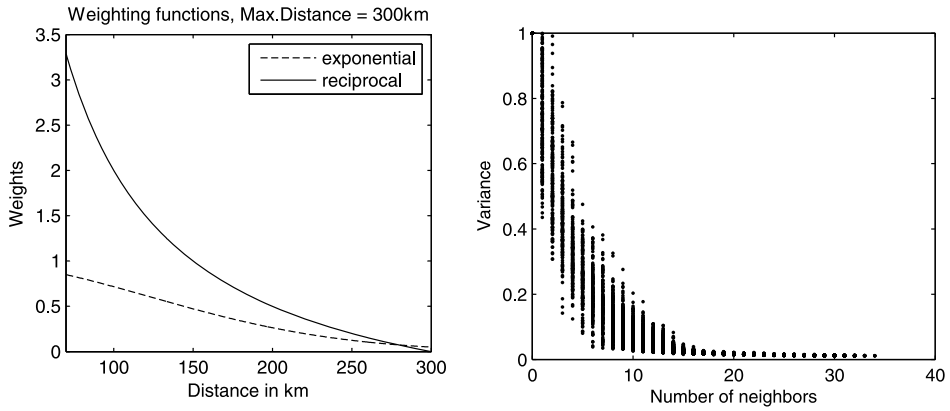

FIG. 3. *Left*: *exponential and reciprocal weight functions for the spatial prior, for* $D = 150$ *km and* $D = 300$ *km*. *Right*: *the trade-off relationship between numbers of neighbors and the prior variance* $\operatorname{diag}(Q^{-1}(\psi))$, $\psi = 10$, $D = 150$ *km*, $w =$ *reciprocal weights*.

Congdon (2003). The left panel of Figure 3 illustrates the weight functions for $D = 300$ km.

Let $\omega(d_{ij})$ be either $w_e(\cdot)$ or $w_r(\cdot)$ in (5). To model the spatial structure of $\boldsymbol{\beta}_{\text{usa}}$ in (3) and (4), a CAR model is used. Following Pettitt, Weir and Hart (2002), let $\boldsymbol{\beta}_{\text{usa}} \sim \mathcal{N}_{d_{\text{usa}}}(\mathbf{0}, \frac{1}{\eta_{\text{usa}}} Q^{-1}(\psi))$ with precision matrix

$$Q_{ij}(\psi) := \begin{cases} 1 + |\psi| \sum_{i:j\sim i} \omega(d_{ij}), & i = j, \\ -\psi\omega(d_{ij}), & i \neq j, i \sim j \text{ for } \psi \in \mathbb{R}. \end{cases} \tag{6}$$

They showed that Q is symmetric and positive definite, and that conditional correlations can be explicitly determined. For $\psi \to 0$, the precision matrix Q converges to the identity matrix, that is, $\psi = 0$ corresponds to independent elements of $\boldsymbol{\beta}$. The precision matrix in (6) for both elliptical and spherical cases indicates anisotropic covariance structure and depends on the distance between nodes, the number of neighbors of each node and the weighting functions. The elliptical precision matrix additionally depends on the orientation. The right panel of Figure 3 shows the trade-off between numbers of neighbors and prior variance, which indicates that the more neighbors the ith node has, the smaller is its prior variance $(Q^{-1}(\psi))_{ii}$. Posterior distribution of velocity parameters from regions with less neighborhood information can be rough, since they are not highly regularized due to the large prior covariances. This may produce sharp edges in the tomographic image. However, this is a more realistic modeling method since one is more sure about the optimization solution if a velocity parameter has more neighbors. Moreover, this prior specification is adapted to the construction of the tetrahedral mesh: regions with many nodes have better ray coverage than regions with less nodes. In

summary, the prior incorporates diverse spatial knowledge about the velocity parameters. Since a precision matrix is defined, which is sparse and positive definite, it provides a computational advantage in sampling from a high-dimensional Gaussian distribution as required in our algorithm (shown in the following sections).

3.2. *A Gibbs–Metropolis sampler for parameter estimation in high dimensions*. To quantify uncertainty, we adopt a Bayesian approach. Posterior inference for the model parameters is facilitated by a Metropolis within Gibbs sampler [Brooks et al. (2011)]]. Recall the linear model in (4),

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}_N\left(\mathbf{0}, \frac{1}{\phi} I_N\right),$$

where $\boldsymbol{\beta} := (\boldsymbol{\beta}_{\text{usa}}, \boldsymbol{\beta}_{\text{hyp}}, \boldsymbol{\beta}_{\text{time}})'$ and $X := (X_{\text{usa}}, X_{\text{hyp}}, X_{\text{time}})$. We now specify the prior distribution of $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta} \sim \mathcal{N}_d(\boldsymbol{\beta}_0, \Sigma_\beta) \qquad \text{with } \boldsymbol{\beta}_0 := (\boldsymbol{\beta}_{0,\text{usa}}, \boldsymbol{\beta}_{0,\text{hyp}}, \boldsymbol{\beta}_{0,\text{time}})'.$$

The prior covariance matrix Σ_β is chosen as

$$\Sigma_\beta := \begin{pmatrix} \frac{1}{\eta_{\text{usa}}} Q^{-1}(\psi) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{1}{\eta_{\text{hyp}}} I_{d_{\text{hyp}}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1}{\eta_{\text{time}}} I_{d_{\text{time}}} \end{pmatrix}. \tag{7}$$

Since we are interested in modeling positive spatial dependence, we impose that the spatial dependence parameter ψ is the truncated normal distribution a priori, that is, $\psi \sim \mathcal{N}(\mu_\psi, \sigma_\psi^2)\mathbb{1}(\psi > 0)$. The priors for the precision scale parameters η_{usa}, η_{hyp}, η_{time} and ϕ are specified in terms of a Gamma distribution $\Gamma(a, b)$ with density $g(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp\{-bx\}$, $x > 0$. The corresponding first two moments are $\frac{a}{b}$ and $\frac{a}{b^2}$, respectively.

The MCMC procedure is derived as follows: the full conditionals of $\boldsymbol{\beta}$ are

$$\begin{aligned} \boldsymbol{\beta} \mid \mathbf{y}, \psi, \boldsymbol{\eta} &\sim \mathcal{N}_d(\Omega_\beta^{-1} \boldsymbol{\xi}_\beta, \Omega_\beta^{-1}), \\ &\text{with } \Omega_\beta := \Sigma_\beta^{-1} + \phi X'X, \boldsymbol{\xi}_\beta := \Sigma_\beta^{-1} \boldsymbol{\beta}_0 + \phi X'\mathbf{y} \end{aligned} \tag{8}$$

and $\boldsymbol{\eta} := (\eta_{\text{usa}}, \eta_{\text{hyp}}, \eta_{\text{time}})$. For η_{usa}, η_{hyp}, η_{time} and ϕ, the full conditionals are again Gamma distributed. The estimation of ψ requires a Metropolis–Hastings (MH) step. The logarithm of the full conditional of ψ is proportional to

$$\begin{aligned} \log \pi(\psi \mid \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\eta}) \propto{}& \frac{1}{2} \log\left|Q(\psi)\right| \\ &- \frac{\eta_{\text{usa}}}{2} (\boldsymbol{\beta}_{\text{usa}} - \boldsymbol{\beta}_{0,\text{usa}})' Q(\psi) (\boldsymbol{\beta}_{\text{usa}} - \boldsymbol{\beta}_{0,\text{usa}}) \\ &- \frac{(\psi - \mu_\psi)^2}{2\sigma_\psi^2}. \end{aligned}$$

For the MH step, we choose a truncated normal random walk proposal for ψ to obtain a new sample, that is, $\mathcal{N}(\psi^{\text{old}}, \bar{\sigma}_{\psi}^2)\mathbb{1}(\psi > 0)$. We use a Cholesky decomposition with permutation to obtain a sample of $\boldsymbol{\beta}$ in (8) (Section 3.4). The method by Pettitt, Weir and Hart (2002), solving a sparse matrix equation, is not useful. Here, computing the determinant of the Cholesky factor of $Q(\psi)$ is much more efficient than calculating its eigenvalues, due to the size and sparseness of $Q(\psi)$.

3.3. *Relationship to ridge regression.* To show the relationship between our approach and ridge regression (also called Tikhonov regularization), we consider only model 1. For simplicity we neglect the notation "usa" in (3). The analysis is also applicable to model 2.

Let $\hat{\boldsymbol{\beta}}^{\text{ridge}}(\lambda) := (X'X + \lambda I_d)^{-1}X'\mathbf{y}$ be the corresponding ordinary ridge regression (ORR) estimate with shrinkage parameter λ [Dębski (2010), Hoerl and Kennard (1970), Swindel (1976)]. For given hyperparameters η, ϕ and ψ, the full conditional of $\boldsymbol{\beta}$ is $\boldsymbol{\beta} \mid \eta, \phi, \psi \sim \mathcal{N}_d(\Omega_\beta^{-1}\boldsymbol{\xi}_\beta, \Omega_\beta^{-1})$ with $\Omega_\beta := \eta Q(\psi) + \phi X'X$ and $\boldsymbol{\xi}_\beta := \eta Q(\psi)\boldsymbol{\beta}_0 + \phi X'\mathbf{y}$. The corresponding full conditional mean can therefore be expressed as

$$E[\boldsymbol{\beta} \mid \mathbf{y}, \psi, \eta] = \left(X'X + \frac{\eta}{\phi}Q(\psi)\right)^{-1}\left(X'y + \frac{\eta}{\phi}Q(\psi)\boldsymbol{\beta}_0\right).$$

This is close to the modified ridge regression estimator $\hat{\boldsymbol{\beta}}^{\text{ridge}}(\lambda, \boldsymbol{\beta}_0) := (X'X + \lambda I_d)^{-1}(X'\mathbf{y} + \lambda\boldsymbol{\beta}_0)$ defined in Swindel (1976). We can see that if $\psi \to 0$, then $\frac{\eta}{\phi}Q(\psi) \to \frac{\eta}{\phi}$, which is the equivalent to λ in the modified ridge regression. This shows that the prior precision matrix $\eta Q(\psi)$ is a regularization matrix with parameter ψ controlling the prior covariance. As discussed in Section 3.1, the prior covariance $\frac{1}{\eta}Q^{-1}(\psi)$ also varies with the specified weights in (5) with maximum distance D and with number of neighboring nodes. For large ψ or large weights function values, as well as large number of neighbors, the prior variances are small, which well reflects the prior knowledge about the data coverage and parameter uncertainty. Thus, the full conditional mean is close to the prior mean in this case.

3.4. *Computational issues.* Since the size of the travel time data requires high-dimensional parameters to be estimated, the traditional method of sampling the parameter vector $\boldsymbol{\beta}$ from $\mathcal{N}_d(\Omega_\beta^{-1}\boldsymbol{\xi}_\beta, \Omega_\beta^{-1})$ directly, as defined in (8), is not efficient with respect to computing time. We instead use a Cholesky decomposition of Ω_β. Since the sensitivity kernel X is sparse, and the prior covariance matrix is sparse and positive definite, the matrix Ω_β remains sparse and symmetric positive definite. Therefore, we can reduce the cost of the Cholesky decompositions. For this we apply an approximate minimum degree ordering algorithm (AMD algorithm) to find a permutation P of Ω_β so that the number of nonzeros in its Cholesky factor is reduced [Amestoy, Davis and Duff (1996)]. In our case, the number of

nonzeros of the full conditional precision matrix Ω_β in (8) is about 5% of all elements. After this permutation the nonzeros of the Cholesky factor are reduced by 50% compared to the original number of nonzeros.

To sample a multivariate normal distributed vector after permutation, we follow Rue and Held (2005). Given the permutation matrix P of Ω_β, we sample a vector $\mathbf{v} := P\boldsymbol{\beta}$, where $\mathbf{v} = (L_p')^{-1}((L_p^{-1})P\boldsymbol{\xi}_\beta + \mathbf{Z})$ with L_p a lower triangular matrix resulting from the Cholesky decomposition of $P\Omega_\beta$, and $\mathbf{Z}$ a standard normal distributed vector, that is, $\mathbf{Z} \sim \mathcal{N}_d(\mathbf{0}, I_d)$. The original parameter vector of interest $\boldsymbol{\beta}$ can be obtained after permuting vector $\mathbf{v}$ again. Rue and Held (2005) suggested finding a permutation such that the matrix is banded. However, we found that in our case the AMD algorithm is more efficient with regard to computing time. Using MATLAB built-in functions, the Cholesky decomposition with an approximate minimum degree ordering takes 8 seconds on a Linux-Cluster 8-way Opteron with 32 cores, while the Cholesky decomposition based on a banded matrix takes 15 seconds. The traditional method without permutation requires 118.5 seconds.

4. Simulation study.

4.1. *Simulation setups*. In this section we examine the performance of our approach for model 1. We want to investigate whether the method works correctly under the correct model assumptions and how much influence the prior has on the posterior estimation. We consider five different prior neighborhood structures of $\boldsymbol{\beta}_{\text{usa}}$:

(0) Independent model of $\boldsymbol{\beta}_{\text{usa}}$, $\psi = 0$ fixed, that is, $\boldsymbol{\beta}_{\text{usa}} \sim \mathcal{N}_{d_{\text{usa}}}(\boldsymbol{\beta}_0, \frac{1}{\eta_{\text{usa}}} I_{d_{\text{usa}}})$,
(1) Spherical neighborhood structure with reciprocal weight function,
(2) Ellipsoidal neighborhood structure with reciprocal weight function,
(3) Spherical neighborhood structure with exponential weight function,
(4) Ellipsoidal neighborhood structure with exponential weight function.

Note that the independent model of $\boldsymbol{\beta}_{\text{usa}}$ corresponds to the Bayesian ridge estimator as described in Section 3.3. For the weight functions in (5), we set $D_x = D_y = 300$ km and $D_z = 150$ km for modeling ellipsoidal neighborhood structures, and $D = 150$ km for the spherical neighborhood distance.

Setup I: Assume the solution by Sigloch, McQuarrie and Nolet (2008), denoted as $\hat{\boldsymbol{\beta}}_{\text{usa}}^{\text{LSQR}}$, represents true mantle structure beneath North America. We use the forward model $X\hat{\boldsymbol{\beta}}_{\text{usa}}^{\text{LSQR}}$ to compute noise-free, synthetic data. Then, we generate two types of noisy data, that is, $\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{usa}}^{\text{LSQR}} + \boldsymbol{\varepsilon}$ with:

(A) Gaussian noise ($\boldsymbol{\varepsilon} \sim \mathcal{N}_N(\mathbf{0}, \frac{1}{\phi_{\text{tr}}} I_N)$, $\phi_{\text{tr}} = 0.4$),
(B) t-noise ($\boldsymbol{\varepsilon} \sim t_N(\mathbf{0}, I_N, \nu_{\text{tr}})$, $\nu_{\text{tr}} = 3$, corresponds to $\phi_{\text{tr}} = 0.333$).

Although we add t-noise to our synthetic earth model $\hat{\boldsymbol{\beta}}_{\text{usa}}^{\text{LSQR}}$, our posterior calculation is based on Gaussian errors. Additionally, we compare two priors for

$\boldsymbol{\beta}_{\text{usa}} \sim \mathcal{N}_{d_{\text{usa}}}(\boldsymbol{\beta}_0, \frac{1}{\eta_{\text{usa}}} Q^{-1}(\psi))$ to examine the sensitivity of the posterior estimates to the prior choices:

(a) $\boldsymbol{\beta}_0 \sim \mathcal{N}_{d_{\text{usa}}}(\hat{\boldsymbol{\beta}}_{\text{usa}}^{\text{LSQR}}, 0.32^2 I_d)$,
(b) $\boldsymbol{\beta}_0 = \mathbf{0}$ (spherically symmetric reference model).

The priors for the hyperparameters are set as follows: $\psi \sim \mathcal{N}(10, 0.2^2)$, $\phi \sim \Gamma(1, 0.1)$ resulting in expectation and standard deviation of 10, $\eta_{\text{usa}} \sim \Gamma(10, 2)$ resulting in expectation of 5 and standard deviation of 1.6.

Setup II: In this case we examine the performance under known prior neighborhood structures. We construct a synthetic true mantle model with two types of known prior neighborhood structures: $\boldsymbol{\beta}_{\text{usa,tr}} \sim \mathcal{N}_{d_{\text{usa}}}(\hat{\boldsymbol{\beta}}_{\text{usa}}^{\text{LSQR}}, \frac{1}{\eta_{\text{usa,tr}}} Q^{-1}(\psi_{\text{tr}}))$ with $\eta_{\text{usa,tr}} = 0.18$ and $\psi_{\text{tr}} = 10$ using:

(a) a spherical neighborhood structure for $\boldsymbol{\beta}_{\text{usa,tr}}$ with reciprocal weights,
(b) an ellipsoidal neighborhood structure for $\boldsymbol{\beta}_{\text{usa,tr}}$ with reciprocal weights.

Again, Gaussian noise is added to the forward model, that is, $\mathbf{Y} = X\hat{\boldsymbol{\beta}}_{\text{usa}}^{\text{LSQR}} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathcal{N}_N(0, \frac{1}{\phi_{\text{tr}}} I_N)$, $\phi_{\text{tr}} = 0.4$. Posterior estimation is carried out assuming the five different prior structures.

The number of MCMC iterations for scenarios in setups I and II is 3000, thinning is 15, and burn-in after thinning is 100. For convergence diagnostics we compute the trace, autocorrelation and estimated density plots as well as the effective sample size (ESS) using `coda` package in R for those samples. According to Brooks et al. (2011), the ESS is defined by $\text{ESS} := \frac{n}{1+2\sum_{k=1}^{\infty} \rho_k}$, with the original sample size n and autocorrelation $\rho_k < 0.05$ at lag k. The infinite sum can be truncated at lag k when ρ_k becomes smaller than 0.05 [Kass et al. (1998), Liu (2008)].

4.2. *Performance evaluation measures*. To evaluate the results, we use the standardized Euclidean norm for both data and model misfits, $\|\cdot\|_{\Sigma_y}$ and $\|\cdot\|_{\Sigma_\beta}$, respectively. The function $\|\mathbf{x}\|_\Sigma$ of a vector $\mathbf{x}$ of mean $\boldsymbol{\mu}$ and covariance Σ is called the *Mahalanobis distance*, defined by $\|\mathbf{x}\|_\Sigma := \sqrt{(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$. To include model complexity, we calculate the deviance information criterion (DIC) [Spiegelhalter et al. (2002)]. Let $\boldsymbol{\theta}$ denote the parameter vector to be estimated. Furthermore, the likelihood of the model is denoted by $\ell(\mathbf{y} \mid \bar{\boldsymbol{\theta}})$, where $\bar{\boldsymbol{\theta}}$ is the estimated posterior mean of $\boldsymbol{\theta}$, estimated by $\frac{1}{R}\sum_{r=1}^{R} \boldsymbol{\theta}^r$ with R number of independent MCMC samples. According to Spiegelhalter et al. (2002) and Celeux et al. (2006), the deviance is defined as $D(\boldsymbol{\theta}) = -2\log(\ell(\mathbf{y} \mid \bar{\boldsymbol{\theta}})) + 2\log h(\mathbf{y})$. The term $h(\mathbf{y})$ is a standardizing term which is a function of the data alone and does not need to be known. Thus, for model comparison we take $D(\boldsymbol{\theta}) = -2\log(\ell(\mathbf{y} \mid \bar{\boldsymbol{\theta}}))$. The effective number of parameters in the model, denoted by p_D, is defined by

$p_D := E_\theta[D(\boldsymbol{\theta})] - D(\bar{\boldsymbol{\theta}})$. The term $E_\theta[D(\boldsymbol{\theta})]$ is the posterior mean deviance and is estimated by $\frac{1}{R}\sum_{r=1}^{R} D(\boldsymbol{\theta}^r)$. This term can be regarded as a Bayesian measure of fit. In summary, the DIC is defined as $\text{DIC} = E_\theta[D(\boldsymbol{\theta})] + p_D = D(\bar{\boldsymbol{\theta}}) + 2p_D$. The model with the smallest DIC is the preferred model under the trade-off of model fit and model complexity.

4.3. *Results and interpretations*. The first two blocks in Table 1 illustrate posterior estimation results for setup I. It shows that the estimation method with ellipsoidal prior structures (2) and (4) turn out to be the most adequate, according to the DIC criterion. The standardized data misfit criteria $\|\cdot\|_{\Sigma_y}$ given the estimated posterior mode $\hat{\boldsymbol{\beta}}$ show similar results in all scenarios. However, this measure ignores the uncertainty of $\boldsymbol{\beta}_{\text{usa}}$. The criteria $\|\mathbf{y} - X\hat{\boldsymbol{\beta}}_L\|_{\Sigma_y}$ and $\|\mathbf{y} - X\hat{\boldsymbol{\beta}}_U\|_{\Sigma_y}$ show the data misfit given the 90% credible interval with lower and upper quantile posterior estimates $\hat{\boldsymbol{\beta}}_L$ and $\hat{\boldsymbol{\beta}}_U$, respectively. These estimates give a range of the data misfit for all possible posterior solutions of $\boldsymbol{\beta}_{\text{usa}}$ and show that methods with independent prior generally yield larger ranges of misfit values than the ones with spatial structures. This indicates that the credible intervals of methods with spatial priors can fit the data better. Further, methods with spatial priors in setup I(b) show smaller model misfit under $\|\cdot\|_{\Sigma_\beta}$ than ones with independent prior, while in setup I(a) results with independent priors are better. Generally, estimated posterior modes of η_{usa} vary considerably due to the different prior assumptions. Models with ellipsoidal neighborhood structures have a stronger prior (in the sense of a smaller prior variance) than models with spherical neighborhood structure. Similarly, models with reciprocal weights have a stronger regularization toward the prior mean than models with exponential weights. This means that the posterior estimates of η_{usa} adapt to different prior settings. Moreover, we notice that the estimate of the spatial dependence parameter ψ depends strongly on its prior, as the prior mean is close to the posterior estimates of ψ in all scenarios. The last two blocks in Table 1 illustrate results from setup II assuming known spatial structure including hyperparameters. The DIC values indicate that our approach correctly detects the underlying prior structures [in (a) it is prior structure (1), in (b) it is prior structure (2)]. We can also observe that our approach estimates the hyperparameters correctly. Estimated posterior modes of the parameters from the identified model are close to their true values.

Generally, tomographic images illustrate velocity parameters as deviation of the solution from the spherically symmetric reference model (in %). Blue colors represent zones that have faster seismic velocities than the reference earth model, while red colors denote slower velocities. Physically, blue colors usually imply that those regions are colder than the default expectation for the corresponding mantle depth, while red regions are hotter than expected. In our simulation study, we assumed the true earth to be represented by the solution of Sigloch, McQuarrie and Nolet (2008), shown in the left column of Figure 4. The middle and right columns of

TABLE 1

Posterior estimation results of the simulation study under setups I *and* II*, using synthetic earth models. The posterior mode of the velocity parameters is denoted as* $\hat{\boldsymbol{\beta}}$*. The quantities* $\hat{\boldsymbol{\beta}}_L$ *and* $\hat{\boldsymbol{\beta}}_U$ *are lower and upper quantiles of the* 90% *credible interval of the MCMC estimates, respectively*

Setup I							
Noises	**Prior struct**	$\|\mathbf{y} - X\hat{\boldsymbol{\beta}}\|_{\boldsymbol{\Sigma}_y}$	$\|\mathbf{y} - X\hat{\boldsymbol{\beta}}_L\|_{\boldsymbol{\Sigma}_y}$	$\|\mathbf{y} - X\hat{\boldsymbol{\beta}}_U\|_{\boldsymbol{\Sigma}_y}$	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathbf{tr}}\|_{\boldsymbol{\Sigma}_\beta}$	**DIC**	**Mode** $\hat{\eta}_{\mathbf{usa}}$
(a) $\boldsymbol{\beta} \sim \mathcal{N}_d(\boldsymbol{\beta}_0, \frac{1}{\eta_{\text{usa}}} Q^{-1}(\psi))$, $\boldsymbol{\beta}_0 \sim \mathcal{N}_d(\hat{\boldsymbol{\beta}}_{\text{usa}}^{\text{LSQR}}, 0.32^2 I_d)$							
(A) Gaussian noise	(0)	232.28	312.82	308.27	91.30	103,748	9.28
$\varepsilon_i \sim N(0, 1/\phi_{\text{tr}})$	(1)	231.71	258.32	255.74	349.70	103,467	2.89
$\phi_{\text{tr}} = 0.4$, η_{usa},	(2)	231.67	264.81	261.59	200.34	103,442	0.11
ψ unknown	(3)	231.77	263.89	260.96	256.52	103,478	2.70
	(4)	231.70	267.81	264.44	185.04	103,456	0.20
(B) t-noises	(0)	227.74	602.35	604.21	46.83	112,436	0.57
$\varepsilon_i \sim t(0, 1, \nu_{\text{tr}})$	(1)	228.58	443.00	443.02	69.50	112,226	0.09
$\phi_{\text{tr}} = 0.33$,	(2)	228.57	437.58	437.80	57.83	112,118	0.01
$\nu_{\text{tr}} = 3$, η_{usa},	(3)	228.51	450.67	450.27	62.03	112,175	0.11
ψ unknown	(4)	228.52	445.22	445.42	55.87	112,126	0.01
(b) $\boldsymbol{\beta} \sim \mathcal{N}_d(\mathbf{0}, \frac{1}{\eta_{\text{usa}}} Q^{-1}(\psi))$							
(A) Gaussian noise	(0)	234.33	635.99	632.19	50.53	106,563	0.59
$\varepsilon_i \sim N(0, 1/\phi_{\text{tr}})$	(1)	233.46	458.55	454.42	40.53	105,365	0.09
$\phi_{\text{tr}} = 0.4$, η_{usa},	(2)	233.36	449.35	444.06	42.45	105,200	0.01
ψ unknown	(3)	233.52	466.36	462.04	38.55	105,357	0.12
	(4)	233.40	458.05	452.60	42.71	105,256	0.01
(B) t-noises	(0)	226.53	831.85	832.84	40.10	113,023	0.19
$\varepsilon_i \sim t(0, 1/\phi_{\text{tr}}, \nu_{\text{tr}})$	(1)	227.60	599.53	598.99	33.50	112,575	0.03
$\phi_{\text{tr}} = 0.33$,	(2)	227.56	596.04	595.78	33.62	112,512	0.00
$\nu_{\text{tr}} = 3$, η_{usa},	(3)	227.61	606.60	605.77	33.48	112,541	0.03
ψ unknown	(4)	227.55	607.03	606.69	34.03	112,536	0.00

TABLE 1
(*Continued*)

Setup II							
Noises	**Prior struct**	$\|\mathbf{y} - X\hat{\boldsymbol{\beta}}\|_{\boldsymbol{\Sigma}_y}$	$\|\mathbf{y} - X\hat{\boldsymbol{\beta}}_L\|_{\boldsymbol{\Sigma}_y}$	$\|\mathbf{y} - X\hat{\boldsymbol{\beta}}_U\|_{\boldsymbol{\Sigma}_y}$	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathbf{tr}}\|_{\boldsymbol{\Sigma}_\beta}$	**DIC**	**Mode $\hat{\eta}_{\mathbf{usa}}$**
(a) $\boldsymbol{\beta} \sim \mathcal{N}_d(\boldsymbol{\beta}_0, \frac{1}{\eta_{\text{usa}}} Q^{-1}(\psi))$ with a spherical neighborhood structure for Q							
Gaussian noise	(0)	279.38	575.55	520.63	40.06	129,433	1.09
$\varepsilon_i \sim N(0, 1/\phi_{\text{tr}})$	(1)	279.82	439.32	389.22	35.83	128,882	0.20
$\phi_{\text{tr}} = 0.4$,	(2)	279.78	475.57	422.49	36.95	129,034	0.01
$\eta_{\text{usa}} = 0.18$,	(3)	279.96	455.81	404.30	36.23	128,986	0.22
$\psi = 10$	(4)	279.79	481.30	427.90	37.10	129,066	0.02
(b) $\boldsymbol{\beta} \sim \mathcal{N}_d(\boldsymbol{\beta}_0, \frac{1}{\eta_{\text{usa}}} Q^{-1}(\psi))$ with an ellipsoidal neighborhood structure for Q							
Gaussian noise	(0)	234.71	305.10	292.47	27.71	104,662	11.84
$\varepsilon_i \sim N(0, 1/\phi_{\text{tr}})$	(1)	234.37	257.34	249.46	26.10	104,262	4.19
$\phi_{\text{tr}} = 0.4$,	(2)	234.27	251.55	244.20	24.59	104,152	0.30
$\eta_{\text{usa}} = 0.18$,	(3)	234.41	260.03	251.59	25.72	104,284	4.22
$\psi = 10$	(4)	234.30	253.50	245.76	24.50	104,173	0.53

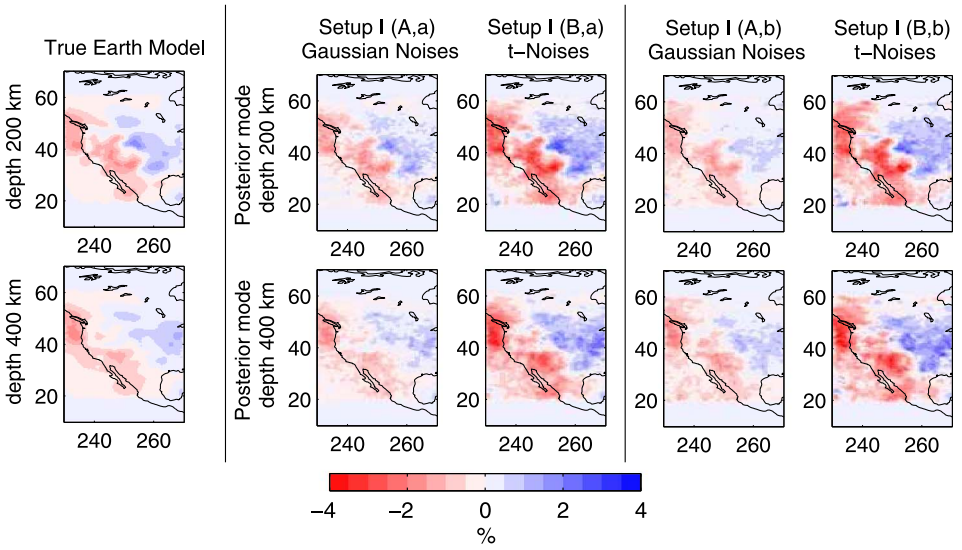

FIG. 4. *Mantle models resulting from the simulation study. Left column shows the "true" model, used to generate the synthetic data. The unit on the color bar is velocity deviation* $\boldsymbol{\beta}$ *in % from the spherically symmetric reference model. All other columns show the posterior mode of velocity deviation* $\boldsymbol{\beta}$*, estimated using ellipsoidal prior structure with reciprocal weights. Middle columns show results for setup* I(a)*, which uses the prior mean* $\boldsymbol{\beta} \neq \mathbf{0}$*. Right columns show results for setup* I(b) *assuming prior mean* $\boldsymbol{\beta}_0 = \mathbf{0}$.

Figure 4 illustrate the estimated posterior modes $\hat{\boldsymbol{\beta}}_{\text{usa}}$ from setup I with ellipsoidal neighborhood structure and reciprocal weight for both Gaussian and t-noises, respectively. They show that the parameter estimates from Gaussian noises are close to the true solution, while the solution from the t-noises tends to overestimate the parameters. The magnitude of mantle anomalies is overestimated but major structures are correctly recovered. The same effect can be seen in the last column of Figure 4 which displays the estimated posterior modes of the tomographic solutions in setup I(b). We have overestimation since the noise is not adequate to the Gaussian model assumption. Moreover, we also observe that tomographic solutions with the prior mean $\boldsymbol{\beta}_0 \neq \mathbf{0}$ are smoother than the ones with the prior mean $\boldsymbol{\beta}_0 = \mathbf{0}$.

Figure 5 shows estimated credible intervals for the solutions of Figure 4. Credible intervals for solutions with t-noises are larger than those for the Gaussian noises, as indicated by the darker shades of blue/red colors, which denote higher/lower quantile estimates. This implies that parameter uncertainty is greater if noise does not fit the model assumption. The same effect can be seen for results with the prior mean $\boldsymbol{\beta}_0 = \mathbf{0}$. The bottom row of Figure 5 maps out how the regions differ from the reference model with 90% posterior probability. For model-conform Gaussian distributed noises and informative prior mean, more regions differ from the reference model with 90% posterior probability than if we added

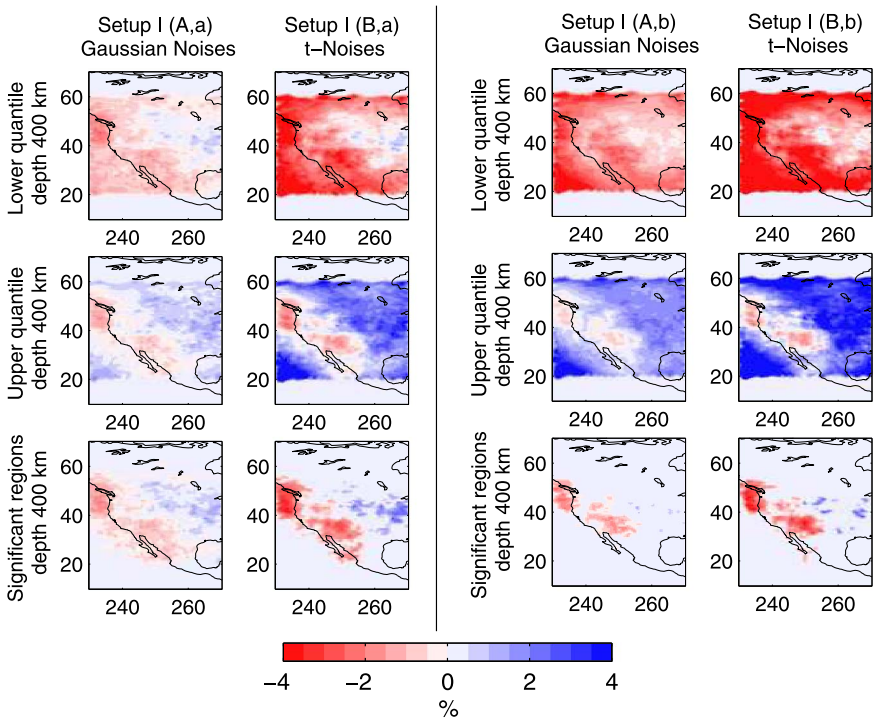

FIG. 5. *Continuation of Figure 4. The maps show velocity deviation in % from the reference Earth model. Left half shows the results under setup* I(a), *which uses the prior mean* $\boldsymbol{\beta}_0 \neq \mathbf{0}$; *right half describes setup* I(b), *which uses prior mean* $\boldsymbol{\beta}_0 = \mathbf{0}$. *First and second rows map out the lower and upper quantiles of the* 90% *confidence interval. Third row shows the posterior mode of velocity structure* $\boldsymbol{\beta}$, *but rendered only in regions that differ significantly from the reference model, according to the* 90% *confidence interval.*

t-noise or used the less informative prior. In the case of an informative prior and/or correctly modeled noise, we achieve more certainty about the velocity deviations from the reference earth model.

5. Application to real seismic travel time data. In this section we apply our MCMC approach to actually measured travel time data.

The measurements are a subset of those generated by Sigloch, McQuarrie and Nolet (2008). We use the same wave paths, but only measurements made on the broadband waveforms, whereas they further bandpassed the data for finite-frequency measurements and also included amplitude data [Sigloch and Nolet (2006)]. Most stations are located in the western U.S., as part of the largest-ever seismological experiment (USArray), which is still in the process of rolling across the continent from west to east. Numerous tomographic studies have incorporated USArray data—the ones most similar to ours are Burdick et al. (2008), Sigloch, McQuarrie and Nolet (2008), Tian, Sigloch and Nolet (2009), and Schmandt

and Humphreys (2010). All prior studies obtained their solutions through least-squares minimization, which yields no uncertainty estimates. Here we use 53,270 broadband travel time observations to estimate velocity structure under western North America (over 11,000 parameters), plus source corrections for 529 events (2116 parameters). We conduct our Bayesian inversion following two different scenarios:

Model 1: We only invert for earth structural parameters. For the velocity parameters we assume $\boldsymbol{\beta} \sim \mathcal{N}_{d_{\text{usa}}}(\hat{\boldsymbol{\beta}}^{\text{LSQR}}_{\text{usa}}, \frac{1}{\eta_{\text{usa}}} Q^{-1}(\psi))$ as in (3) with $\psi \sim \mathcal{N}(10, 0.5^2)\mathbb{1}(\psi > 0)$, $\phi \sim \Gamma(1, 0.1)$ and $\eta_{\text{usa}} \sim \Gamma(10, 2)$.

Model 2: We invert for both earth structural parameters and the source corrections. The prior distributions are set to $\boldsymbol{\beta} \sim \mathcal{N}_d(\hat{\boldsymbol{\beta}}^{\text{LSQR}}, \Sigma_\beta)$ as in (4) and Σ_β as defined in (7). For ψ, ϕ and η_{usa}, we adopt the same distribution as in model 1. For the parameters of the source corrections we adopt $\eta_{\text{hyp}} \sim \Gamma(1, 5)$ and $\eta_{\text{time}} \sim \Gamma(10, 2)$.

We use the same five prior structures (0)–(4) as in the simulation study and run the MCMC algorithm for 10,000 iterations. The high-dimensional $\boldsymbol{\beta}$ vector can be sampled efficiently in terms of ESS with low burn-in and thinning rates thanks to the efficient Gibbs sampling scheme in (8). However, the hyperparameters, for example, ψ_β, are more difficult to sample. To achieve a good mixing, we applied a burn-in of 200 and a thinning rate of 25 (393 samples for each parameter) in our analysis. On average, the effective sample size ESS values for $\boldsymbol{\beta}_{\text{usa}}$, $\boldsymbol{\beta}_{\text{hyp}}$ and $\boldsymbol{\beta}_{\text{time}}$ are about 393, 393 and 327, respectively, which indicate very low autocorrelations for most of the parameters. The ESS of both η_{usa} and ψ_β is about 103, while both η_{hyp} and ϕ have good mixing characteristics with ESS values equal to the sample size, and η_{time} has ESS value equal to 165. Figure 6 shows as examples the parameters $\beta_{\text{usa},955}$ at node 955, η_{usa} and ψ_β. The computing cost of our algorithm is about $O(n^4)$. Sampling model 1 with about 9000 parameters, our algorithm needs 12 hours in 10,000 runs on a 32-core cluster, while under the same condition it needs 38 hours for model 2.

Table 2 shows the results from model 1 (estimation of earth structure) and model 2 (earth structure plus source corrections). For both models, results from the independent prior structures, corresponding to the Bayesian ridge estimator, provide the best fit according to the DIC criterion. We also run the model 1 with prior mean $\boldsymbol{\beta}_0 = \mathbf{0}$ (the spherically symmetric reference model) and different covariance structures (0)–(2). The DIC results for priors (0), (1) and (2) are 103,100, 103,700 and 103,370, respectively. Two reasons may explain the selection of prior (0): (1) the data has generally more correlation structure than the i.i.d. Gaussian assumption, which can not be solely explained by the spatial prior structure of the $\boldsymbol{\beta}$-fields. However, in our simulation study where different prior structures and the corrected data error are applied (Table 1), the DIC was able to identify the correct models; (2) Since the data are noisy, fitting could be difficult without a shrinkage prior. The prior in (0) can be compared to shrinkage in the ridge regression, which is the limiting case of priors in (1) to (4). Priors in (1) to (4) do not shrink the

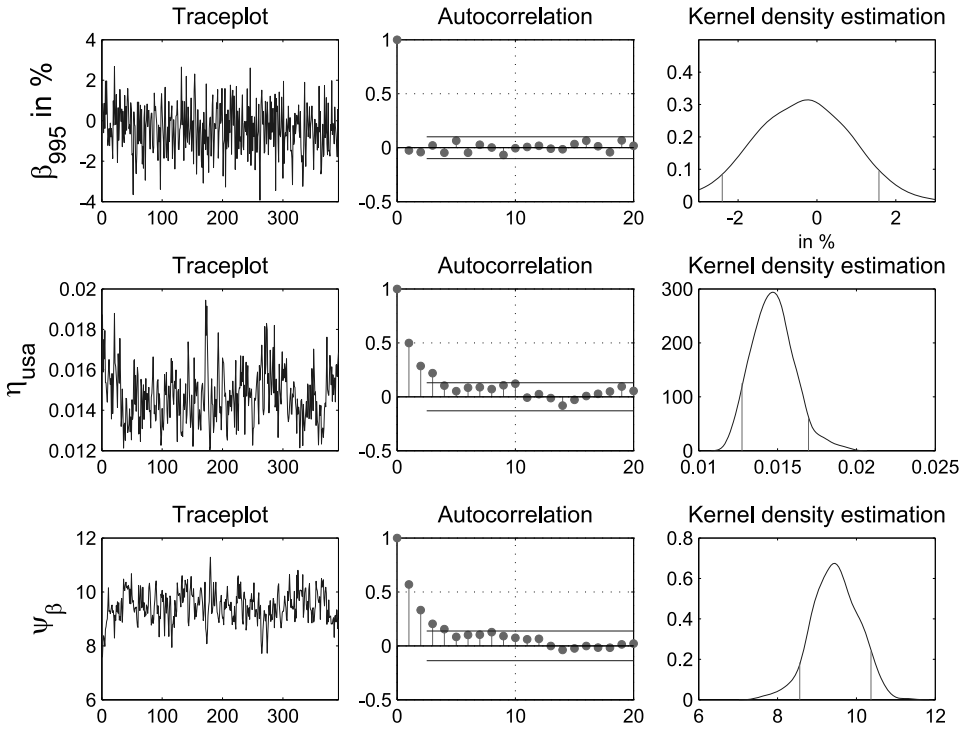

FIG. 6. *Convergence diagnostics*: *trace plot*, *autocorrelation and kernel density estimation of the parameters* $\boldsymbol{\beta}_{\text{usa}}$ *at node* 955, η_{usa} *and* ψ_β. *For* 10,000 *MCMC iterations the samples shown in plots are based on a burn-in of* 200 *and a thinning rate of* 25.

solutions of $\boldsymbol{\beta}$-fields as much as prior (0). They better reflect the uncertainty since the prior covariances in (1)–(4) are larger than variances in prior (0) in regions that have no data (no neighboring nodes), and smaller in regions with lots of data (lots of neighboring nodes).

Furthermore, the standardized data misfit criteria $\|\cdot\|_{\Sigma_y}$ do not show much difference between models with different prior specifications. According to the estimated 90% credible interval, estimates using spherical prior structure show a smaller range of data misfit in model 1, whereas in model 2, the independence prior shows a better result. Since our method assumes i.i.d. Gaussian errors, the resulting residuals might not be optimally fitted as expected. With regard to computational time, the independent prior model has a definite advantage over other priors in both models 1 and 2. The general advantage of our Bayesian method is that the independent model yields an estimate given as the ratio between the variance of the data and the variance of the priors corresponding to ridge estimates with *automatically chosen shrinkage* described in Section 3.3, whereas in Aster, Borchers and Thurber (2005), Sigloch, McQuarrie and Nolet (2008), Bodin et al.

TABLE 2
Posterior estimation results for the inversion using real data, under models 1 *and* 2 *specifications*

Model 1

Prior struct	$\|\mathbf{y} - X\hat{\boldsymbol{\beta}}\|_{\boldsymbol{\Sigma}_y}$	$\|\mathbf{y} - X\hat{\boldsymbol{\beta}}_L\|_{\boldsymbol{\Sigma}_y}$	$\|\mathbf{y} - X\hat{\boldsymbol{\beta}}_U\|_{\boldsymbol{\Sigma}_y}$	DIC	Mode $\hat{\boldsymbol{\phi}}$	Mode $\hat{\eta}_{\text{usa}}$	Mode $\hat{\boldsymbol{\psi}}$
(0)	228.46	490.92	490.46	102,928	0.40	1.40	–
(1)	229.14	389.73	390.21	104,096	0.39	0.20	9.63
(2)	228.72	464.73	465.67	103,466	0.40	0.01	9.98
(3)	228.90	430.78	431.34	103,749	0.39	0.17	9.63
(4)	228.74	471.50	472.37	103,408	0.40	0.01	9.98

Model 2

Prior struct	$\|\mathbf{y} - X\hat{\boldsymbol{\beta}}\|_{\boldsymbol{\Sigma}_y}$	$\|\mathbf{y} - X\hat{\boldsymbol{\beta}}_L\|_{\boldsymbol{\Sigma}_y}$	$\|\mathbf{y} - X\hat{\boldsymbol{\beta}}_U\|_{\boldsymbol{\Sigma}_y}$	DIC	Mode $\hat{\boldsymbol{\phi}}$	Mode $\hat{\eta}_{\text{usa}}$	Mode $\hat{\boldsymbol{\psi}}$	Mode $\hat{\eta}_{\text{hyp}}$	Mode $\hat{\eta}_{\text{time}}$
(0)	225.40	483.96	488.35	93,788	0.49	1.15	–	0.01	5.01
(1)	225.76	515.29	524.61	94,993	0.48	0.10	9.63	0.01	4.53
(2)	225.48	498.96	501.45	94,374	0.48	0.00	9.55	0.01	4.70
(3)	225.61	503.20	512.01	94,669	0.48	0.11	9.63	0.01	4.53
(4)	225.44	496.69	498.97	94,312	0.49	0.01	10.00	0.01	4.70

(2012a) and all other prior work, the shrinkage parameter (strength of regularization) had to be chosen by the user a priori.

Figure 7 shows the estimated posterior and prior densities of parameters in model 2, at four different locations of varying depth. We see that parameters at locations with good ray coverage, for example, node 5400 and node 3188, have smaller credible intervals than parameters at locations with no ray coverage, for example, node 5564 and node 995 beneath the uninstrumented oceans. Geologically, the regions between node 5400 and node 3188 are well known to represent the hot upper mantle, where seismic waves travel slower than the reference velocity. This is consistent with our results in Figure 7: the fact that $\boldsymbol{\beta} = \mathbf{0}$ does not fall inside the 90% credible intervals indicates a velocity deviation from the spherically symmetric reference model with high posterior probability. Figure 7 shows that the posterior is more diffuse than the prior. As mentioned in Section 3.1, the spatial prior for $\boldsymbol{\beta}$ depends on distance of neighboring nodes, number of neighbors and orientation. The variance can be very small if the number of neighbors is very large, as shown in Figure 3. Incorporating data, the information about $\boldsymbol{\beta}$ is updated and thus may yield more diffuse posteriors than the priors, as we see here. The left half of Figure 8 shows the estimated posterior modes of mantle structure obtained by model 2, for independent and for ellipsoidal priors with reciprocal weights. The right half of Figure 8 extracts only those regions that differ from the

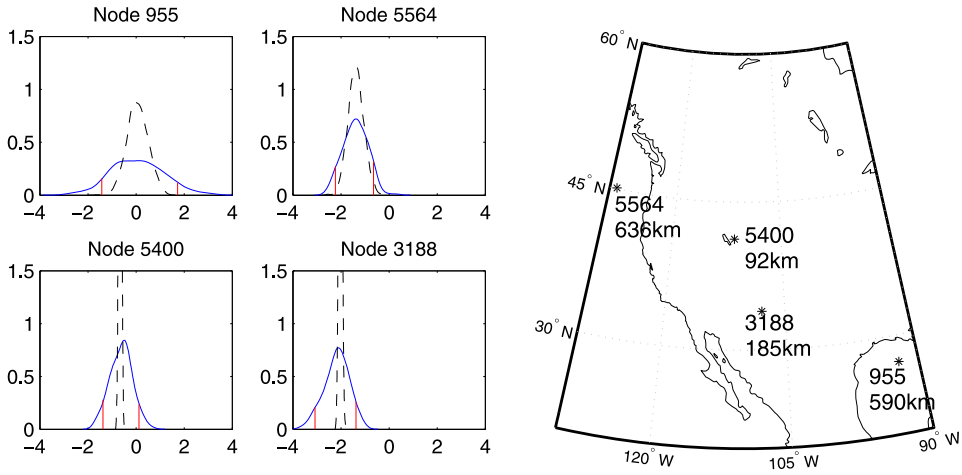

FIG. 7. *Results of the Bayesian tomography using real travel time observations. Left*: *estimated posterior density of* $\boldsymbol{\beta}_{\text{usa}}$ *at a few selected model nodes, whose locations and depths are indicated on the map. Unit on the x-axes is velocity deviation in %. Dashed lines*: *prior density, the prior variance can be very small if number of neighbors is large. Solid lines*: *posterior density with* 90% *credible intervals.*

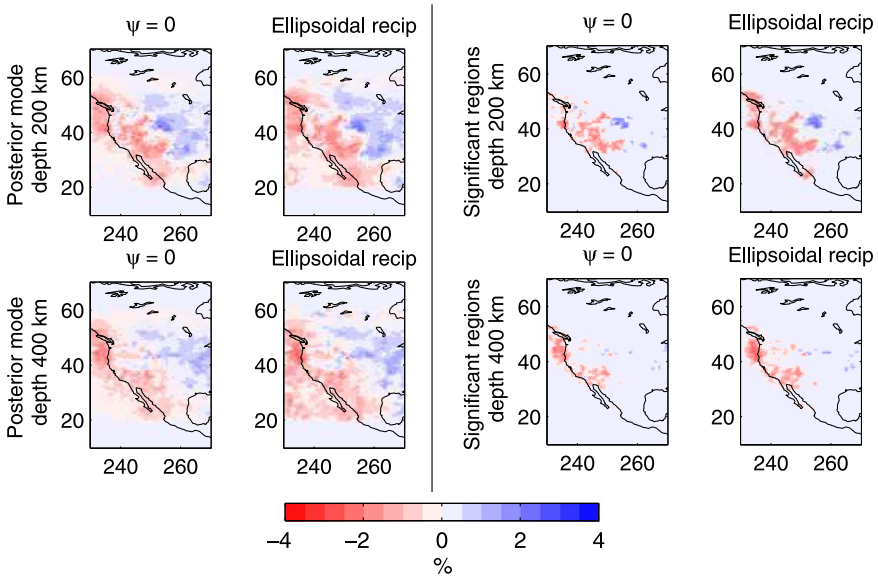

FIG. 8. *Results of the Bayesian tomography using real travel time data. All maps show estimated velocity deviation from the reference earth model IASP*91 (*in %*). *Left columns*: *estimated posterior mode of velocity deviation, for the scenario of model* 2. *Right columns*: *same posterior mode, but rendered are only regions that differ from the reference model with* 90% *posterior probability.*

reference model according to the 90% credible interval. The ellipsoidal prior results in higher certainty of velocity deviations at a depth of 200 km compared to the independence prior. At a depth of 400 km, the credible regions resemble each other more strongly. This confirms geological arguments that deeper regions of the mantle are more homogeneous and do not differ as much from the spherically symmetric reference model as shallower regions.

Many lines of geoscientific investigations provide independent confirmation of the significantly anomalous regions of Figure 8. The red areas map out the hot upper mantle under the volcanic, extensional Basin and Range province and Yellowstone; the blue anomalies map out the western edge of the old and cool North American craton.

The overall comparison of our solutions to earlier least-squares inversions, for example, the model by Sigloch (2008) shown in the left column of Figure 4, confirms that Bayesian inversion successfully retrieves the major features of mantle structure. The images are similar, but the major advantage and novelty of our approach is that it also quantifies uncertainties in the solution (which we have chosen to visualize as credible intervals here).

6. Discussion and outlook. Uncertainty quantification in underdetermined, large inverse problems is important, since a single solution is not sufficient for making conclusive judgements. Two central difficulties for MCMC methods have always been the dimensionality of the problem (number of parameters to sample) or the evaluation of the complex physical forward model (nonlinear problems) in each MCMC iteration [Bui-Thanh, Ghattas and Higdon (2011), Martin et al. (2012), Tarantola (2004)].

Consider the model $\mathbf{Y} = f(\boldsymbol{\beta}) + \boldsymbol{\varepsilon}$ with the physical forward model $f(\cdot)$, high-dimensional parameter $\boldsymbol{\beta}$ and error $\boldsymbol{\varepsilon}$. In general, if the physical problem is linear ($f(\boldsymbol{\beta}) = X\boldsymbol{\beta}$) and the full conditional of $\boldsymbol{\beta}$ is Gaussian, efficient sampling from the high-dimensional Gaussian conditional distribution is essential for the exploration of model space. In this case the error $\boldsymbol{\varepsilon}$ need not necessarily be Gaussian, but may be t or skewed-t distributed [Frühwirth-Schnatter and Pyne (2010), Sahu, Dey and Branco (2003)], or a Gaussian error with a spatial correlation such as considered in Banerjee, Gelfand and Carlin (2003). Given a sparse posterior precision matrix [e.g., (8)], efficient sampling from a multivariate normal can be carried out by Cholesky decomposition of a permuted precision matrix as discussed in Wilkinson and Yeung (2002) or Rue and Held (2005), by using an approximate minimum-degree ordering algorithm. A further improvement to the current sampling approach might be to apply the Krylov subspace method from Simpson, Turner and Pettitt (2008). This would require substantial implementation efforts and is the subject of further research. If the forward matrix or the prior precision matrix is not sparse, a dense posterior precision matrix for $\boldsymbol{\beta}$ will result. In this case our sampling scheme is inefficient, but the model-space reduction method developed by Flath et al. (2011) might be used instead. They exploit the low-rank structure of the preconditioned Hessian matrix of the data misfit, involving eigenvalue

calculations. However, this approximation quantifies uncertainty of large-scale linear inverse problems only for known hyperparameters, thus ignoring uncertainty in those parameters. Eigenvalue calculation in each MCMC step can be time consuming and prohibitive for hierarchical models with unknown hyperparameters when the posterior covariance matrix in every MCMC step changes. Here additional research is needed.

If the full conditionals cannot be written as Gaussian [this case includes the cases of a nonlinear $f(\cdot)$, a non-Gaussian prior of $\boldsymbol{\beta}$ or non-Gaussian, nonelliptical distributed errors], using the standard MH algorithm to sample from the high-dimensional posterior distribution is often computationally infeasible. Constructing proposal density that provides a good approximation of the stationary distribution while keeping the high-dimensional forward model $f(\cdot)$ inexpensive to evaluate has been the focus of the research over the past years: Lieberman, Willcox and Ghattas (2010) have drawn samples from an approximate posterior density on a reduced parameter space using a projection-based reduced-order model. In the adaptive rejection sampling technique by Cui, Fox and O'Sullivan (2011), the exact posterior density is evaluated only if its approximation is accepted. The stochastic Newton approach proposed by Martin et al. (2012) approximates the posterior density by local Hessian information, thus resulting in an improvement of the Langevin MCMC by Stramer and Tweedie (1999). Other random-walk-free, optimization-based MCMC techniques for improving the proposal and reducing correlation between parameters have been developed, such as Hamiltonian Monte Carlo (HMC) [Neal (2010)], Adaptive Monte Carlo (AM) [Andrieu and Thoms (2008), Haario, Saksman and Tamminen (2001)] and several variations, for example, delay rejection AM (DRAM) [Haario et al. (2006)], differential evolution MC (DEMC) [Ter Braak (2006)] and differential evolution adaptive Metropolis (DREAM) [Vrugt et al. (2009)], just to mention a few. However, MCMC sampling of high-dimensional problems still requires a massive amount of computing time and resources. For example, the quasi three-dimensional nonlinear model of Herbei, McKeague and Speer (2008) contains about 9000 parameters on a 37×19 grid. We expect a long computing time since they use standard MCMC sampling methods. The example by Cui, Fox and O'Sullivan (2011) shows that their algorithm achieves a significant improvement in both computing time and efficiency of parameter space sampling for a large nonlinear system of PDEs that includes about 10,000 parameters. However, their algorithm gives 11,200 iterations in about 40 days, while our problem requires only 38 hours (on a 32-core cluster) for the same number of iterations for about 11,000 parameters.

While the future may be in effective uncertainty quantification of nonlinear physical problems using model reduction and optimization techniques, the computing time and resources at the moment are too demanding to explore the large model space. This paper demonstrates effective Bayesian analysis tailored to a realistically large seismic tomographic problem, featuring over 11,000 structural and source parameters. We deliver a precise uncertainty quantification of tomographic

models in terms of posterior distribution and credible intervals using the MCMC samples, which allows us to detect regions that differ from the reference earth model with high posterior probability. Our approach is the first to solve seismic tomographic problems in such high dimensions on a fine grid, and thus provides ground work in this important research area.

Acknowledgments. The authors acknowledge two referees, the Associate Editor and the Editor for helpful remarks and suggestions which led to a significant improved manuscript. The authors would like to thank the support of the Leibniz-Rechenzentrum in Garching, Germany.

REFERENCES

AKI, K. and LEE, W. (1976). Determination of three-dimensional velocity anomalies under a seismic array using first P arrival times from local earthquakes: 1. A homogeneous initial model. *Journal of Geophysical Research* **81** 4381–4399.

AMESTOY, P. R., DAVIS, T. A. and DUFF, I. S. (1996). An approximate minimum degree ordering algorithm. *SIAM J. Matrix Anal. Appl.* **17** 886–905. MR1410707

ANDRIEU, C. and THOMS, J. (2008). A tutorial on adaptive MCMC. *Stat. Comput.* **18** 343–373. MR2461882

ASTER, R., BORCHERS, B. and THURBER, C. (2005). *Parameter Estimation and Inverse Problems* (*International Geophysics*), Har/cdr ed. Academic Press, San Diego.

BANERJEE, S., GELFAND, A. E. and CARLIN, B. P. (2003). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton, FL.

BARDSLEY, J. M. (2012). MCMC-based image reconstruction with uncertainty quantification. *SIAM J. Sci. Comput.* **34** A1316–A1332. MR2970254

BODIN, T. and SAMBRIDGE, M. (2009). Seismic tomography with the reversible jump algorithm. *Geophysical Journal International* **178** 1411–1436.

BODIN, T., SAMBRIDGE, M., TKALČIĆ, H., ARROUCAU, P., GALLAGHER, K. and RAWLINSON, N. (2012a). Transdimensional inversion of receiver functions and surface wave dispersion. *Journal of Geophysical Research* **117** B02301.

BODIN, T., SAMBRIDGE, M., RAWLINSON, N. and ARROUCAU, P. (2012b). Transdimensional tomography with unknown data noise. *Geophysical Journal International* **189** 1536–1556.

BROOKS, S., GELMAN, A., JONES, G. L. and MENG, X.-L., eds. (2011). *Handbook of Markov Chain Monte Carlo*. CRC Press, Boca Raton, FL. MR2742422

BUI-THANH, T., GHATTAS, O. and HIGDON, D. (2011). Adaptive Hessian-based non-stationary Gaussian process response surface method for probability density approximation with application to Bayesian solution of large-scale inverse problems. Technical report, Institut for computational engineering and sciences, Univ. Texas, Austin.

BURDICK, S., LI, C., MARTYNOV, V., COX, T., EAKINS, J., MULDER, T., ASTIZ, L., VERNON, F. L., PAVLIS, G. L. and VAN DER HILST, R. D. (2008). Upper mantle heterogeneity beneath North America from travel time tomography with global and USArray transportable array data. *Seismological Research Letters* **79** 389–392.

CELEUX, G., FORBES, F., ROBERT, C. P. and TITTERINGTON, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Anal.* **1** 651–673 (electronic). MR2282197

CONGDON, P. (2003). *Applied Bayesian Modelling*. Wiley, Chichester. MR1990543

CROSSON, R. S. (1976). Crustal structure modeling of earthquake data 1. Simultaneous least squares estimation of hypocenter and velocity parameters. *Journal of Geophysical Research* **81** 3036–3046.

CUI, T., FOX, C. and O'SULLIVAN, M. J. (2011). Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance Metropolis–Hastings algorithm. *Water Resources Research* **47** 26 pp.

DAHLEN, F. A., HUNG, S. H. and NOLET, G. (2000). Fréhet kernels for finite-frequency traveltimes-I. Theory. *Geophysical Journal International* **141** 157–174.

DĘBSKI, W. (2010). Seismic tomography by Monte Carlo sampling. *Pure and Applied Geophysics* **167** 131–152.

DZIEWONSKI, A. M. and ANDERSON, D. L. (1981). Preliminary reference Earth model. *Physics of the Earth and Planetary Interiors* **25** 297–356.

FICHTNER, A., KENNETT, B. L. N., IGEL, H. and BUNGE, H. P. (2009). Full waveform tomography for upper-mantle structure in the Australasian region using adjoint methods. *Geophysical Journal International* **179** 1703–1725.

FLATH, H. P., WILCOX, L. C., AKÇELIK, V., HILL, J., VAN BLOEMEN WAANDERS, B. and GHATTAS, O. (2011). Fast algorithms for Bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial Hessian approximations. *SIAM J. Sci. Comput.* **33** 407–432. MR2783201

FRÜHWIRTH-SCHNATTER, S. and PYNE, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics* **11** 317–336.

HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (2001). An adaptive metropolis algorithm. *Bernoulli* **7** 223–242. MR1828504

HAARIO, H., LAINE, M., MIRA, A. and SAKSMAN, E. (2006). DRAM: Efficient adaptive MCMC. *Stat. Comput.* **16** 339–354. MR2297535

HÄMÄLÄINEN, M. S. and ILMONIEMI, R. J. (1994). Interpreting magnetic fields of the brain: Minimum norm estimates. *Medical and Biolgical Engineering and Computing* **32** 35–42.

HERBEI, R., MCKEAGUE, I. W. and SPEER, K. G. (2008). Gyres and jets: Inversion of tracer data for ocean circulation structure. *Journal of Physical Oceanography* **38** 1180–1202.

HOERL, E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

KAIPIO, J. and SOMERSALO, E. (2007). Statistical inverse problems: Discretization, model reduction and inverse crimes. *J. Comput. Appl. Math.* **198** 493–504. MR2260683

KASS, R. E., CARLIN, B. P., GELMAN, A. and NEAL, R. M. (1998). Markov chain Monte Carlo in practice: A roundtable discussion. *Amer. Statist.* **52** 93–100. MR1628427

KENNETT, B. L. N. and ENGDAHL, E. R. (1991). Traveltimes for global earthquake location and phase identification. *Geophysical Journal International* **105** 429–465.

KHAN, A., ZUNINO, A. and DESCHAMPS, F. (2011). The thermo-chemical and physical structure beneath the North American continent from Bayesian inversion of surface-wave phase velocitie. *Journal of Geophysical Research* **116** 23 pp.

KOLEHMAINEN, V., VANNE, A., SILTANEN, S., JÄRVENPÄÄ, S., KAIPIO, J. P., LASSAS, M. and KALKE, M. (2007). Bayesian inversion method for 3D dental X-ray imaging. *e & i Elektrotechnik und Informationstechnik* **124** 248–253.

LIEBERMAN, C., WILLCOX, K. and GHATTAS, O. (2010). Parameter and state model reduction for large-scale statistical inverse problems. *SIAM J. Sci. Comput.* **32** 2523–2542. MR2684726

LIU, J. S. (2008). *Monte Carlo Strategies in Scientific Computing*. Springer, New York. MR2401592

MARTIN, J., WILCOX, L. C., BURSTEDDE, C. and GHATTAS, O. (2012). A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM J. Sci. Comput.* **34** A1460–A1487. MR2970260

MERCERAT, D. and NOLET, G. (2013). On the linearity of cross-correlation delay times in finite-frequency tomography. *Geophysical Journal International* **192** 681–687.

MOHAMMAD-DJAFARI, A. (2012). Bayesian approach with prior models which enforce sparsity in signal and image processing. *EURASIP Journal on Advances in Signal Processing* **2012** 52.

MONTELLI, R., NOLET, G., DAHLEN, F. A., MASTERS, G., ENGDAHL, E. R. and HUNG, S. H. (2004). Finite-frequency tomography reveals a variety of plumes in the mantle. *Science* **303** 338–343.

MOSCA, I., COBDEN, L., DEUSS, A., RITSEMA, J. and TRAMPERT, J. (2012). Seismic and mineralogical structures of the lower mantle from probabilistic tomography. *Journal of Geophysical Research* **117** 26 pp.

MOSEGAARD, K. and TARANTOLA, A. (1995). Monte Carlo sampling of solutions to inverse problems. *Journal of Geophysical Research* **100** 12431–12447.

MOSEGAARD, K. and TARANTOLA, A. (2002). *International Handbook of Earthquake and Engineering Seismology*: *Probabilistic Approach to Inverse Problems* 237–265. Academic Press, San Diego.

NEAL, R. M. (2010). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo* (S. Brooks, A. Gelman, G. Jones and X. L. Meng, eds.). Chapman & Hall/CRC, Boca Raton, FL.

NOLET, G. (2008). *A Breviary of Seismic Tomography*: *Imaging the Interior of the Earth and Sun*. Cambridge Univ. Press, Cambridge.

PAIGE, C. C. and SAUNDERS, M. A. (1982). LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Trans. Math. Software* **8** 43–71. MR0661121

PETTITT, A. N., WEIR, I. S. and HART, A. G. (2002). A conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. *Stat. Comput.* **12** 353–367. MR1951708

RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields*: *Theory and Applications*. *Monographs on Statistics and Applied Probability* **104**. Chapman & Hall/CRC, Boca Raton, FL. MR2130347

SAHU, S. K., DEY, D. K. and BRANCO, M. D. (2003). A new class of multivariate skew distributions with applications to Bayesian regression models. *Canad. J. Statist.* **31** 129–150. MR2016224

SAMBRIDGE, M. and MOSEGAARD, K. (2002). Monte Carlo methods in geophysical inverse problems. *Reviews of Geophysics* **40** 1–29.

SCHMANDT, B. and HUMPHREYS, E. (2010). Complex subduction and small-scale convection revealed by body-wave tomography of the western United States upper mantle. *Earth and Planetary Science Letters* **297** 435–445.

SIEMINSKI, A., LIU, Q., TRAMPERT, J. and TROMP, J. (2007). Finite-frequency sensitivity of body waves to anisotropy based upon adjoint methods. *Geophysical Journal International* **171** 368–389.

SIGLOCH, K. (2008). Multiple-frequency body-wave tomography. Ph.D. thesis, Princeton Univ., Princeton, NJ.

SIGLOCH, K. (2011). Mantle provinces under North America from multi-frequency P-wave tomography. *Geochemistry*, *Geophysics*, *Geosystems* **12** Q02W08.

SIGLOCH, K., MCQUARRIE, N. and NOLET, G. (2008). Two-stage subduction history under North America inferred from finite-frequency tomography. *Nature GEO* **1** 458–462.

SIGLOCH, K. and NOLET, G. (2006). Measuring finite-frequency body-wave amplitudes and traveltimes. *Geophysical Journal International* **167** 271–287.

SIMPSON, D. P., TURNER, I. W. and PETTITT, A. N. (2008). Fast sampling from a Gaussian Markov random field using Krylov subspace approaches. QUT ePrints ID 14376, Queensland Univ. Technology, Brisbane, Australia.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 583–639. MR1979380

STRAMER, O. and TWEEDIE, R. L. (1999). Langevin-type models. II. Self-targeting candidates for MCMC algorithms. *Methodol. Comput. Appl. Probab.* **1** 307–328. MR1730652

SWINDEL, B. F. (1976). Good ridge estimators based on prior information. *Comm. Statist. Theory Methods* **A5** 1065–1075. MR0440806

TARANTOLA, A. (2004). *Inverse Problem Theory and Methods for Model Parameter Estimation*, 1st ed. SIAM, Philadelphia.

TER BRAAK, C. J. F. (2006). A Markov chain Monte Carlo version of the genetic algorithm differential evolution: Easy Bayesian computing for real parameter spaces. *Stat. Comput.* **16** 239–249. MR2242236

TIAN, Y., SIGLOCH, K. and NOLET, G. (2009). Multiple-frequency SH-wave tomography of the western US upper mantle. *Geophysical Journal International* **178** 1384–1402.

TIAN, Y., MONTELLI, R., NOLET, G. and DAHLEN, F. A. (2007). Computing traveltime and amplitude sensitivity kernels in finite-frequency tomography. *J. Comput. Phys.* **226** 2271–2288. MR2356411

TROMP, J., TAPE, C. and LIU, Q. (2005). Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels. *Geophysical Journal International* **160** 195–216.

UUTELA, K., HÄÄMÄLÄINEN, M. and SOMERSALO, E. (1999). Visualization of magnetoencephalographic data using minimum current estimates. *NeuroImage* **10** 173–180.

VRUGT, J. A., TER BRAAK, C. J. F., DIKS, C. G. H., ROBINSON, B. A., HYMAN, J. M. and HIGDON, D. (2009). Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *International Journal of Nonlinear Sciences & Numerical Simulation* **10** 273–290.

WATZENIG, D. and FOX, C. (2009). A review of statistical modelling and inference for electrical capacitance tomography. *Measurement Science and Technology* **20** 22 pp.

WILKINSON, D. J. and YEUNG, S. K. H. (2002). Conditional simulation from highly structured Gaussian systems, with application to blocking-MCMC for the Bayesian analysis of very large linear models. *Stat. Comput.* **12** 287–300. MR1933515

R. ZHANG
C. CZADO
CENTER FOR MATHEMATICAL SCIENCES
TECHNISCHE UNIVERSITÄT MÜNCHEN
BOLTZMANNSTRASSE 3
85748 GARCHING BEI MÜNCHEN
GERMANY
E-MAIL: ran.zhang@ma.tum.de
cczado@ma.tum.de

K. SIGLOCH
DEPARTMENT OF EARTH
AND ENVIRONMENTAL SCIENCES
LUDWIG-MAXIMILIANS-UNIVERSITÄT
THERESIENSTRASSE 41
80333 MUNICH
GERMANY
E-MAIL: karin.sigloch@geophysik.uni-muenchen.de