GEOPHYSICS, VOL. 86, NO. 1 (JANUARY-FEBRUARY 2021); P. R15–R30, 14 FIGS., 2 TABLES.
10.1190/GEO2019-0585.1

A gradient-based Markov chain Monte Carlo method for full-waveform inversion and uncertainty analysis

Zeyu Zhao[1] and Mrinal K. Sen[1]

ABSTRACT

Traditional full-waveform inversion (FWI) methods only render a "best-fit" model that cannot account for uncertainties of the ill-posed inverse problem. Additionally, local optimization-based FWI methods cannot always converge to a geologically meaningful solution unless the inversion starts with an accurate background model. We seek the solution for FWI in the Bayesian inference framework to address those two issues. In Bayesian inference, the model space is directly probed by sampling methods such that we obtain a reliable uncertainty appraisal, determine optimal models, and avoid entrapment in a small local region of the model space. The solution of such a statistical inverse method is completely described by the posterior distribution, which quantifies the distributions for parameters and inversion uncertainties. To efficiently sample the posterior distribution, we introduce a sampling algorithm in which the proposal distribution is constructed by the local gradient and the diagonal approximate Hessian of the local log posterior. Our algorithm is called the gradient-based Markov chain Monte Carlo (GMCMC) method. The GMCMC FWI method can quantify inversion uncertainties with estimated posterior distribution given sufficiently long Markov chains. By directly sampling the posterior distribution, we obtain a global view of the model space. Theoretically speaking, statistical assessments do not depend on starting models. Our method is applied to the 2D Marmousi model with the frequency-domain FWI setting. Numerical results suggest that our method can be readily applied to 2D cases with affordable computational efforts.

INTRODUCTION

An important application of the geophysical inverse theory is seismic full-waveform inversion (FWI), in which seismic data recorded at the surface, seafloors, or in boreholes are used to estimate subsurface compressional wave and shear wave velocities, anisotropy parameters, attenuation parameters, etc. In FWI, seismic waveforms are exploited to update subsurface model parameters by trying to match the recorded data/observed data with estimated data. Since the early development of the theory (Lailly, 1983; Tarantola, 1984), FWI has been successfully demonstrated in recovering subsurface structures and images at different scales (Pratt and Shipp, 1999; Fichtner et al., 2008; Brossier et al., 2009; Tape et al., 2010; Zhu et al., 2012; Vigh et al., 2014; Operto et al., 2015).

There are two critical issues often related to the traditional FWI problem. First, most optimization-based FWI methods are designed to only find the "best-fit" model. Nevertheless, FWI problems are often ill-posed, observed data can be noisy and incomplete, modeling methods can be inaccurate, prior knowledge can be insufficient, and model parameterization strategies can be inappropriate. All of these factors introduce uncertainties into the inversion results (Scales et al., 1992; Sen and Stoffa, 1996; Sambridge and Mosegaard, 2002; Sen and Stoffa, 2013). Multiple plausible solutions might explain the observed data equally well. In such cases, obtaining the unique "best" solution is not sufficient to fully describe the underlying problem. Instead of trying to find the actual values of model parameters that explain the data (a deterministic fashion), one might need to infer information on model parameters provided by the observed data and to estimate the uncertainties associated with the inference (a probabilistic fashion) (Tarantola, 2005). Second, in traditional local optimization-based FWI methods, an objective function, which measures the misfit between observed data and estimated data, is minimized with respect to model

Manuscript received by the Editor 13 September 2019; revised manuscript received 28 August 2020; published ahead of production 4 October 2020; published online 16 December 2020.

[1]The University of Texas at Austin, John A. and Katherine G. Jackson School of Geosciences, Institute for Geophysics, Austin, Texas 78758-4445, USA. E-mail: zeyu.zhao@utexas.edu (corresponding author); mrinal@ig.utexas.edu.

parameters (Santosa et al., 1987; Pratt et al., 1998). The commonly used L2-norm objective function can have multiple local minima because of the highly nonlinear forward mapping. Hence, the inversion is very likely to be trapped into one of the local minima (Virieux and Operto, 2009). Alternative objective functions and multiscale inversion strategies have been proposed to mitigate the issue with varying degrees of success (Luo and Schuster, 1991; Bunks et al., 1995; Ravaut et al., 2004; Sirgue and Pratt, 2004; Engquist and Froese, 2013; Fichtner et al., 2013; Wu et al., 2014; Métivier et al., 2016; Warner and Guasch, 2016; Xue et al., 2016; Zhu and Fomel, 2016; Zhao and Sen, 2019). As argued by Fichtner and Trampert (2011) and Ray et al. (2017), however, the convergence to geologically meaningful models is not always guaranteed for local optimization-based methods if the starting point of the inversion is obtained from very uninformative prior information.

Variants of Bayesian inference methods have been adopted in geophysical inverse problems to quantify uncertainties for the inversion (Duijndam, 1988; Sen and Stoffa, 1996; Ulrych et al., 2001; Petra et al., 2014; Menke, 2018). By combining prior information with observed data and modeling errors, the result of the inference is described by the posterior distribution, which accounts for inversion uncertainties. Most of the current implementations of Bayesian inference for FWI are limited to linearizing the forward mapping. The approximate posterior probability density (PPD) is then estimated by fitting a normal distribution around the maximum a posteriori model, which is obtained by local optimization methods (Gouveia and Scales, 1998; Bui-Thanh et al., 2013; Fang et al., 2014; Zhu et al., 2016; Fang et al., 2018). PPDs obtained by this strategy would be affected by the starting point of the inversion, and they may not represent the complete possible solutions (Sen and Stoffa, 1996; Fichtner et al., 2018).

Alternatively, one can take the Bayesian inference approach to frame FWI to be a statistical inverse problem using Markov chain Monte Carlo (MCMC) sampling methods, in which the posterior distribution is estimated by directly probing the model space, usually according to certain proposal distributions (Mosegaard and Tarantola, 1995; Mosegaard and Sambridge, 2002; Mosegaard and Tarantola, 2002; Sen and Stoffa, 2013). Early studies of Monte Carlo (MC) methods to geophysical inverse problems are pioneered by Keilis-Borok and Yanovskaja (1967) and Press (1968). Recent applications of MCMC methods can be found in Bodin and Sambridge (2009), Sen and Stoffa (2013), Sajeva et al. (2016), Stuart et al. (2016), Aleardi and Mazzotti (2017), Ray et al. (2017), Sen and Biswas (2017), Ely et al. (2018), Hunziker et al. (2019), and Stuart et al. (2019). One of the difficulties of implementing MCMC methods for large-scale inverse problems is the high computational cost for sampling the posterior distribution in high dimensions. Traditional sampling methods might require solving the forward mapping more than billions of times before converging to the target posterior distribution. Efficient sampling strategies can help to address the "curse of dimensionality" issue for MCMC methods. The Hamiltonian Monte Carlo (HMC) method (Neal, 2011), using derivatives, can make large independent jumps, at the same time maintaining a high acceptance ratio when sampling the posterior distribution, which promotes its application to geophysical inverse problems (Biswas and Sen, 2017; Fichtner and Simutė, 2018).

To tackle the computational issue of MCMC methods related to FWI, we propose here a gradient-based Markov chain Monte Carlo (GMCMC) sampling method for FWI in the framework of Bayesian inference. We tailor the method proposed by Geweke and Tanizaki (1999) and Martin et al. (2012) in which the first- and second-order derivative information is used to construct a proposal distribution to sample the posterior distribution through the Metropolis-Hastings (M-H) algorithm. The method is originally discussed in the low-dimensional model space with less computationally demanding problems. The application of similar MCMC methods to large-scale inverse problems is rarely studied and reported. Here, we propose to use the local gradient and the diagonal approximate Hessian of the log posterior distribution to construct a proposal distribution. Because the local geometric information of the model space is considered, the proposal distribution is expected to be a good local approximation of the underlying posterior distribution. As a result, high posterior probability regions tend to be more frequently visited and samples drawn from the proposal distribution are more likely to be accepted. It helps the Markov chain to more efficiently sample the posterior distribution. Additionally, MCMC sampling methods provide the global view of the model space; hence, the inversion avoids the entrapment in a local region. In theory, the GMCMC method can accurately estimate the posterior distribution given sufficiently long Markov chains with arbitrary starting points.

In the following sections, we first briefly review the Bayesian inference framework. Based on the local approximation for the log posterior distribution, we introduce the GMCMC sampling method and its implementation for frequency-domain acoustic FWI. The 2D Marmousi model is used to demonstrate the feasibility of the proposed method for an inverse problem where the number of model parameters is approximately 2×10^4. We show that the proposed method achieves approximate convergence with different starting models. Geologically meaningful statistical assessments are produced to account for the inversion uncertainty.

METHODOLOGY

Bayesian inference framework

Given a model parameter set $\mathbf{m} \in \mathbb{R}^p$, the "forward problem" predicts the data $\mathbf{d}_{\mathrm{cal}} \in \mathbb{R}^n$ generated by the system $\mathbf{d}_{\mathrm{cal}} = f(\mathbf{m})$. The forward mapping operator f depicts the linear or nonlinear relationship between $\mathbf{m}$ and $\mathbf{d}_{\mathrm{cal}}$. In the "inverse problem," we are provided with actual measurement/observed data $\mathbf{d}_{\mathrm{obs}} \in \mathbb{R}^n$ to infer $\mathbf{m}$. In the Bayesian inference framework, we combine the prior knowledge of the unknown model parameter $\mathbf{m}$ with the information provided by the measurement $\mathbf{d}_{\mathrm{obs}}$ to define the posterior distribution $\pi_{\mathrm{post}}(\mathbf{m}|\mathbf{d}_{\mathrm{obs}})$,

$$\pi_{\mathrm{post}}(\mathbf{m}|\mathbf{d}_{\mathrm{obs}}) \propto \pi_{\mathrm{prior}}(\mathbf{m})\pi_{\mathrm{like}}(\mathbf{d}_{\mathrm{obs}}|\mathbf{m}), \tag{1}$$

where $\pi_{\mathrm{prior}}(\mathbf{m})$ is the prior probability density, which describes a priori information on $\mathbf{m}$, and $\pi_{\mathrm{like}}(\mathbf{d}_{\mathrm{obs}}|\mathbf{m})$ is the likelihood function, which describes the conditional probability of $\mathbf{d}_{\mathrm{obs}}$ given $\mathbf{m}$. The posterior distribution $\pi_{\mathrm{post}}(\mathbf{m}|\mathbf{d}_{\mathrm{obs}})$, often written as $\pi_{\mathrm{post}}(\mathbf{m})$, describes the conditional probability of $\mathbf{m}$ given the observed data $\mathbf{d}_{\mathrm{obs}}$. The term $\pi_{\mathrm{post}}(\mathbf{m})$ is the solution to the Bayesian inverse problem consisting of all the information inferred from our prior knowledge and observed data. Such a solution takes the prior information, noise, discretization errors, and errors in modeling theory into consideration and provides statistical assessments to quantify the inversion uncertainties. In this research, a priori information and ambient noise are accounted for; we ignore the discretization errors and theoretical errors present in the forward mapping operation.

Under the assumption that the noise and the model prior can be represented by multidimensional normal distributions, we have

$$\pi_{\text{prior}}(\mathbf{m}) \propto \exp\left(-\frac{1}{2}(\mathbf{m}-\bar{\mathbf{m}}_{\text{prior}})^T\Gamma_{\text{prior}}^{-1}(\mathbf{m}-\bar{\mathbf{m}}_{\text{prior}})\right), \tag{2}$$

$$\pi_{\text{like}}(\mathbf{d}_{\text{obs}}|\mathbf{m}) \propto \exp\left(-\frac{1}{2}(\mathbf{d}_{\text{cal}}-\mathbf{d}_{\text{obs}})^T\Gamma_{\text{noise}}^{-1}(\mathbf{d}_{\text{cal}}-\mathbf{d}_{\text{obs}})\right), \tag{3}$$

where the superscript T represents the matrix transpose, $\bar{\mathbf{m}}_{\text{prior}} \in \mathbb{R}^p$ is the prior mean model and $\Gamma_{\text{prior}} \in \mathbb{R}^{p\times p}$ and $\Gamma_{\text{noise}} \in \mathbb{R}^{n\times n}$ are covariance matrices of prior model and data noise, respectively. Then, we obtain

$$\pi_{\text{post}}(\mathbf{m}|\mathbf{d}_{\text{obs}}) \propto \exp\left(-\frac{1}{2}((\mathbf{m}-\bar{\mathbf{m}}_{\text{prior}})^T\Gamma_{\text{prior}}^{-1}(\mathbf{m}-\bar{\mathbf{m}}_{\text{prior}}) + (\mathbf{d}_{\text{cal}}-\mathbf{d}_{\text{obs}})^T\Gamma_{\text{noise}}^{-1}(\mathbf{d}_{\text{cal}}-\mathbf{d}_{\text{obs}}))\right). \tag{4}$$

Note that, for problems in which $f(\mathbf{m})$ is nonlinear, $\pi_{\text{post}}(\mathbf{m})$ is not Gaussian. Directly computing $\pi_{\text{post}}(\mathbf{m})$ according to equation 4 requires evaluating $f(\mathbf{m})$ for every possible $\mathbf{m}$ in the model space. For applications in which the numbers of model parameters are large and the forward mapping is governed by computationally expensive partial differential equations, directly computing $\pi_{\text{post}}(\mathbf{m})$ poses tremendous challenges to current computational resources. To efficiently approximate $\pi_{\text{post}}(\mathbf{m})$, sampling techniques such as MCMC methods are developed to generate samples $\mathbf{y} \in \mathbb{R}^p$ from the posterior distribution with the M-H algorithm (Metropolis et al., 1953; Hastings, 1970; Tierney, 1994; Kaipio and Somersalo, 2006). Here, a sample is drawn from a proposal distribution $q(\mathbf{m_k}, \mathbf{y})$ at the current model $\mathbf{m}_k \in \mathbb{R}^p$. The generated sample $\mathbf{y}$ is then subjected to be accepted or rejected according to the M-H criterion. In this way, a chain of samples can be obtained, and the ensembles of the samples from multiple chains are used to estimate $\pi_{\text{post}}(\mathbf{m})$.

The mismatch between $q(\mathbf{m_k}, \mathbf{y})$ and $\pi_{\text{post}}(\mathbf{m})$ has a great impact on the performance of an MCMC method (Tierney, 1994; Chib and Greenberg, 1995; Gilks et al., 1995; Roberts et al., 1997). A proposal distribution that is not a good approximation of the underlying posterior distribution might lead to poor MCMC performance, especially in high dimensions. To efficiently estimate $\pi_{\text{post}}(\mathbf{m})$, a proposal distribution that well represents the underlying $\pi_{\text{post}}(\mathbf{m})$ becomes critical. In addition, sampling from the proposal distribution should be easy to achieve so that the computational cost for drawing samples is acceptable. In the following sections, we exploit the local approximation of the log posterior distribution and derive a new proposal distribution.

Local approximation for the posterior distribution

Several nomenclatures are summarized in Table 1 to make the derivations easy to follow.

We rewrite equation 4 as

$$\pi_{\text{post}}(\mathbf{m}|\mathbf{d}_{\text{obs}}) \propto \exp(-E(\mathbf{m})), \tag{5}$$

with

$$E(\mathbf{m}) = \frac{1}{2}((\mathbf{m}-\bar{\mathbf{m}}_{\text{prior}})^T\Gamma_{\text{prior}}^{-1}(\mathbf{m}-\bar{\mathbf{m}}_{\text{prior}}) + (\mathbf{d}_{\text{cal}}-\mathbf{d}_{\text{obs}})^T\Gamma_{\text{noise}}^{-1}(\mathbf{d}_{\text{cal}}-\mathbf{d}_{\text{obs}})). \tag{6}$$

Equation 6 can be regarded as the regularized L2 misfit function for the deterministic inverse problem where matrices $\Gamma_{\text{prior}}^{-1}$ and $\Gamma_{\text{noise}}^{-1}$ represent weighting matrices for the model regularization and the data regularization, respectively. For any given $\mathbf{m}_k$, we can approximate $E(\mathbf{m})$ with the Taylor series and retaining up to the quadratic term as

$$\begin{aligned} E(\mathbf{m}) &= E(\mathbf{m}_k + \Delta\mathbf{m}) \approx \tilde{E}(\mathbf{m}) \\ &= E(\mathbf{m}_k) + \Delta\mathbf{m}^T\mathbf{g} + \frac{1}{2}\Delta\mathbf{m}^T\mathbf{H}\Delta\mathbf{m} + O(\|\Delta\mathbf{m}\|^3), \end{aligned} \tag{7}$$

where $\Delta\mathbf{m} = (\mathbf{m} - \mathbf{m}_k) \in \mathbb{R}^p$, and $\mathbf{g} \in \mathbb{R}^p$ and $\mathbf{H} \in \mathbb{R}^{p\times p}$ are the gradient and the Hessian of $E(\mathbf{m})$ at $\mathbf{m}_k$, respectively. We have

$$\mathbf{g} = \nabla_{\mathbf{m}}E(\mathbf{m}_k) = \mathbf{J}^T\Gamma_{\text{noise}}^{-1}\Delta\mathbf{d}(\mathbf{m}_k) + \Gamma_{\text{prior}}^{-1}(\mathbf{m}-\bar{\mathbf{m}}_{\text{prior}}), \tag{8}$$

and

$$\mathbf{H} = \nabla_{\mathbf{m}}^2E(\mathbf{m}_k) = \mathbf{H}_a + \mathbf{R} + \Gamma_{\text{prior}}^{-1}, \tag{9}$$

with

$$\mathbf{H}_a = \mathbf{J}^T\Gamma_{\text{noise}}^{-1}\mathbf{J}, \quad \mathbf{R} = \frac{\partial\mathbf{J}^T}{\partial\mathbf{m}^T}\Gamma_{\text{noise}}^{-1}(\Delta\mathbf{d}(\mathbf{m}_k)\ldots\Delta\mathbf{d}(\mathbf{m}_k)), \tag{10}$$

where $\mathbf{J} \in \mathbb{R}^{n\times p}$ is the Jacobian matrix representing the partial derivatives of wavefields with respect to the model parameters, that is, $J_{ij} = \partial d_{\text{cal}_i}/\partial m_j, i = (1, 2, \ldots, n); j = (1, 2, \ldots, p)$.

Table 1. Nomenclature.

Symbol	Type	Description
$\mathbf{m}$	Vector	Model parameter
$\bar{\mathbf{m}}_{\text{prior}}$	Vector	Prior mean model
$E(\mathbf{m})$	Scalar	Log posterior/objective function
$\tilde{E}(\mathbf{m})$	Scalar	Local approximation of the log posterior/objective function
$\Gamma_{\text{prior}}^{-1}$	Matrix	Covariance matrix of the prior model
$\Gamma_{\text{noise}}^{-1}$	Matrix	Covariance matrix of data noise
$\mathbf{H}_a$	Matrix	$\mathbf{J}^T\Gamma_{\text{noise}}^{-1}\mathbf{J}$
$\mathbf{R}$	Matrix	$\frac{\partial\mathbf{J}^T}{\partial\mathbf{m}^T}\Gamma_{\text{noise}}^{-1}(\Delta\mathbf{d}(\mathbf{m}_k)\ldots\Delta\mathbf{d}(\mathbf{m}_k))$
$\mathbf{H}$	Matrix	$\mathbf{H}_a + \mathbf{R} + \Gamma_{\text{prior}}^{-1}$
$\mathbf{m}^*$	Vector	$-\mathbf{H}^{-1}\mathbf{g}$, local minimizer for $\tilde{E}(\mathbf{m})$, assuming positive-definite matrix $\mathbf{H}$
$\mathbf{K}$	Matrix	Diagonal approximation of $\mathbf{H}_a$
$\tilde{\mathbf{H}}$	Matrix	$\mathbf{K} + \Gamma_{\text{prior}}^{-1}$

The data misfit vector $\Delta \mathbf{d}(\mathbf{m}_k) = (\mathbf{d}_{\text{cal}}(\mathbf{m}_k) - \mathbf{d}_{\text{obs}}) \in \mathbb{R}^n$, $\mathbf{H}_a = \mathbf{J}^T \Gamma_{\text{noise}}^{-1} \mathbf{J} \in \mathbb{R}^{p\times p}$ represents the approximate Hessian, and $\mathbf{R} = \partial \mathbf{J}^T / \partial \mathbf{m}^T \Gamma_{\text{noise}}^{-1} (\Delta \mathbf{d}(\mathbf{m}_k) \ldots \Delta \mathbf{d}(\mathbf{m}_k)) \in \mathbb{R}^{p\times p}$ represents terms related to the second-order partial derivatives (Pratt et al., 1998).

Because $\tilde{E}(\mathbf{m})$ is the local quadratic approximation of $E(\mathbf{m})$, it can be written in quadratic form as

$$\tilde{E}(\mathbf{m}) = \frac{1}{2}(\Delta \mathbf{m} - \mathbf{b})^T \mathbf{A} (\Delta \mathbf{m} - \mathbf{b}) + c, \tag{11}$$

where c represents a constant number and $\mathbf{b}$ and $\mathbf{A}$ are the vector and the matrix that need to be determined. By expanding equation 11 and matching terms with equation 7, we have

$$\tilde{E}(\mathbf{m}) = \frac{1}{2}(\Delta \mathbf{m} - \mathbf{m}^*)^T \mathbf{H} (\Delta \mathbf{m} - \mathbf{m}^*) + c, \tag{12}$$

where $\mathbf{A} = \mathbf{H}$, $\mathbf{b} = \mathbf{m}^* = -\mathbf{H}^{-1}\mathbf{g}$, and $\mathbf{m}^* \in \mathbb{R}^p$ is the stationary point where $\nabla_{\Delta \mathbf{m}} E(\mathbf{m}^*) = \mathbf{0}$. Assuming that $\mathbf{H}$ is positive-definite, $\mathbf{m}^*$ minimizes $\tilde{E}(\mathbf{m})$. Substituting $\Delta \mathbf{m}$ and $\mathbf{m}^*$ into equation 12, we have

$$\tilde{E}(\mathbf{m}) = \frac{1}{2}(\mathbf{m} - (\mathbf{m}_k - \mathbf{H}^{-1}\mathbf{g}))^T \mathbf{H} (\mathbf{m} - (\mathbf{m}_k - \mathbf{H}^{-1}\mathbf{g})) + c. \tag{13}$$

Substituting equation 13 into equation 5, we obtain the local approximation to $\pi_{\text{post}}(\mathbf{m})$ as

$$\pi_{\text{post}}(\mathbf{m}) \approx \tilde{\pi}(\mathbf{m}) \propto \exp\left(-\frac{1}{2}(\mathbf{m} - (\mathbf{m}_k - \mathbf{H}^{-1}\mathbf{g}))^T \mathbf{H} (\mathbf{m} - (\mathbf{m}_k - \mathbf{H}^{-1}\mathbf{g}))\right). \tag{14}$$

Note that $\tilde{\pi}(\mathbf{m})$ is a multidimensional normal distribution $\mathcal{N}(\mathbf{m}_k - \mathbf{H}^{-1}\mathbf{g}, \mathbf{H}^{-1})$ when $\mathbf{H}$ is positive-definite. For cases in which $\mathbf{H}$ is not positive-definite, Martin et al. (2012) propose to use the low-rank approximation of $\mathbf{H}$, in which small and negative eigenvalues of $\mathbf{H}$ are replaced with a positive threshold value. The posterior distribution at the vicinity of $\mathbf{m}_k$ is now approximated by $\tilde{\pi}(\mathbf{m})$ with the local gradient and the local Hessian information. The mean value of the proposal distribution is equivalent to the model after a Newton's update from $\mathbf{m}_k$. When the forward mapping $f(\mathbf{m})$ is linear or weakly nonlinear, $\tilde{\pi}(\mathbf{m})$ approximates the underlying $\pi_{\text{post}}(\mathbf{m})$. When $f(\mathbf{m})$ presents high nonlinearity, as in the case of FWI, $\tilde{\pi}(\mathbf{m})$ is a local approximation to $\pi_{\text{post}}(\mathbf{m})$. Qi and Minka (2002), Geweke and Tanizaki (2003), Martin et al. (2012), and Robert and Casella (2013) suggest using $\tilde{\pi}(\mathbf{m})$ as a proposal distribution to draw samples from the posterior distribution. Nevertheless, directly drawing samples according to equation 14 poses great computational challenges in high dimensions due to the high computational cost related to the Hessian computation and manipulations (inverse and square-root operation).

GMCMC sampling method

In this section, we propose a new proposal distribution that is easy to construct and is computationally efficient for drawing samples. The exact Hessian $\mathbf{H}$ (i.e., equation 9) consists of three parts. (1) The term $\mathbf{H}_a$ represents the correlations for two partial-derivative wavefields with respect to the parameters. Partial-derivative wavefields are generally uncorrelated if the two model parameters are far away from each other, and they are perfectly self-correlated (Pratt et al., 1998). Therefore, $\mathbf{H}_a$ is mostly diagonally dominant. (2) The term $\mathbf{R}$ represents the changes of partial-derivative wavefields with respect to the changes of model parameters weighted by the data misfit. As pointed out by Tarantola (2005), $\mathbf{R}$ is in general small if the data misfit is small or if changes in model parameters cause few changes in the partial-derivative wavefields (the second-order scattering effect is weak). (3) The term $\Gamma_{\text{prior}}^{-1}$ contains a priori information without taking into account the observed data.

Here, we approximate $\mathbf{H}$ with the diagonal of $\mathbf{H}_a$,

$$\mathbf{H} \approx \tilde{\mathbf{H}} = \mathbf{K} + \Gamma_{\text{prior}}^{-1}, \tag{15}$$

where $\mathbf{K} \in \mathbb{R}^{p\times p}$ is the diagonal matrix of $\mathbf{H}_a$. Covariance matrix $\Gamma_{\text{prior}}^{-1}$ is positive-definite, $\mathbf{K}$ has nonzero positive terms on diagonal terms because of the autocorrelation, and $\tilde{\mathbf{H}} \in \mathbb{R}^{p\times p}$ is also positive-definite. With the information encoded in $\mathbf{K}$ provided by the observed data, we add more knowledge into the inference, leading to reduced uncertainties compared to the prior. If the data provide little information for the parameters, we gain no additional knowledge in the model space.

With the diagonal approximate Hessian, we now define a new proposal distribution based on equation 14. Because $\tilde{\mathbf{H}}$ is only an approximation to the full Hessian, $-\tilde{\mathbf{H}}^{-1}\mathbf{g}$ does not give a full Newton update. We scale $-\tilde{\mathbf{H}}^{-1}\mathbf{g}$ with α to act as a step length along the negative gradient direction from $\mathbf{m}_k$. With the same argument, we scale $\tilde{\mathbf{H}}$ with $1/\beta^2$ to obtain the covariance matrix for the proposal distribution. Therefore, we have the new proposal distribution defined as

$$q(\mathbf{m}_k, \mathbf{y}) = \exp\left(-\frac{1}{2}(\mathbf{y} - (\mathbf{m}_k - \alpha \tilde{\mathbf{H}}^{-1}\mathbf{g}))^T \frac{\tilde{\mathbf{H}}}{\beta^2} (\mathbf{y} - (\mathbf{m}_k - \alpha \tilde{\mathbf{H}}^{-1}\mathbf{g}))\right), \tag{16}$$

which is a normal distribution $\mathcal{N}(\mathbf{m}_k - \alpha \tilde{\mathbf{H}}^{-1}\mathbf{g}, \beta^2 \tilde{\mathbf{H}}^{-1})$. Considering the decomposition of $\tilde{\mathbf{H}}^{-1} = \tilde{\mathbf{H}}^{-1/2} (\tilde{\mathbf{H}}^{-1/2})^T$, we can draw samples from the multivariate proposal distribution according to Gentle (2009),

$$\mathbf{y} = \mathbf{m}_k - \alpha \tilde{\mathbf{H}}^{-1}\mathbf{g} + \beta \tilde{\mathbf{H}}^{-1/2}\mathbf{r}, \tag{17}$$

where vector $\mathbf{r} \in \mathbb{R}^p$ is a random vector drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Equation 17 indicates that a sample is obtained by adding $-\alpha \tilde{\mathbf{H}}^{-1}\mathbf{g}$ and $\beta \tilde{\mathbf{H}}^{-1/2}\mathbf{r}$ on the current model $\mathbf{m}_k$. It can be interpreted as a model update using the preconditioned gradient (i.e., $-\alpha \tilde{\mathbf{H}}^{-1}\mathbf{g}$) plus a constrained random perturbation term (i.e., $\beta \tilde{\mathbf{H}}^{-1/2}\mathbf{r}$). The preconditioning and constraining terms come from $\tilde{\mathbf{H}}$: large $\tilde{\mathbf{H}}$ values would lead to low variances or more constraints, and small $\tilde{\mathbf{H}}$ values would lead to high variances or fewer constraints. Drawing samples with equation 17 can also be interpreted in the framework of the Langevin MC method (Grenander and Miller, 1994; Roberts and Tweedie, 1996; Stuart et al., 2004) or the one-step HMC method (Neal, 2011). In Appendix A, we give a short comparison between the proposed sampling method and the Langevin MC and HMC methods.

The values of α and β determine how far to go along the negative gradient direction and how much the model is randomly perturbed. They can be used as tunable parameters to adjust the acceptance rate. We normalize $\tilde{\mathbf{H}}^{-1}\mathbf{g}$ and $\tilde{\mathbf{H}}^{-1/2}\mathbf{r}$ to bring all values into the range $[-1, 1]$ km/s so that the values of α and β can be easily tuned. Tuning parameter α controls the step length for the preconditioned gradient. Its value should be large enough to make the gradient contribute meaningful information to the model update, but small enough so that the sampling process is not dominated by the deterministic information. Tuning parameter β controls the maximum magnitude for the random walk. Its value should be large enough to make sufficient jumps from the current position, but small enough so that the gradient information is not completely masked by the random walk. In fact, if $\beta = 0$, equation 17 becomes the preconditioned gradient-descent update. However, if $\alpha = 0$, equation 17 simulates a random walk MCMC algorithm with some constraints given by $\tilde{\mathbf{H}}$. The proposal distribution equation 16 clearly shows the connection between deterministic problems and statistical problems. For acoustic waveform inversion, our experiences suggest that $\alpha = 0.05 \sim 0.15$ and β is $5\alpha \sim 10\alpha$, or β is $2\sigma \sim 3\sigma$, where σ is the expected standard deviation of the posterior distribution, are good starting values for tuning the GMCMC. Further analysis related to the tuning parameters can be found in the "Discussion" section.

In the traditional MCMC methods, no local geometric information is used in constructing the proposal distribution, which might lead to a large mismatch between the proposal distribution and the posterior distribution, resulting in low acceptance rates and an inefficient sampling process. In contrast, GMCMC exploits the derivative information to obtain the proposal distribution. The proposal distribution carries a good local representation of the underlying posterior distribution. In this way, the sampling process can focus on sampling regions matched closely with the posterior distribution, leading to high convergence rates and improved MCMC performance. In Figure 1, we plot three proposal distributions overlaid with the Rosenbrock function where the red contours represent high-probability regions of the target distribution. The random walk MCMC proposal distribution would explore the target distribution in a very inefficient fashion. The proposal distributions with local geometric information match the target distribution well. Note that the proposal distribution with only the diagonal Hessian (Figure 1d) shows no correlations. In other words, each parameter is independently perturbed. However, the proposal distribution with the full Hessian (Figure 1c) contains correlation information.

With $\tilde{\mathbf{H}}^{-1}$ and $\tilde{\mathbf{H}}^{-1/2}$ defined, we can draw a sample $\mathbf{y}$ from the proposal distribution. The sample is then subjected to the accept/reject M-H criterion. Given that an MCMC chain is long enough, the sampling process is able to sufficiently explore the model space and accurately estimate the posterior distribution. We demonstrate such a workflow in Algorithm 1, and we call the proposed sampling method GMCMC.

Algorithm 1. GMCMC algorithm to sample $\pi_{\text{post}}(\mathbf{m}|\mathbf{d}_{\text{obs}})$.

```
Choose $\mathbf{m}_0$, $\alpha$, and $\beta$
Compute $\pi_{\text{post}}(\mathbf{m}_0)$, $\mathbf{g}(\mathbf{m})_0$, and $\tilde{\mathbf{H}}(\mathbf{m}_0)$
for $k$=0, ..., $N-1$ do
  Define $q(\mathbf{m}_k, \mathbf{y}) = \tilde{\pi}(\mathbf{y})$ as in equation 16
  Draw sample $\mathbf{y}$ from the proposal distribution $q(\mathbf{m}_k, \bullet)$ with equation 17
  Compute $\pi_{\text{post}}(\mathbf{y})$, $\mathbf{g}(\mathbf{y})$, and $\tilde{\mathbf{H}}(\mathbf{y})$
  Compute $\gamma(\mathbf{m}_k, \mathbf{y}) = \min\left(1, \frac{\pi_{\text{post}}(\mathbf{y})q(\mathbf{y},\mathbf{m}_k,)}{\pi_{\text{post}}(\mathbf{m}_k)q(\mathbf{m}_k,\mathbf{y})}\right)$
  Draw random number u $\sim\mu([0, 1])$
  if u $< \gamma(\mathbf{m}_k, \mathbf{y})$ then
    Accept: Set $\mathbf{m}_{k+1} = \mathbf{y}$
  else
    Reject: Set $\mathbf{m}_{k+1} = \mathbf{m}_k$
  end if
end for
```

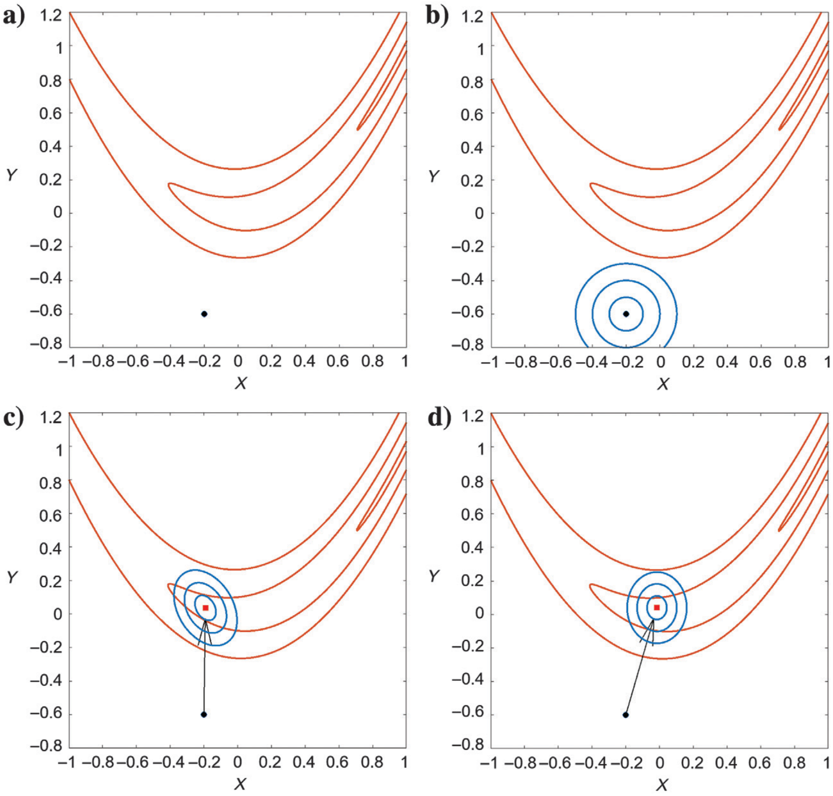


Figure 1. (a) Rosenbrock function $f(x, y) = (1 - x)^2 + 100(y - x^2)^2$, the black dot represents the current point in the model space, (b) proposal distribution of the random-walk MCMC method, (c) proposal distribution of the stochastic Newton MCMC method, and (d) proposal distribution of the GMCMC method. The blue contours represent σ, 2σ, and 3σ of the proposal distribution. The red squares in (c and d) are the point after the preconditioned gradient updates.

FREQUENCY-DOMAIN SEISMIC WAVEFORM INVERSION

In this section, we briefly describe several necessary components for implementing the proposed GMCMC method in the frequency-domain acoustic FWI settings. Only the velocity v is treated as the model parameter for the inversion.

In the frequency domain, the constant-density acoustic wave equation is (Pratt et al., 1998)

$$-\left(\frac{\omega^2}{v^2(\mathbf{x})}+\nabla^2\right)u(\mathbf{x}_s,\mathbf{x},\omega)=f(\omega)\delta(\mathbf{x}-\mathbf{x}_s), \tag{18}$$

where ω is the angular frequency, v is the velocity, ∇^2 represents the Laplacian operator, f is the source term, $\mathbf{x}_s$ is the source location, and $\mathbf{x}$ is the model parameter location. The gradient of the misfit function is given by

$$\begin{aligned} g(\mathbf{x}) = \Re\Bigg(&\sum_{\omega}\sum_{\mathbf{x}_s}\sum_{\mathbf{x}_r}\omega^2 f(\omega)G(\mathbf{x},\mathbf{x}_s,\omega)G(\mathbf{x},\mathbf{x}_r,\omega)\\ &\times\Gamma^{-1}_{\text{noise}}(\mathbf{x}_s,\mathbf{x}_r,\omega)\Delta d^*(\mathbf{x}_s,\mathbf{x}_r,\omega)\\ &+\sum_{\mathbf{x}_j}\Gamma^{-1}_{\text{prior}}(\mathbf{x},\mathbf{x}_j)(v(\mathbf{x}_j)-v_{\text{prior}}(\mathbf{x}_j))\Bigg), \end{aligned} \tag{19}$$

where $\Gamma^{-1}_{\text{noise}}(\mathbf{x}_s,\mathbf{x}_r,\omega)$ can be regarded as the weight, $\Delta d(\mathbf{x}_s,\mathbf{x}_r,\omega)=d_{\text{obs}}(\mathbf{x}_s,\mathbf{x}_r,\omega)-d_{\text{cal}}(\mathbf{x}_s,\mathbf{x}_r,\omega)$, the superscript $*$ represents the complex conjugate, and $G(\mathbf{x},\mathbf{x}_s,\omega)$ and $G(\mathbf{x},\mathbf{x}_r,\omega)$ are the source- and receiver-side Green's functions, respectively. Equation 19 is often implemented by the adjoint-state method. One can also use the plane-wave domain gradient computation method proposed by Zhao and Sen (2017) to improve the computational efficiency.

The approximate Hessian $\mathbf{H}_a$ can be computed by

$$\begin{aligned} H_a(\mathbf{x}_i,\mathbf{x}_j) = \Re\Bigg(&\sum_{\omega}\sum_{\mathbf{x}_s}\sum_{\mathbf{x}_r}\omega^4 f^2(\omega)G^*(\mathbf{x}_i,\mathbf{x}_s,\omega)\\ &\times G(\mathbf{x}_j,\mathbf{x}_s,\omega)\Gamma^{-1}_{\text{noise}}(\mathbf{x}_s,\mathbf{x}_r,\omega)\\ &\times G^*(\mathbf{x}_i,\mathbf{x}_r,\omega)G(\mathbf{x}_j,\mathbf{x}_r,\omega). \end{aligned} \tag{20}$$

Taking the diagonal part of $\mathbf{H}_a$ and combining it with $\Gamma^{-1}_{\text{prior}}$, we obtain $\tilde{\mathbf{H}}$ as

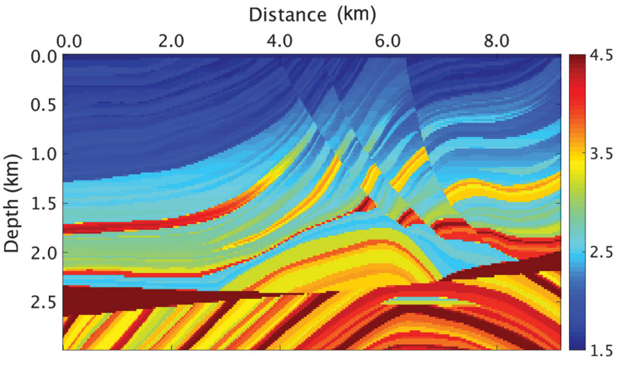


Figure 2. The 2D Marmousi velocity model.

$$\tilde{H}(\boldsymbol{x}_i,\boldsymbol{x}_j)=K(\boldsymbol{x}_i,\boldsymbol{x}_j)+\Gamma^{-1}_{\text{prior}}(\boldsymbol{x}_i,\boldsymbol{x}_j), \tag{21}$$

with

$$\begin{aligned} K(\mathbf{x}_i,\mathbf{x}_j) = \Re\Bigg(&\sum_{\omega}\sum_{\mathbf{x}_s}\sum_{\mathbf{x}_r}\omega^4 f^2(\omega)G^*(\mathbf{x}_i,\mathbf{x}_s,\omega)\\ &\times G(\mathbf{x}_j,\mathbf{x}_s,\omega)\Gamma^{-1}_{\text{noise}}(\mathbf{x}_s,\mathbf{x}_r,\omega)\\ &\times G^*(\mathbf{x}_i,\mathbf{x}_r,\omega)G(\mathbf{x}_j,\mathbf{x}_r,\omega)\delta_{ij}\Bigg), \end{aligned} \tag{22}$$

where δ_{ij} is the delta function; it is nonzero only at locations $i=j$. Hence, $\mathbf{K}$ is a diagonal matrix. Computing $\mathbf{K}$ according to equation 22 can be computationally expensive when the number of sources and receivers is large. Plessix and Mulder (2004) suggest several approximations of $\mathbf{K}$. We implement the "type 3" approximation (i.e., $\mathbf{K}_3$ in Plessix and Mulder, 2004) in the waveform inversion. The off-diagonal terms of $\mathbf{K}_3$ are zeros, and the diagonal terms of $\mathbf{K}_3$ can be written as

$$\begin{aligned} K_3(\mathbf{x}_i,\mathbf{x}_i) = \Re\Bigg(&\sum_{\omega}\sum_{\mathbf{x}_s}\omega^4 f^2(\omega)G^*(\mathbf{x}_i,\mathbf{x}_s,\omega)G(\mathbf{x}_i,\mathbf{x}_s,\omega)\\ &\times 1/2(\Gamma^{-1}_{\text{noise}}(\mathbf{x}_s,x_r^{\max}(\mathbf{x}_s),\omega)+\Gamma^{-1}_{\text{noise}}(\mathbf{x}_s,x_r^{\min}(\mathbf{x}_s),\omega))\\ &\times\left(\mathrm{a\,sinh}\left(\frac{x_r^{\max}(\mathbf{x}_s)-\mathbf{x}_i(x)}{\mathbf{x}_i(z)}\right)-\mathrm{a\,sinh}\left(\frac{x_r^{\min}(\mathbf{x}_s)-\mathbf{x}_i(x)}{\mathbf{x}_i(z)}\right)\right)\Bigg), \end{aligned} \tag{23}$$

where $x_r^{\max}(\mathbf{x}_s)$ and $x_r^{\min}(\mathbf{x}_s)$ are the maximum and minimum receiver locations for shot $\mathbf{x}_s$, respectively, and $\mathbf{x}_i(x)$ and $\mathbf{x}_i(z)$ are the horizontal and vertical location of $\mathbf{x}_i$, respectively. Here, we take the average of $\Gamma^{-1}_{\text{noise}}$ for weighting.

Of note, $\tilde{\mathbf{H}}=\mathbf{K}+\Gamma^{-1}_{\text{prior}}$ generally contains off-diagonal terms due to $\Gamma^{-1}_{\text{prior}}$; drawing samples with such $\tilde{\mathbf{H}}$ requires large matrix manipulations, which can be computationally expensive in high dimensions. Therefore, we drop the $\Gamma^{-1}_{\text{prior}}$ term for the covariance matrix of the proposal distribution. As a result, a sample drawn from the proposal distribution contains no correlation information from the prior.

NUMERICAL EXAMPLES

We demonstrate the proposed GMCMC FWI method on the 2D Marmousi model (shown in Figure 2). In the examples, the observed data and estimated data are generated using different modeling methods on different computational grids. The source wavelet is assumed to be known for the inversion.

Inversion setup

Observed data generation

The grid size for generating the observed data is $460(nx_0)\times 150$ (nz_0), with grid spacing of 0.02×0.02 km. We use the time-domain rapid expansion method (Pestana and Stoffa, 2010) to generate the shot gathers. The spatial derivatives are computed by the pseudo-spectral method. A Ricker wavelet with the peak frequency at 10 Hz is used as the source. We simulate 92 shots with a 0.1 km shot interval; each shot is recorded by 460 receivers with a 0.02 km receiver interval. Noncorrelated Gaussian white noise is added to the observed data. One of the observed shot gathers is shown in Figure 3.

Model parameterization and computation

The model parameters are set to be on a $245(nx) \times 79(nz)$ regular grid, with grid spacing of 0.0375×0.0375 km. So, there are 19,355 model parameters in total. The same grid is used to estimate $\mathbf{d}_{\mathrm{cal}}$, $\mathbf{g}$, and $\tilde{\mathbf{H}}$. The top 0.15 km of the model is replaced with the ground truth, assuming that we have accurate shallow subsurface velocity estimations. This part of the model is kept unchanged during the inversion. The frequency-domain finite-difference method (Chen et al., 2013) is used as the modeling engine in the inversion. We simultaneously invert data for six frequencies ranging from 5.0 to 7.5 Hz with a 0.5 Hz interval. We use six cores on an Intel Xeon E5-2690 CPU to run the numerical tests where each core processes a single frequency. Each iteration, including computing the gradient, the approximate diagonal Hessian, and drawing a sample, takes approximately 1.1 s wall clock time. We run several chains with 200,000 iterations, and each chain can finish running within 61 h.

Likelihood and prior information

With the noncorrelated Gaussian white noise, the likelihood function is defined as

$$\pi_{\mathrm{like}}(\mathbf{d}_{\mathrm{obs}}|\mathbf{m}) \propto \exp\left(-\frac{1}{2}(\mathbf{d}_{\mathrm{cal}} - \mathbf{d}_{\mathrm{obs}})^{\dagger}\Gamma_{\mathrm{noise}}^{-1}(\mathbf{d}_{\mathrm{cal}} - \mathbf{d}_{\mathrm{obs}})\right), \tag{24}$$

where † represents the complex conjugate transpose and the noise covariance matrix $\Gamma_{\mathrm{noise}}^{-1}$ is a diagonal matrix with constant variance $\sigma = 0.06$.

We assume the model parameters to be smooth and specify as little a priori knowledge as possible. The prior distribution is defined as

$$\pi_{\mathrm{prior}}(\mathbf{m}) \propto \begin{cases} \exp\left(-\frac{1}{2}\mathbf{m}^T\Gamma_{\mathrm{prior}}^{-1}\mathbf{m}\right) & \text{if } v\min_j \leq m_l \leq v\max_j \ \forall\ l, \\ 0 & \text{otherwise,} \end{cases} \tag{25}$$

where $l = i \times nz + j$, i and j indicate the horizontal and vertical locations, respectively, $v\min_j$ and $v\max_j$ are the minimum and maximum velocity values for given j, respectively, and the $v\min_j$ and $v\max_j$ values can be found in Figure 5b. The matrix $\Gamma_{\mathrm{prior}}^{-1}$ is defined as the second-order differential operator to penalize the roughness of a sample.

Tuning parameters α and β

We first demonstrate the effect of α and β on generating a sample. In Figure 4, we plot $\tilde{\mathbf{H}}$, $\alpha\tilde{\mathbf{H}}^{-1}\mathbf{g}$, $\beta\tilde{\mathbf{H}}^{-\frac{1}{2}}\mathbf{r}$, and $\alpha\tilde{\mathbf{H}}^{-1}\mathbf{g} + \beta\tilde{\mathbf{H}}^{-\frac{1}{2}}\mathbf{r}$ for a homogeneous model with $v = 2.25$ km/s. Figure 4a shows $\tilde{\mathbf{H}}$, which is also recognized as the illumination compensation. The cold colors, corresponding to small values, represent poorly illuminated areas. Hence, applying $\tilde{\mathbf{H}}^{-1}$ and $\tilde{\mathbf{H}}^{-\frac{1}{2}}$ to $\mathbf{g}$ and $\mathbf{r}$ means that a proposal sample takes the illumination compensation into consideration. As a result, a new sample has mostly smaller updates for small variance areas than that for large variance areas, as shown in Figure 4c. In Figure 4d, we plot the total model update.

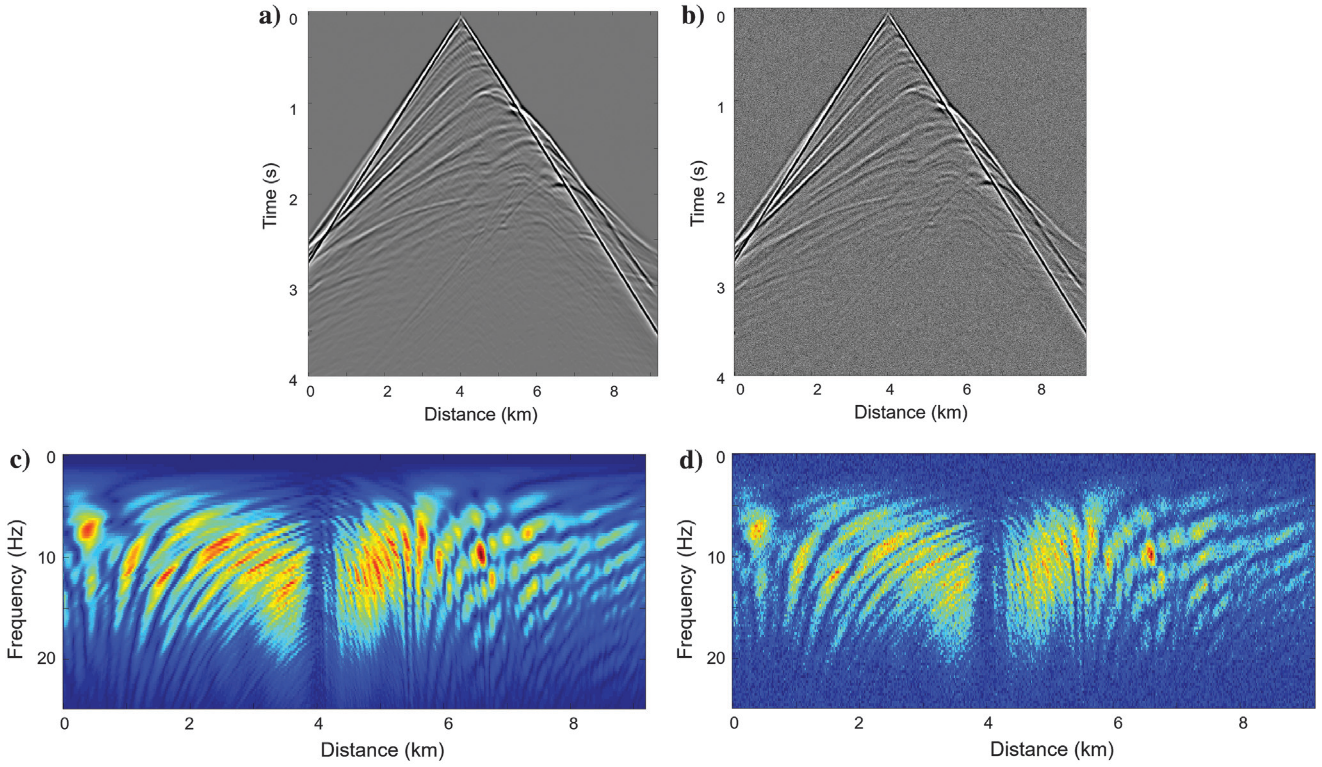


Figure 3. (a and b) Clean and noisy data in the time domain. (c and d) The corresponding data in the frequency domain.

Starting from different models

We use four models shown in Figure 5a as the starting models to test the behavior of Markov chains with different starting points. Here, we set $\alpha = 0.08, \beta = 0.8$. The models are: V_{smooth}, a smoothed version of the true velocity model; V_{constant}, a constant velocity model with v = 2.25 km/s; V_{gradient}, a 1D velocity model in which the velocity values increase linearly with a small gradient; and V_{random}, a random velocity model drawn from the defined prior bounds. In Figure 5b, we plot vertical profiles at four locations for the starting models versus the true model. Three of the four starting models are deliberately chosen to be far away from the true model. The V_{constant} and V_{gradient} models overestimate the shallow parts of the model and largely underestimate the deep parts of the model, although the V_{gradient} model is closer to the true model in the shallow parts. The V_{random} model overestimates most parts of the model, especially the shallow parts, by more than 2 km/s.

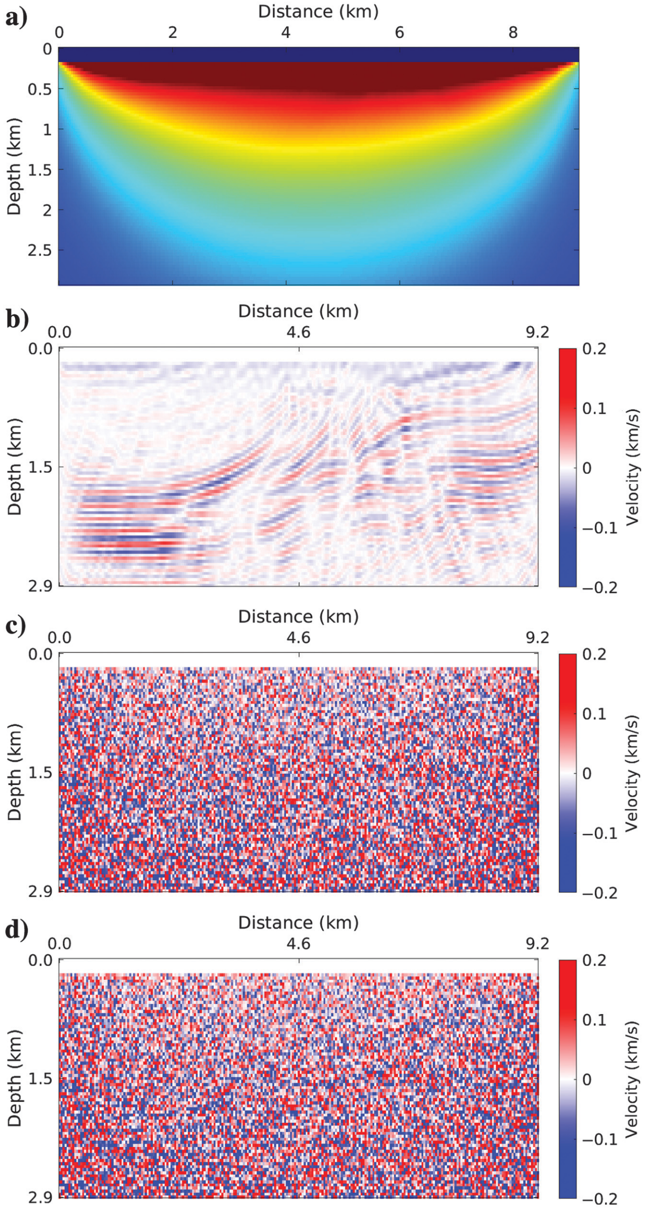


Figure 4. Plots of (a) $\tilde{\mathbf{H}}$, (b) $\tilde{\mathbf{H}}^{-1}\mathbf{g}$, (c) $\tilde{\mathbf{H}}^{-\frac{1}{2}}\mathbf{r}$, and (d) $\alpha\tilde{\mathbf{H}}^{-1}\mathbf{g} + \beta\tilde{\mathbf{H}}^{-\frac{1}{2}}\mathbf{r}$, where $\alpha = 0.08$ and $\beta = 0.4$.

We first initiate one chain for each starting model and test 50,000 samples for each chain. The L2 error curves for the tests are shown in Figure 6. All curves reach the steady state and fluctuate approximately 0.4–0.5, although with a different number of iterations. The V_{smooth} case has the smallest initial misfit, and it reaches the steady state very fast. Because V_{smooth} is very close to the true model, this chain starts sampling high-probability regions of the posterior distribution at the early stage in this case. The V_{gradient} reaches the steady state at approximately 3000 iterations, which is slower than the V_{smooth} case. Because the shallow parts of the V_{gradient} model approximate the true model, it is easy for the chain to find the regions of high posterior probability for the shallow parts. However, the deep parts of this starting model are far away from the true model, which makes the chain take more steps than the V_{smooth} case to reach the same steady state. The V_{constant} model is a very uninformative starting model. As a result, the initial error is higher, and it takes 6000 iterations for this chain to reach the steady state. The V_{random} case has the largest initial error, and it takes more than 20,000 iterations for the chain to reach the steady state, suggesting that the V_{random} model is located in a very low probability region in the model space due to the very large deviations from the true model. The chain spends a very long time exploring the model space before identifying regions of high posterior probability.

We discard the first 10,000 samples for each chain after it reaches the steady state because they are still at the early stage in the chains. For each chain, we use the following 10,000 samples to generate the mean model for each chain as shown in Figure 7. Note that we have not shown the convergence for the chains, and likely they have not converged, at least not for all of the locations. However, even though the chains have not formally converged, by plotting the mean models, we see that the mean models for the four chains are overall close to each other and they all resemble the true model. This indicates that all chains have reached the high posterior probability regions.

Final results

As demonstrated in the previous section, the initial model V_{random} lies in the very low probability region of the model space, which makes the chain spend too much time before sampling the high posterior probability regions. Thus, the V_{random} model is excluded in the following test. We keep the other three chains running to 200,000 iterations, and we initiate one more chain for each of the three models. Each new chain is set to run 200,000 iterations as well. A total of 1,200,000 samples are tested for the six chains, and the accept ratio is approximately 50%. The L2 error curves for the six chains are plotted in Figure 8. All chains reach the same steady state.

We discard the first 20,000 samples in each chain, treating them as the burn-in samples. Post burn-in samples are used to generate the mean model and standard deviation map shown in Figure 9. We also include an absolute difference map between the mean model and the true model in Figure 9. Note that the mean model is mostly close to the true model, with some differences at the edges and the bottom of the model. The standard deviation map, which is very similar to the inverse of the illumination map, suggests small deviations (approximately 0.0–0.3 km/s) in the shallow parts of the model, where we have good data coverage and good illumination.

This can be verified by examining marginal PPDs, shown in Figure 10, at x = 3.00, 4.31, 5.62, and 6.94 km. Small deviations indicate that velocity values are well constrained in these areas and the inversion results have less uncertainties. In contrast, the deep parts and the edges of the model, especially the lower left and lower right corners, have the largest variations (0.8 km/s and above) due to poor data coverage and poor illumination, for example, marginal PPDs at x = 1.12 km and x = 8.06 km in Figure 10. Inverted velocity values are less constrained and are spread out over a wider range of values, suggesting large uncertainties for the inverted values. Examining the marginal PPDs, we find that the mean model generally predicts the true model very well, especially in areas with good data constraints. In most parts of the model, the true velocity values fall in the 90% probability intervals of marginal PPDs.

Convergence analysis

Here, we examine several characteristics of the chains to diagnose the convergence of the tests. We first plot marginal PPDs at several locations at different stages in the chains after the burn-in phase. Locations at the shallow, middle, and deep parts of the model are selected, and their corresponding marginal PPDs are shown in Figure 11. We see few differences in marginal PPDs at the same location while increasing numbers of samples are included, an indication of convergence to the stationary distributions for the selected locations. Similar behavior is observed for marginal PPDs at other locations as well. We examine the potential scale reduction factor (PSRF) $\hat{R}$, an MCMC convergence diagnostic tool proposed by Brooks and Gelman (1998), at different locations. PSRF

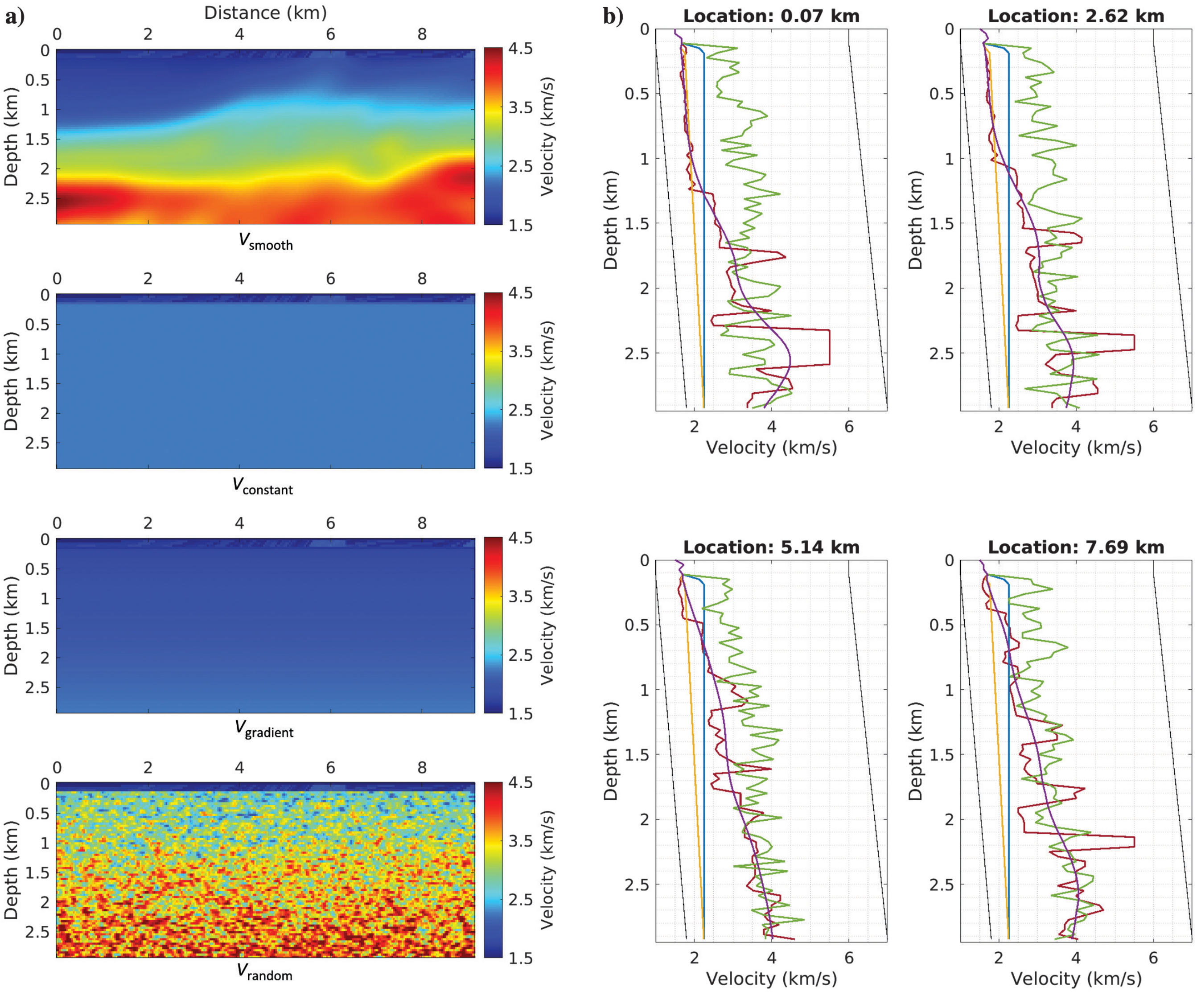
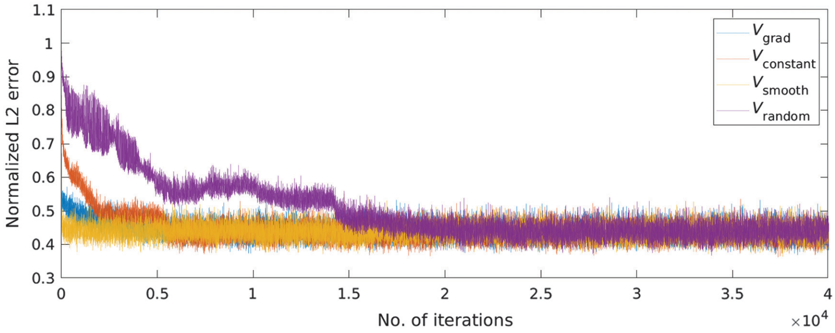


Figure 5. (a) Four starting models for the inversion: a smoothed version of the true model, V_{smooth}; homogeneous velocity model with v = 2.25 km/s, V_{constant}; a velocity model in which the velocity increases with depth, V_{gradient}; and a random starting model drawn from the lower and upper bounds, V_{random}. (b) Vertical profiles comparing the four starting models with the true model and the lower and upper bounds. The red, magenta, blue, yellow, and green curves represent the true model, V_{smooth}, V_{constant}, V_{gradient}, and V_{random}, respectively. The solid black lines are the prior information, that is, the lower and upper velocity bounds.

Figure 6. The normalized L2 error curves for the four tests. The yellow, blue, red, and magenta curves correspond to the V_{smooth}, V_{gradient}, V_{constant}, and V_{random} cases, respectively. Because all chains have stabilized toward the end of the iterations, we truncate the display at the 40,000th iteration.



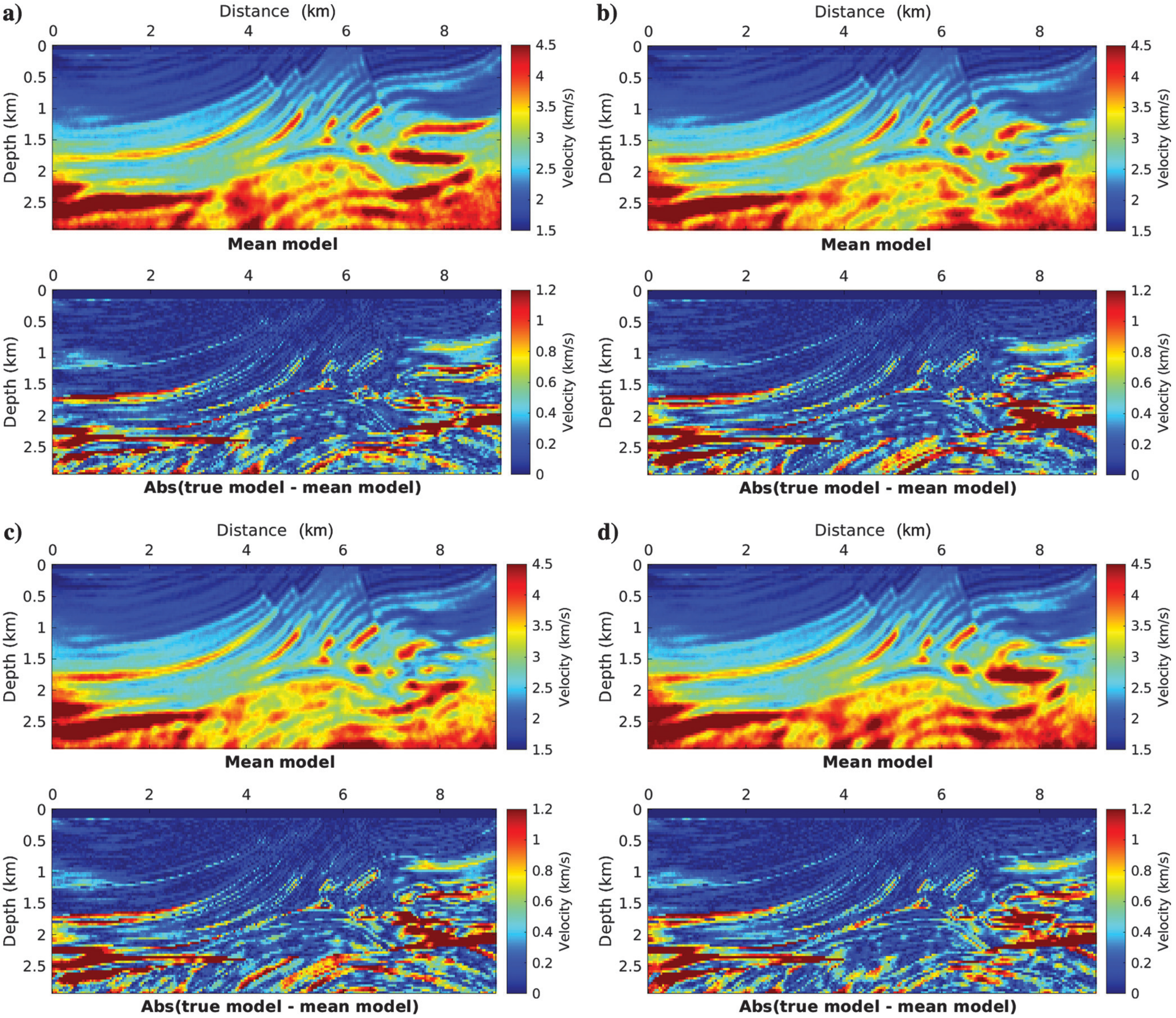


Figure 7. The mean models of the 10,000 samples for each test, and the absolute error maps with respect to the true model. (a) V_{smooth} test, (b) V_{constant} test, (c) V_{gradient} test, and (d) V_{random} test.

compares within-chain variances to the variance computed from all mixed chains for a parameter. Its value should be close to one if the parameter is close to its posterior distribution. Often in practice, one can consider the convergence for parameters of interest when $\hat{R} < 1.1$ (Brooks et al., 2011). The formal definition of $\hat{R}$ can be found in Brooks and Gelman (1998). We select one well-illuminated area and one poorly illuminated area, indicated by the white lines in Figure 12a, to generate their $\hat{R}$ values, shown in Figure 12b. The $\hat{R}$ values for the shallow part drop below the threshold value 1.1 very quickly, whereas it takes a while before $\hat{R}$ values for the deep part become smaller than 1.1. It takes more steps to achieve the stationary distributions for poorly illuminated areas. This is due to the nature of the surface seismic data inversion, where shallow parts of a model tend to have better data coverage and illumination, hence better constraints than deep parts of the model. Therefore, it usually takes fewer iterations for the well-illuminated areas to reach stationary distributions. If poorly illuminated areas are of little interest, one might be able to run shorter chains. If the formal convergence for the entire model is desired, however, longer chains are necessary.

The multivariate potential scale reduction factor (MPSRF), a convergence diagnostic tool for multiple parameters (Brooks and Gelman, 1998), is used here to analyze the convergence for a greater area. MPSRF compares within-chain covariance matrix $\mathbf{W}$ with the pooled sample chain covariance matrix $\hat{\mathbf{V}}$, estimated from all of the chains. It summarizes all PSRF sequences of interest in a single sequence, which can be used to assess the convergence for all interested parameters (Martin et al., 2012; Stuart et al., 2019). MPSRF is defined as

$$\hat{R}^p = \max_{\mathbf{a}} \frac{\mathbf{a}^T\hat{\mathbf{V}}\mathbf{a}}{\mathbf{a}^T\mathbf{W}\mathbf{a}} = \frac{n-1}{n} + \frac{m+1}{m}\lambda_1, \qquad (26)$$

where n is the number of iterations, m is the number of chains, and λ_1 is the eigenvalue of the matrix $\mathbf{W}^{-1}B/n$. Definitions for $\mathbf{W}$, $\hat{\mathbf{V}}$, and $\mathbf{B}$ can be found in Brooks and Gelman (1998). The value of $\hat{R}^p$ should approach one as the chains converge. Here, we compute MPSRF values at different stages in the chains, i.e., Figure 13, for parameters within the range of x = [1.1258.0625] km and z = [0 2.4375] km. The edges and the bottom of the model are excluded because they are typically of little interest. When computing $\hat{R}^p$, we use every other point on the inversion grid to reduce the cost of computing the large matrices $\mathbf{W}$, $\hat{\mathbf{V}}$, and $\mathbf{B}$. Otherwise, it would be computationally expensive to perform the convergence analysis for this high-dimensional problem (Brooks and Gelman, 1998). In Figure 13, $\hat{R}^p$ drops rapidly from large values and it approaches one as the chains evolve. Because $\hat{R}^p$ is not computed using every parameter within the target range, we infer that the chains achieve at least approximate convergence. Of note, we do recommend always running chains longer than might be necessary to ensure the convergence. Because none of the convergence diagnostic methods is free of deficiencies, the theoretically full convergence might only be achieved asymptotically. Here, we believe that our analyses are adequate for our FWI problem.

DISCUSSION

In MCMC sampling methods, it is necessary to sample the posterior distribution adequately to sufficiently approximate its distribution. In large-scale inverse problems, the computational cost is very high due to the high model dimensions. As shown in the numerical tests, the computational cost related to implementing

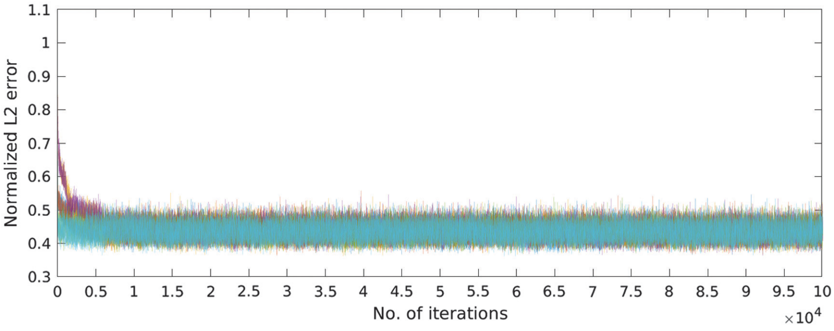


Figure 8. The normalized L2 error curves for the six chains. Because all chains have stabilized toward the end of the iterations, we truncate the display at the 100,000th iteration.

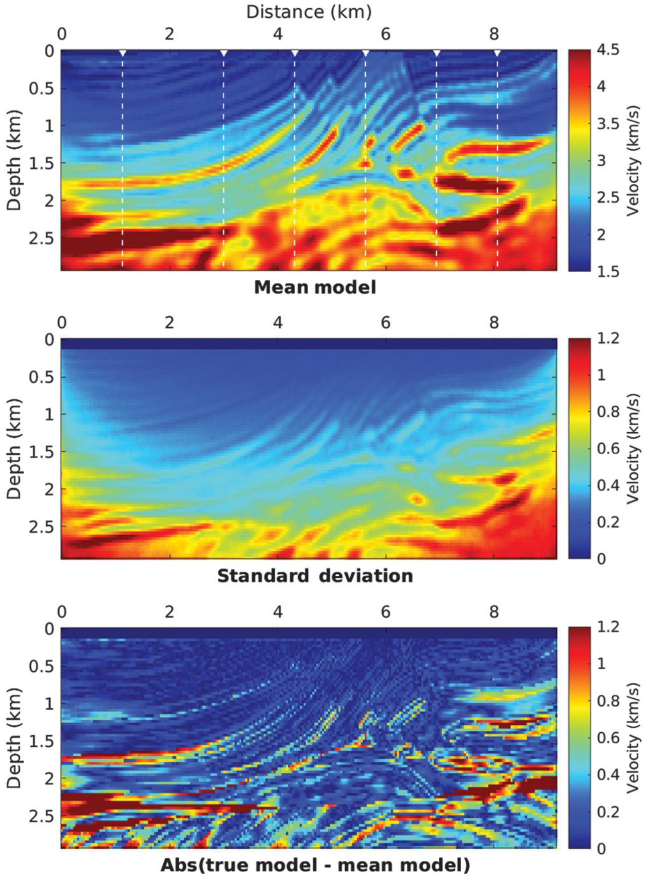


Figure 9. The mean model, the standard deviation map for samples after the burn-in phase, and the absolute difference between the true model and the mean model. The dashed white lines indicate the locations for plotting the marginal PPDs as shown in Figure 10.

GMCMC for 2D frequency-domain acoustic FWI is very high compared to local optimization-based methods that typically run no more than 1000 gradient updates. However, we do show that it is still computationally affordable to run the proposed method for a high-dimensional nonlinear sampling problem with only one CPU. With faster forward solvers, better code implementations, and faster CPUs, the computing time of GMCMC FWI can be further reduced, although it is still very computationally demanding. Realistic 3D FWI problems will pose challenges to any available MCMC algorithms. Therefore, we recommend several strategies to tackle the computational challenges, especially for the large-scale 3D problems. (1) Optimizing code implementations for wave equation solvers can improve computational efficiencies. Etienne et al. (2014) present several strategies by which the computational cost for time-domain 3D FWI can be reduced by more than an order of magnitude. (2) Using fast computing units, for instance, graphic computing units, can speed up calculations. (3) Using sparse representation methods can reduce model dimensions. Several researches (Sajeva et al., 2016; Ray et al., 2017; Hunziker et al., 2019) show the effectiveness of this dimension reduction strategy for FWI with different MCMC algorithms. Additionally, one can carry out MCMC FWI in a layer-stripping fashion to first estimate the posterior distribution for the shallow parts of a model and then use the obtained distributions as prior knowledge to constrain the sampling for the deep parts of the model. (4) Using advanced sampling methods can improve the efficiency for sampling the posterior distribution, for instance, the full HMC method. Well-designed sampling methods can efficiently explore the model space, resulting in improved convergence rates.

The a priori knowledge has a significant impact on inversion results. Too general a priori information providing little useful information can significantly increase the computational effort. Too restricted prior knowledge might lead to inversion results biased toward specific information. In the numerical examples, we deliberately make the inversion difficult by using uninformative prior knowledge and poor starting models. As shown in the V_{smooth} case, starting with an accurate background model helps to sample the regions of high posterior probability from the early stage, resulting in very short burn-in phases. Furthermore, informative and geologically meaningful prior constraints also help in regularizing samples that resemble plausible earth models. In real applications, models representing subsurface velocity trends, tomography results, and existing wells are good prior knowledge for constructing the starting point and the prior distribution.

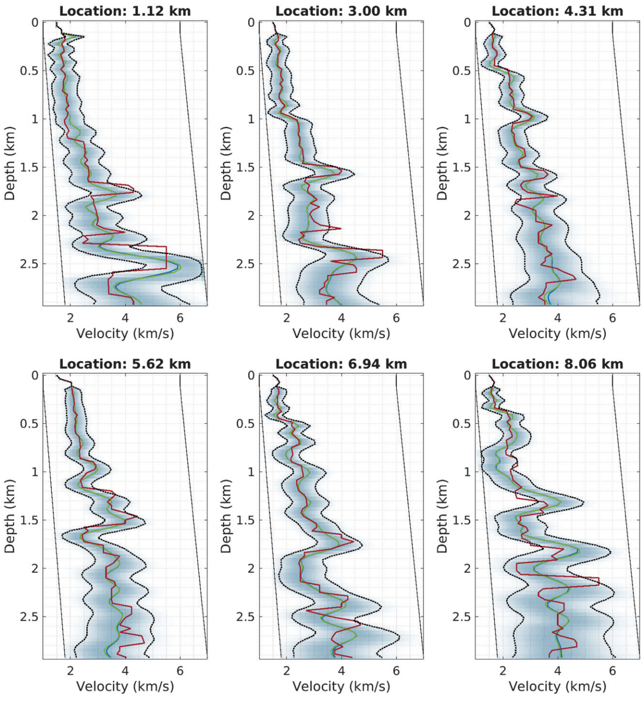


Figure 10. Selected marginal PPDs at the six locations labeled in Figure 9. Marginal PPDs at the locations 1.12 and 8.06 km represent poorly illuminated areas, whereas other marginal PPDs represent areas with better illumination. The blue, green, and red curves represent the mean, median, and the true values, respectively. The darker and lighter colors of the cloud plots correspond to the high and low probability densities, respectively. The black dotted curves represent the 5% and 95% probability interval bounds. The solid black lines are the lower and upper velocity bounds.

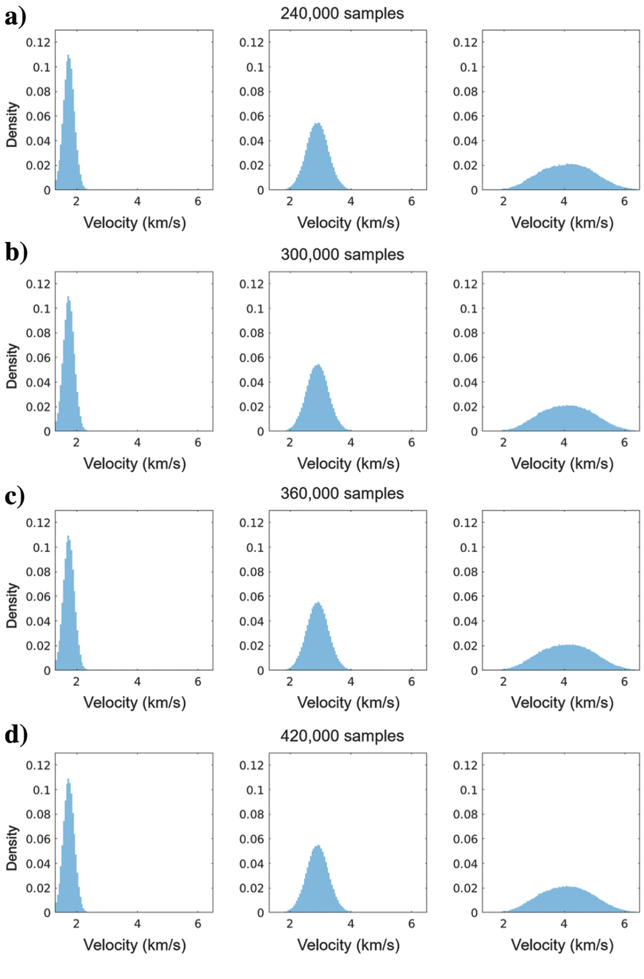


Figure 11. From left to right are the marginal PPDs at different stages in the chains for (x = 3.975, z = 0.375) km, (x = 3.975, z = 1.5) km, and (x = 3.975, z = 2.625) km: (a) 40,000 samples per chain, of the total 240,000 samples, (b) 50,000 samples per chain, of the total 300,000 samples, (c) 60,000 samples per chain, of the total 360,000 samples, and (d) 70,000 samples per chain, of the total 420,000 samples. Chains are thinned by jumping over every other sample to save storage space.

In the numerical tests, parameters are defined on the same grid as that used for computing the data residual and the gradient. One might choose to define parameters on different grid settings. A sparse inversion grid leads to reduced model dimensions; as a result, faster convergence rates might be achieved at the expense of the spatial resolution. From the optimization-based FWI point of view, changes of velocity values on a sparse grid might have more impacts on the traveltime for wavefields. A sparse inversion grid might be beneficial for low-wavenumber components/traveltime update, which might reduce the time to identify regions with high posterior probability, resulting in a shorter burn-in phase. On the contrary, when parameters are defined on a very dense grid, the details of subsurface structures can be better described. However, the MCMC sampling process would take longer to converge to the posterior distribution due to the increased model dimensions. An alternative to the fixed inversion grid, which fixes the number of parameters, is the transdimensional inversion strategy (Bodin and Sambridge, 2009; Biswas and Sen, 2017; Ray et al., 2017) in which the number of model parameters is regarded as an inversion parameter. In this way, the number of model parameters is determined by the observed data and the uncertainty associated with the number of parameters would also be taken into consideration.

When defining the proposal distribution, we drop several components contributing to the off-diagonal terms of the covariance matrix, i.e., equation 16, including $\mathbf{R}$, $\Gamma_{\text{prior}}^{-1}$, and off-diagonal parts of $\mathbf{H}_a$, to only retain $\mathbf{K}$, which makes the sampling process computationally efficient. As a result, a sample drawn from this proposal distribution contains no correlation information between different parameters. This proposal distribution is a good local approximation to the target distribution in cases in which parameters exhibit little correlations. However, if strong correlations exist between parameters, mismatches between the proposal distribution and the underlying target distribution would increase, which might have negative impacts on the convergence rate. Therefore, when parameters are highly correlated, one might want to include off-diagonal terms for the covariance matrix to construct the proposal distribution that better matches the posterior distribution. For instance, in Figure 14, the proposal distribution constructed with the full Hessian simulates the positive correlation between the two parameters, hence better representing the target distribution than that

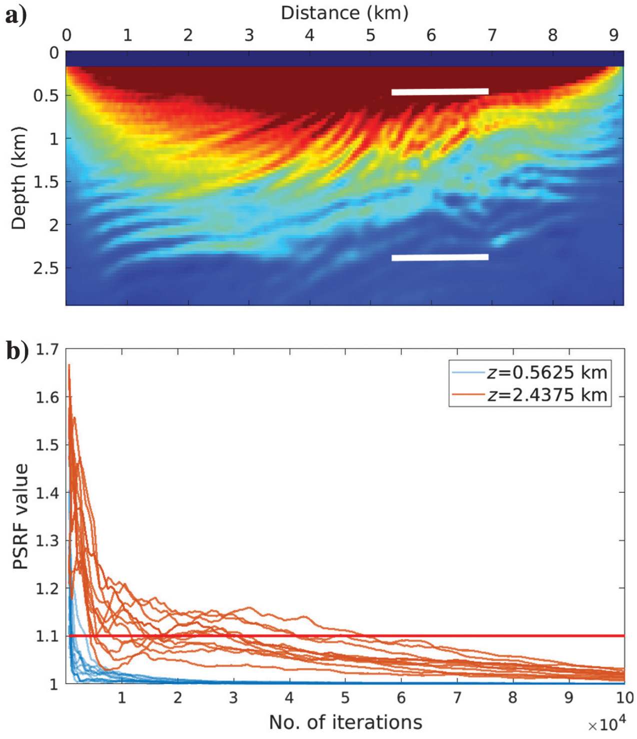


Figure 12. (a) The illumination map for the final mean model. The hot and cold colors represent good and poor illumination, respectively. The white lines indicate a well-illuminated area and a poorly illuminated area. There are 10 locations evenly distributed on each of the white lines whose $\hat{R}$ values are plotted in (b). (b) The 20 $\hat{R}$ curves computed for the locations indicated by the white lines. The blue lines represent $\hat{R}$ values at the depth $z = 0.6525$ km, and orange lines represent $\hat{R}$ values at the depth $z = 2.4375$ km. The red line indicates the threshold $\hat{R} = 1.1$.

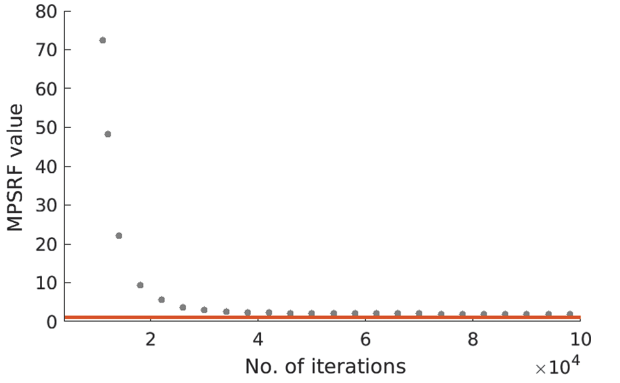


Figure 13. Plot of $\hat{R}^p$, MPSRF values, at different stages in the chains. The red line indicates the value one.

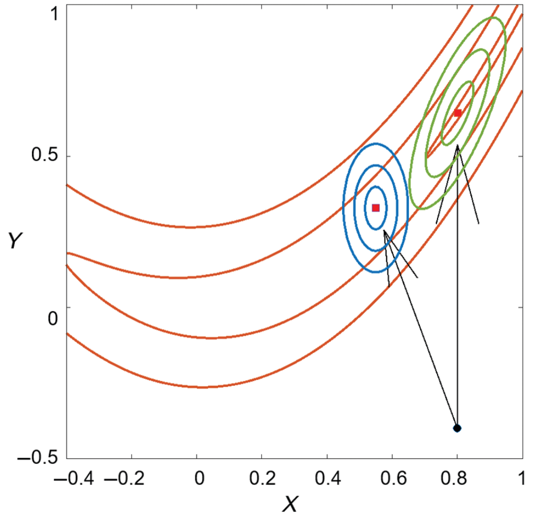


Figure 14. Two proposal distributions overlaid with the target distribution, Rosenbrock function as shown in Figure 1. The black dot is the current location. The blue and green contours represent σ, 2σ, and 3σ of the proposal distributions with only the diagonal Hessian and the full Hessian, respectively. The proposal distribution constructed with the full Hessian reflects the positive correlation between the parameters. It better represents the target distribution at the current step, whereas the proposal distribution constructed with only the diagonal Hessian shows no correlation.

constructed with only the diagonal terms of the Hessian. Sampling from the proposal distribution constructed with the full Hessian would improve the convergence rate in this scenario. For FWI problems in which the off-diagonal terms of $\mathbf{H}$ are strong, using a proposal distribution that considers the correlations between parameters might speed up the convergence. Nevertheless, one should also note that sampling from a proposal distribution that contains off-diagonal terms in the covariance matrix might require large matrix manipulations for large-scale problems, which increases the computational cost for sampling.

Tuning parameters α and β are important to the performance of the sampling process. Parameter α acts as the step length in gradient-descent methods. Its value should be similar to the one used in gradient-based local optimization methods, so that the linearized Taylor expansion is still locally valid. Parameter β controls the maximum variance of the proposal distribution. A too small β would make the chain move too slowly, whereas a too large β would make proposed samples be rejected too often. A good initial trial for choosing β is to set it close to 2σ or 3σ, where σ is the expected standard deviation of the posterior distribution. Poorly chosen α and β values would result in chains that converge too slowly to the posterior distribution. Good combinations of α and β would make the chains explore the model space efficiently and sufficiently, rendering reasonable convergence rates. Indeed, it is difficult, especially in high dimensions, to manually find the optimal values for α and β. The adaptive algorithms (Haario et al., 2001; Atchadé and Rosenthal, 2005; Atchadé, 2006), in which proposal distributions are automatically adjusted during the sampling process, might be beneficial to the proposed method.

CONCLUSION

We have presented a GMCMC sampling method based on the Bayesian inference framework to solve the ill-posed inverse problem in high dimensions. The main idea of the method is to construct a proposal distribution that is locally a good approximation to the posterior distribution. We show that, with the help of the local gradient and the diagonal approximate Hessian information, such a proposal distribution is easy to construct, and samples can be drawn from the proposal distribution efficiently. Drawing samples from such a proposal distribution can be regarded as updating the current model parameters with the preconditioned gradient plus a constrained random perturbation term. The preconditioned gradient guides the misfit going toward a “better” point, whereas the random perturbation term explores model space to avoid entrapment in a local region. The resultant GMCMC method samples the posterior distribution more efficiently. We implement the proposed method for the acoustic FWI problem in the frequency domain. In the synthetic example, we demonstrated that the mean and median values of the inverted statistical results well represent the ground truth even with different starting points that contain no informative prior information. It suggests that the proposed GMCMC method has the potential to make FWI a fully automatic process. Unlike traditional local optimization-based FWI methods, the results of the proposed GMCMC FWI provide statistical assessments by which the uncertainties related to the inversion can be estimated. We showed that within the 2D frequency-domain acoustic waveform inversion framework, the computational cost of the GMCMC method is affordable for high-dimensional problems.

ACKNOWLEDGMENTS

We would like to thank the associate editor P. Routh, reviewer Musa Maharramov, and three anonymous reviewers for their valuable comments and suggestions. This work was partially supported by an NSF grant EAR 1723019.

DATA AND MATERIALS AVAILABILITY

Data associated with this research are available and can be obtained by contacting the corresponding author.

APPENDIX A

COMPARISON BETWEEN GMCMC WITH LANGEVIN MC AND HMC

Similarity can be drawn between the proposed sampling method with Langevin MC (Grenander and Miller, 1994; Roberts and Tweedie, 1996; Stuart et al., 2004) and HMC methods (Neal, 2011). We rewrite equation 19 as

$$\boldsymbol{y} = \boldsymbol{m}_k - \alpha\tilde{\boldsymbol{H}}^{-1}\boldsymbol{g} + \beta\tilde{\boldsymbol{H}}^{-\frac{1}{2}}\boldsymbol{r}. \tag{A-1}$$

In Langevin MC, samples are produced according to the reversible Langevin diffusion process $\mathbf{X}_t$ that satisfies the stochastic differential equation

$$d\mathbf{X}_t = \frac{\sigma^2}{2}\nabla \log \pi_n(\mathbf{X})dt + \sigma\mathbf{W}_t, \tag{A-2}$$

where $\pi_n(\mathbf{X})$ is the n-dimensional posterior distribution, with variance σ^2, and $\mathbf{W}$ is the standard independent n-dimensional Brownian motion. Discretizing the diffusion, the sample of the next time step can be drawn by (Roberts and Tweedie, 1996; Stuart et al., 2004)

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \frac{\sigma^2}{2}\nabla \log \pi_n(\mathbf{x}_t) + \sigma\mathbf{W}. \tag{A-3}$$

One can precondition the update by a positive-definite matrix $\mathbf{A}$ as (Stuart et al., 2004)

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{A}\nabla \log \pi_n(\mathbf{x}_t) + \sqrt{2}\mathbf{A}^{1/2}\mathbf{W}. \tag{A-4}$$

We can clearly draw the connection between equations A-1 and A-3 or A-4. Dropping the tuning factors, equation A-1 can be interpreted as a preconditioned Langevin MC with the preconditioning matrix $\tilde{\mathbf{H}}^{-1}$.

If we set $\tilde{\mathbf{H}}^{-1} = \mathbf{I}$ and choose $\alpha = \sqrt{(2\beta)}$, we recover the first leapfrog step of HMC. In fact, Neal (2011) recognizes Langevin MC as a special case of HMC and makes a detailed comparison.

REFERENCES

Aleardi, M., and A. Mazzotti, 2017, 1D elastic full-waveform inversion and uncertainty estimation by means of a hybrid genetic algorithm-Gibbs sampler approach: Geophysical Prospecting, **65**, 64–85, doi: 10.1111/1365-2478.12397.

Atchadé, Y. F., 2006, An adaptive version for the metropolis adjusted Langevin algorithm with a truncated drift: Methodology and Computing in Applied Probability, **8**, 235–254, doi: 10.1007/s11009-006-8550-0.

Atchadé, Y. F., and J. S. Rosenthal, 2005, On adaptive Markov chain Monte Carlo algorithms: Bernoulli, **11**, 815–828, doi: 10.3150/bj/1130077595.

Biswas, R., and M. K. Sen, 2017, 2D full-waveform inversion and uncertainty estimation using the reversible jump Hamiltonian Monte Carlo: 87th Annual International Meeting, SEG, Expanded Abstracts, 1280–1285, doi: 10.1190/segam2017-17680416.1.

Bodin, T., and M. Sambridge, 2009, Seismic tomography with the reversible jump algorithm: Geophysical Journal International, **178**, 1411–1436, doi: 10.1111/j.1365-246X.2009.04226.x.

Brooks, S., A. Gelman, G. Jones, and X.-L. Meng, 2011, Handbook of Markov chain Monte Carlo: CRC press.

Brooks, S. P., and A. Gelman, 1998, General methods for monitoring convergence of iterative simulations: Journal of Computational and Graphical Statistics, **7**, 434–455, doi: 10.2307/1390675.

Brossier, R., S. Operto, and J. Virieux, 2009, Seismic imaging of complex onshore structures by 2D elastic frequency-domain full-waveform inversion: Geophysics, **74**, no. 6, WCC105–WCC118, doi: 10.1190/1.3215771.

Bui-Thanh, T., O. Ghattas, J. Martin, and G. Stadler, 2013, A computational framework for infinite-dimensional Bayesian inverse problems — Part 1: The linearized case, with application to global seismic inversion: SIAM Journal on Scientific Computing, **35**, A2494–A2523, doi: 10.1137/12089586X.

Bunks, C., F. M. Saleck, S. Zaleski, and G. Chavent, 1995, Multiscale seismic waveform inversion: Geophysics, **60**, 1457–1473, doi: 10.1190/1.1443880.

Chen, Z., D. Cheng, W. Feng, and T. Wu, 2013, An optimal 9-point finite difference SCHEME for the Helmholtz equation with PML: International Journal of Numerical Analysis and Modeling, **10**, 389–410.

Chib, S., and E. Greenberg, 1995, Understanding the metropolis-Hastings algorithm: The American Statistician, **49**, 327–335, doi: 10.1080/00031305.1995.10476177.

Duijndam, A., 1988, Bayesian estimation in seismic inversion — Part 1: Principles 1: Geophysical Prospecting, **36**, 878–898, doi: 10.1111/j.1365-2478.1988.tb02198.x.

Ely, G., A. Malcolm, and O. V. Poliannikov, 2018, Assessing uncertainties in velocity models and images with a fast nonlinear uncertainty quantification method: Geophysics, **83**, no. 2, R63–R75, doi: 10.1190/geo2017-0321.1.

Engquist, B., and B. D. Froese, 2013, Application of the Wasserstein metric to seismic signals: arXiv preprint arXiv:1311.4581.

Etienne, V., T. Tonellot, P. Thierry, V. Berthoumieux, and C. Andreolli, 2014, Speeding-up FWI by one order of magnitude: EAGE Workshop on High Performance Computing for Upstream, cp-426.

Fang, Z., C. Da Silva, R. Kuske, and F. J. Herrmann, 2018, Uncertainty quantification for inverse problems with weak partial-differential-equation constraints: Geophysics, **83**, no. 6, R629–R647, doi: 10.1190/geo2017-0824.1.

Fang, Z., F. J. Herrmann, and C. D. Silva, 2014, Fast uncertainty quantification for 2D full-waveform inversion with randomized source subsampling: 76th Annual International Conference and Exhibition, EAGE, Extended Abstracts, doi: 10.3997/2214-4609.20140715.

Fichtner, A., B. L. Kennett, H. Igel, and H.-P. Bunge, 2008, Theoretical background for continental-and global-scale full-waveform inversion in the time-frequency domain: Geophysical Journal International, **175**, 665–685, doi: 10.1111/j.1365-246X.2008.03923.x.

Fichtner, A., and S. Simutė, 2018, Hamiltonian Monte Carlo inversion of seismic sources in complex media: Journal of Geophysical Research, Solid Earth, **123**, 2984–2999, doi: 10.1002/2017JB015249.

Fichtner, A., and J. Trampert, 2011, Resolution analysis in full waveform inversion: Geophysical Journal International, **187**, 1604–1624, doi: 10.1111/j.1365-246X.2011.05218.x.

Fichtner, A., J. Trampert, P. Cupillard, E. Saygin, T. Taymaz, Y. Capdeville, and A. Villasenor, 2013, Multiscale full waveform inversion: Geophysical Journal International, **194**, 534–556, doi: 10.1093/gji/ggt118.

Fichtner, A., A. Zunino, and L. Gebraad, 2018, Hamiltonian Monte Carlo solution of tomographic inverse problems: Geophysical Journal International, **216**, 1344–1363, doi: 10.1093/gji/ggy496.

Gentle, J. E., 2009, Computational statistics: Springer.

Geweke, J., and H. Tanizaki, 1999, On Markov chain Monte Carlo methods for nonlinear and non-Gaussian state-space models: Communications in Statistics-Simulation and Computation, **28**, 867–894, doi: 10.1080/03610919908813583.

Geweke, J., and H. Tanizaki, 2003, Note on the sampling distribution for the Metropolis-Hastings algorithm: Communications in Statistics-Theory and Methods, **32**, 775–789, doi: 10.1081/STA-120018828.

Gilks, W. R., S. Richardson, and D. Spiegelhalter, 1995, Markov chain Monte Carlo in practice: Chapman and Hall/CRC.

Gouveia, W. P., and J. A. Scales, 1998, Bayesian seismic waveform inversion: Parameter estimation and uncertainty analysis: Journal of Geophysical Research, Solid Earth, **103**, 2759–2779, doi: 10.1029/97JB02933.

Grenander, U., and M. I. Miller, 1994, Representations of knowledge in complex systems: Journal of the Royal Statistical Society: Series B (Methodological), **56**, 549–581, doi: 10.2307/2346184.

Haario, H., E. Saksman, and J. Tamminen, 2001, An adaptive Metropolis algorithm: Bernoulli, **7**, 223–242, doi: 10.2307/3318737.

Hastings, W. K., 1970, Monte Carlo sampling methods using Markov chains and their applications: Biometrika, **57**, 97–109, doi: 10.1093/biomet/57.1.97.

Hunziker, J., E. Laloy, and N. Linde, 2019, Bayesian full-waveform tomography with application to crosshole ground penetrating radar data: Geophysical Journal International, **218**, 913–931, doi: 10.1093/gji/ggz194.

Kaipio, J., and E. Somersalo, 2006, Statistical and computational inverse problems: Springer Science and Business Media.

Keilis-Borok, V., and T. Yanovskaja, 1967, Inverse problems of seismology (structural review): Geophysical Journal International, **13**, 223–234, doi: 10.1111/j.1365-246X.1967.tb02156.x.

Lailly, P., 1983, The seismic inverse problem as a sequence of before stack migrations: Conference on Inverse Scattering, Theory and Application, Expanded Abstracts, Society for Industrial and Applied Mathematics, 206–220.

Luo, Y., and G. T. Schuster, 1991, Wave-equation traveltime inversion: Geophysics, **56**, 645–653, doi: 10.1190/1.1443081.

Martin, J., L. C. Wilcox, C. Burstedde, and O. Ghattas, 2012, A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion: SIAM Journal on Scientific Computing, **34**, A1460–A1487, doi: 10.1137/110845598.

Menke, W., 2018, Geophysical data analysis: Discrete inverse theory: Academic Press.

Métivier, L., R. Brossier, Q. Merigot, E. Oudet, and J. Virieux, 2016, An optimal transport approach for seismic tomography: Application to 3D full waveform inversion: Inverse Problems, **32**, 115008, doi: 10.1088/0266-5611/32/11/115008.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, 1953, Equation of state calculations by fast computing machines: The Journal of Chemical Physics, **21**, 1087–1092, doi: 10.1063/1.1699114.

Mosegaard, K., and M. Sambridge, 2002, Monte Carlo analysis of inverse problems: Inverse Problems, **18**, R29, doi: 10.1088/0266-5611/18/3/201.

Mosegaard, K., and A. Tarantola, 1995, Monte Carlo sampling of solutions to inverse problems: Journal of Geophysical Research, Solid Earth, **100**, 12431–12447, doi: 10.1029/94JB03097.

Mosegaard, K., and A. Tarantola, 2002, Probabilistic approach to inverse problems: International Geophysics Series, **81**, 237–268.

Neal, R. M., 2011, MCMC using Hamiltonian dynamics, *in* S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, eds., Handbook of Markov chain Monte Carlo: Chapman and Hall/CRC, 113–162.

Operto, S., A. Miniussi, R. Brossier, L. Combe, L. Métivier, V. Monteiller, A. Ribodetti, and J. Virieux, 2015, Efficient 3-D frequency-domain mono-parameter full-waveform inversion of ocean-bottom cable data: Application to Valhall in the visco-acoustic vertical transverse isotropic approximation: Geophysical Journal International, **202**, 1362–1391, doi: 10.1093/gji/ggv226.

Pestana, R. C., and P. L. Stoffa, 2010, Time evolution of the wave equation using rapid expansion method: Geophysics, **75**, no. 4, T121–T131, doi: 10.1190/1.3449091.

Petra, N., J. Martin, G. Stadler, and O. Ghattas, 2014, A computational framework for infinite-dimensional Bayesian inverse problems — Part 2: Stochastic Newton MCMC with application to ice sheet flow inverse problems: SIAM Journal on Scientific Computing, **36**, A1525–A1555, doi: 10.1137/130934805.

Plessix, R.-E., and W. Mulder, 2004, Frequency-domain finite-difference amplitude preserving migration: Geophysical Journal International, **157**, 975–987, doi: 10.1111/j.1365-246X.2004.02282.x.

Pratt, R. G., C. Shin, and G. Hick, 1998, Gauss-Newton and full Newton methods in frequency-space seismic waveform inversion: Geophysical Journal International, **133**, 341–362, doi: 10.1046/j.1365-246X.1998.00498.x.

Pratt, R. G., and R. M. Shipp, 1999, Seismic waveform inversion in the frequency domain — Part 2: Fault delineation in sediments using crosshole data: Geophysics, **64**, 902–914, doi: 10.1190/1.1444598.

Press, F., 1968, Earth models obtained by Monte Carlo inversion: Journal of Geophysical Research, **73**, 5223–5234, doi: 10.1029/JB073i016p05223.

Qi, Y., and T. P. Minka, 2002, Hessian-based Markov Chain Monte Carlo algorithms: Proceedings of the First Cape Cod Workshop on Monte Carlo Methods.

Ravaut, C., S. Operto, L. Improta, J. Virieux, A. Herrero, and P. Dell'Aversana, 2004, Multiscale imaging of complex structures from multifold wide-aperture seismic data by frequency-domain full-waveform tomography: Application to a thrust belt: Geophysical Journal International, **159**, 1032–1056, doi: 10.1111/j.1365-246X.2004.02442.x.

Ray, A., S. Kaplan, J. Washbourne, and U. Albertin, 2017, Low frequency full waveform seismic inversion within a tree based Bayesian framework: Geophysical Journal International, **212**, 522–542, doi: 10.1093/gji/ggx428.

Robert, C., and G. Casella, 2013, Monte Carlo statistical methods: Springer Science & Business Media.

Roberts, G. O., A. Gelman, and W. R. Gilks, 1997, Weak convergence and optimal scaling of random walk Metropolis algorithms: The Annals of Applied Probability, **7**, 110–120, doi: 10.1214/aoap/1034625254.

Roberts, G. O., and R. L. Tweedie, 1996, Exponential convergence of Langevin distributions and their discrete approximations: Bernoulli, **2**, 341–363, doi: 10.2307/3318418.

Sajeva, A., M. Aleardi, E. Stucchi, N. Bienati, and A. Mazzotti, 2016, Estimation of acoustic macro models using a genetic full-waveform inversion: Applications to the Marmousi model genetic FWI for acoustic macro models: Geophysics, **81**, no. 4, R173–R184, doi: 10.1190/geo2015-0198.1.

Sambridge, M., and K. Mosegaard, 2002, Monte Carlo methods in geophysical inverse problems: Reviews of Geophysics, **40**, 3-1–3-29, doi: 10.1029/2000RG000089.

Santosa, F., W. Symes, and G. Raggio, 1987, Inversion of band-limited reflection seismograms using stacking velocities as constraints: Inverse Problems, **3**, 477–499, doi: 10.1088/0266-5611/3/3/015.

Scales, J. A., M. L. Smith, and T. L. Fischer, 1992, Global optimization methods for multimodal inverse problems: Journal of Computational Physics, **103**, 258–268, doi: 10.1016/0021-9991(92)90400-S.

Sen, M. K., and R. Biswas, 2017, Transdimensional seismic inversion using the reversible jump Hamiltonian Monte Carlo algorithm: Geophysics, **82**, no. 3, R119–R134, doi: 10.1190/geo2016-0010.1.

Sen, M. K., and P. L. Stoffa, 1996, Bayesian inference, Gibbs' sampler and uncertainty estimation in geophysical inversion: Geophysical Prospecting, **44**, 313–350, doi: 10.1111/j.1365-2478.1996.tb00152.x.

Sen, M. K., and P. L. Stoffa, 2013, Global optimization methods in geophysical inversion: Cambridge University Press.

Sirgue, L., and R. G. Pratt, 2004, Efficient waveform inversion and imaging: A strategy for selecting temporal frequencies: Geophysics, **69**, 231–248, doi: 10.1190/1.1649391.

Stuart, A. M., J. Voss, and P. Wilberg, 2004, Conditional path sampling of SDEs and the Langevin MCMC method: Communications in Mathematical Sciences, **2**, 685–697, doi: 10.4310/CMS.2004.v2.n4.a7.

Stuart, G., W. Yang, S. Minkoff, and F. Pereira, 2016, A two-stage Markov chain Monte Carlo method for velocity estimation and uncertainty quantification: 86th Annual International Meeting, SEG, Expanded Abstracts, 3682–3687, doi: 10.1190/segam2016-13865449.1.

Stuart, G. K., S. E. Minkoff, and F. Pereira, 2019, A two-stage Markov chain Monte Carlo method for seismic inversion and uncertainty quantification: Geophysics, **84**, no. 6, R1003–R1020, doi: 10.1190/geo2018-0893.1.

Tape, C., Q. Liu, A. Maggi, and J. Tromp, 2010, Seismic tomography of the southern California crust based on spectral-element and adjoint methods: Geophysical Journal International, **180**, 433–462, doi: 10.1111/j.1365-246X.2009.04429.x.

Tarantola, A., 1984, Inversion of seismic reflection data in the acoustic approximation: Geophysics, **49**, 1259–1266, doi: 10.1190/1.1441754.

Tarantola, A., 2005, Inverse problem theory and methods for model parameter estimation: SIAM.

Tierney, L., 1994, Markov chains for exploring posterior distributions: The Annals of Statistics, **22**, 1701–1728, doi: 10.1214/aos/1176325750.

Ulrych, T. J., M. D. Sacchi, and A. Woodbury, 2001, A Bayes tour of inversion: A tutorial: Geophysics, **66**, 55–69, doi: 10.1190/1.1444923.

Vigh, D., K. Jiao, D. Watts, and D. Sun, 2014, Elastic full-waveform inversion application using multicomponent measurements of seismic data collection: Geophysics, **79**, no. 2, R63–R77, doi: 10.1190/geo2013-0055.1.

Virieux, J., and S. Operto, 2009, An overview of full-waveform inversion in exploration geophysics: Geophysics, **74**, no. 6, WCC1–WCC26, doi: 10.1190/1.3238367.

Warner, M., and L. Guasch, 2016, Adaptive waveform inversion: Theory: Geophysics, **81**, no. 6, R429–R445, doi: 10.1190/geo2015-0387.1.

Wu, R.-S., J. Luo, and B. Wu, 2014, Seismic envelope inversion and modulation signal model: Geophysics, **79**, no. 3, WA13–WA24, doi: 10.1190/geo2013-0294.1.

Xue, Z., N. Alger, and S. Fomel, 2016, Full-waveform inversion using smoothing kernels: 86th Annual International Meeting, SEG, Expanded Abstracts, 1358–1363, doi: 10.1190/segam2016-13948739.1.

Zhao, Z., and M. Sen, 2017, Fast double plane wave full-waveform inversion using the scattering-integral method in frequency domain: 87th Annual International Meeting, SEG, Expanded Abstracts, 1324–1329, doi: 10.1190/segam2017-17790005.1.

Zhao, Z., and M. K. Sen, 2019, A multi-scale full waveform inversion method — Staging wavenumber components and layer-stripping: 89th Annual International Meeting, SEG, Expanded Abstracts, 1470–1474, doi: 10.1190/segam2019-3216581.1.

Zhu, H., E. Bozdağ, D. Peter, and J. Tromp, 2012, Structure of the European upper mantle revealed by adjoint tomography: Nature Geoscience, **5**, 493, doi: 10.1038/ngeo1501.

Zhu, H., and S. Fomel, 2016, Building good starting models for full-waveform inversion using adaptive matching filtering misfit: Geophysics, **81**, no. 5, U61–U72, doi: 10.1190/geo2015-0596.1.

Zhu, H., S. Li, S. Fomel, G. Stadler, and O. Ghattas, 2016, A Bayesian approach to estimate uncertainty for full-waveform inversion using a priori information from depth migration: Geophysics, **81**, no. 5, R307–R323, doi: 10.1190/geo2015-0641.1.

Biographies and photographs of the authors are not available.