Elastic full waveform inversion results uncertainty analysis: a comparison between the model uncertainty given by conventional FWI and machine learning methods

Tianze Zhang, Scott Keating, Kristopher Innanen

ABSTRACT

Uncertainty analysis is an important aspect of quantifying the results of the inversion problem. In this report, we compare the uncertainty analysis given by two methods for full waveform inversion. The first method is by using the approximation of the inverse Hessian to perform the uncertainty analysis, as the approximation of the inverse Hessian is closely related to the posterior model covariance matrix. The second method is based on a machine learning-based method, which uses the Bayesian neural network (BNN) to generate elastic models and then performs the inversion. In the BNN, each trainable weight is represented as a Gaussian distributed probability distribution function (pdf). When BNN is well trained, we can forward calculate the BNN several times and perform the statistic analysis for the prediction results and give the uncertainty analysis for the generated models. Our numerical results suggested that both methods can generate promising inversion results and reasonable uncertainty quantification when compared with the true model errors.

INTRODUCTION

Full-waveform inversion (FWI) addresses the geophysical inverse problem of estimating subsurface model parameters from observed waveform data. In most geophysical applications, FWI is introduced as an iterative, local optimization problem that attempts to minimize the least-squares residuals between observed and synthetic data. Mathematically, the inverse problem is ill-posed, leading to a non-uniqueness of the solutions. It remains challenging to solve inverse problems piratically due to limitations in data acquisition, measurement uncertainties and the non-uniqueness of the solution (Tarantola, 1984; Lailly, 1983).

Estimations of the resolution or uncertainty in seismic inversions have a long history in geophysics and can be analyzed with mathematical tools such as the posterior covariance matrix. The posterior covariance matrix is closely related to the inverse Hessian (Fichtner and Trampert, 2011; Zhu et al., 2016). However, for practical problems with millions of parameters, it is unfeasible to store such vast matrices. With least-squares QR factorization Zhang and McMechan (1995) modify classic inversion algorithms with least-squares QR factorization to handle large-scale inverse problems. The spatial resolution lengths with a Gaussian approximation to the resolution matrix were introduced by An (2012). Trampert et al. (2013) sample the tomographic models for resolution lengths with random probing and analyze the direction-dependent resolution lengths of waveform tomography by Fichtner and Leeuwen (2015) autocorrelating the randomly sampled Hessian. Rawlinson et al. (2014) also gives a detailed explanation for the uncertainty estimation for the seismic inversion problem.

Randomized singular-value decomposition (SVD) also attracted attention to geophysi-

cists, Halko et al. (2011), with the development of matrix probing theories in applied mathematics. The Bayesian inference workflow for waveform tomography is formulated by Bui-Thanh et al. (2013) by deriving an approximation to the posterior covariance matrix by decomposing the data-misfit Hessian into eigenvalues and eigenvectors with randomized SVD. Zhu et al. (2016) improve the efficiency of the Hessian computation by exploiting point-spread function (PSF) tests. More recent ensemble-based approaches have been introduced to tomography problems by Jordan (2015) with the utilization of the Kalman Filter (KF) theory (Kalman, 1960; Evensen, 1994). Liu and Peter (2019) used the Square-root variable metric-based elastic FWI to quantify the uncertainties for V_p and V_s. The theory part of this report is referred to the paper of Liu and Peter (2019) to provide a recent review of uncertainty estimations.

This report will first give a brief review of how we can use the inverse Hessian approximation to give the uncertainty quantification for the two parameters of elastic FWI. As the inverse Hessian approximation in this paper is calculated by the Quasi-Newton method, which requires starting point given manually, we discuss how this initial guessing of the inverse Hessian could influence the final uncertainty quantification. Next, we briefly introduce the BNN-based EIFWI, which uses the BNN to generate elastic models. Uncertainty analysis can be given by forward calculating the well-trained neural network several times can perform the statistical analysis of the generated elastic models. Then, we will compare the uncertainty quantification given by these two methods.

REVIEW OF THE BAYESIAN INFERENCE

Bayesian inference of FWI

In the FWI workflows, Bayesian inference allows us to incorporate the prior information into waveform tomography to estimate posterior uncertainties of the inverted results. To review, we start with a short glance at the forward problem. For a forward modeling problem we have:

$$\mathbf{d} = \mathbf{g}(\mathbf{m}). \tag{1}$$

$\mathbf{d}$ is the observed data calculated with physics model $\mathbf{m}$, via the operator $\mathbf{g}$. $\mathbf{g}$ is usually nonlinear. The forward problem gives what should be observed for a particular model, whereas the inverse problem calculates the particular physics models for a set of observations. Equation 1 relates the physics model and the data observations. In this paper, the operator $\mathbf{g}$ represents the numerical solution of the isotropic elastic wave equation based on the finite difference method using stress-velocity format (Virieux, 1986).

We first make the assumption that the prior model probability distribution function (PDF) is Gaussian distributed (Gauss, 1877). Thus the prior model probability distribution function (PDF) can be expressed as:

$$\rho(\mathbf{m}) \propto \exp(\mathbf{m} - \mathbf{m}_{prior})^{\mathrm{T}} \mathbf{C}_m^{\mathrm{T}} (\mathbf{m} - \mathbf{m}_{prior}), \tag{2}$$

where $\mathbf{m}_{prior}$ is the prior model, with its mean being the (initial) model $\mathbf{m}_0$, and $\mathbf{C}_m$ is the prior model covariance matrix. Due to the central limit theorem, the choice of Gaussian priors is very difficult to avoid.

In Bayesian inference, the inherent discrepancy between the observed and synthetic data is also considered to be distributed as Gaussian noise. We represent the likelihood function of the data as

$$\rho(\mathbf{d}|\mathbf{m}) \propto \exp\left(-\frac{1}{2}(\mathbf{d}-\mathbf{g}(\mathbf{m}))^{\mathrm{T}}\mathbf{C}_d^{-1}(\mathbf{d}-\mathbf{g}(\mathbf{m}))\right). \tag{3}$$

$\mathbf{C}_d$ is the data covariance matrix indicating the data uncertainties. According to the Bayesian inference, the solution to an inverse problem yields the PDF, which is defined as:

$$\rho(\mathbf{m}|\mathbf{d}) \propto \rho(\mathbf{d}|\mathbf{m})\rho(\mathbf{m}), \tag{4}$$

which the term $\rho(\mathbf{m}|\mathbf{d})$ indicates the probability of $\mathbf{m}$ when the data $\mathbf{d}$ is observed, indicating the posterior probability. $\rho(\mathbf{d}|\mathbf{m})$ is the likelihood, and $\rho(\mathbf{m})$ is the prior probability. $\rho(\mathbf{m})$ is related to the model misfit, and the $\rho(\mathbf{d}|\mathbf{m})$ is related to the data misfit. Combining equation 1 and 2 we have the posterior distribution :

$$\rho(\mathbf{m}|\mathbf{d}) \propto \exp(-\mathrm{f}(\mathbf{m})), \tag{5}$$

where the $f(\mathbf{m})$ is the objective function in the least-square sense, namely

$$f(\mathbf{m}) = \frac{1}{2}(\mathbf{d}-\mathbf{g}(\mathbf{m}))^{\mathrm{T}}\mathbf{C}_d^{-1}(\mathbf{d}-\mathbf{g}(\mathbf{m})) + (\mathbf{m}-\mathbf{m}_{prior})^{\mathrm{T}}\mathbf{C}_m^{-1}(\mathbf{m}-\mathbf{m}_{prior}) \tag{6}$$

Note that even with the data error and the model prior are Gaussian distributed, the posterior distribution could not be Gaussian distributed, due to that the forward modeling $\mathbf{g}(\mathbf{m})$ is non-linear in the likelihood function of equation 3. One simple approach is to linearize the forward modeling operator around the maximum posterior (MAP) model (Gouveia and Scales, 1998; Petra et al., 2014). This indicates that the posterior PDF as:

$$\rho(\mathbf{m}|\mathbf{d}) \propto \exp\left(-\frac{1}{2}(\mathbf{m}-\bar{\mathbf{m}})^{\mathrm{T}}\mathbf{C}_m^{-1}(\mathbf{m}-\bar{\mathbf{m}})\right) \tag{7}$$

with the MAP model being:

$$\bar{\mathbf{m}} = \mathbf{m}_{prior} + (\mathbf{G}^{\mathrm{T}}\mathbf{C}_d^{-1}\mathbf{G} + \mathbf{C}_m^{-1})^{-1}\mathbf{G}^{T}\mathbf{C}_d^{-1}(\mathbf{d}-\mathbf{G}\mathbf{m}_{prior}) \tag{8}$$

and the posterior covariance being:

$$\mathbf{C}_m = (\mathbf{G}^{\mathrm{T}}\mathbf{C}_d^{-1}\mathbf{G} + \mathbf{C}_m^{-1})^{-1} = (\mathbf{H}_d + \mathbf{C}_m^{-1})^{-1} \tag{9}$$

The term $\mathbf{G}$ in equation 8 and 9 are the Fréchet derivatives: $\mathbf{G} = \frac{\partial \mathbf{g}(\mathbf{m})}{\partial \mathbf{m}}$. In equation 9, we also have $\mathbf{H}_d = \mathbf{G}^{\mathrm{T}}\mathbf{C}_d^{-1}\mathbf{G}$, which is known as the Gaussian Newton approximation method to the data-misfit Hessian (Pratt, 1999; Virieux and Operto, 2009). The direct calculation of the Fréchet derivatives is too expensive to achieve. An efficient way of calculating the Fréchet derivatives is to calculate the gradient by using the adjoint state method, (Tarantola, 1984, 1986; Tromp et al., 2005), which uses the zero-lag correlation between the forward modeling waveform and the backpropagation wavefield to calculate the Fréchet derivatives. The isotropic elastic wave equation is self-adjoint, meaning that the forward modeling operator and the backpropagation operator for solving the forward modeling wavefield and the backpropagation wavefields can be the same. The covariance matrix of $\mathbf{C}_m$ reflects the key information of the uncertainty information of the inverted models. According to equation 8, and 9, the effective quantification of the uncertainty by using the Bayesian inference relies on the efficient way of the estimation of the Hessian $\mathbf{H}_d$.

Inverse Hessian approximation

In the Newtown's method, the model update $\delta\mathbf{m}$ in each iteration is related to the gradient $\mathbf{g}$ through the inverse of the Hessian $\mathbf{H}^{-1}$ by:

$$\delta\mathbf{m} = -\mathbf{H}^{-1}\mathbf{g} \approx -\mathbf{B}\mathbf{g}, \tag{10}$$

where the matrix $\mathbf{B}$ is the approximation of the inverse Hessian: $\mathbf{H}^{-1} \approx \mathbf{B}$. Gauss-Newton method is one of the most popular optimization methods to perform full-waveform inversion (Pan et al., 2016; Chen and Sacchi, 2020). In this study, we are particularly interested in the Quasi-Newton method to calculate the inverse Hessian approximation, which is a relatively less computationally intensive way of explicitly evaluating the inverse Hessian approximation.

The standard Davidon–Fletcher–Powell (DFP) method (Fletcher and Powell, 1963) gives an iterative approach to update the inverse Hessian:

$$\mathbf{B}_{k+1} = \mathbf{B}_k - \frac{\mathbf{B}_k\mathbf{y}_k\mathbf{y}_k^T\mathbf{B}_k}{\mathbf{y}_k^T\mathbf{B}_k\mathbf{y}_k} + \frac{\mathbf{s}_k\mathbf{s}_k^T}{\mathbf{y}_k^T\mathbf{y}_k} \tag{11}$$

Here the $\mathbf{B}_k$ is the inverse Hessian approximation at k iteration. If $k = 0$, $\mathbf{B}_0$ is the initial guessing of the inverse approximation Hessian. $\mathbf{s}_k = \mathbf{m}_{k+1} - \mathbf{m}_k$, which is the model update vector. $\mathbf{y}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$, which is the gradient update vector. A stable and practical algorithmic form modified from the DFP method is the vector-version SRVM, which is a positive definite approximation to collect the information about the inverse Hessian over all the n iterations Tarantola (2005):

$$\mathbf{B}_{n+1} = \mathbf{B}_0 + \sum_{k=0}^{n}\left(\frac{\mathbf{s}_k\Delta\mathbf{s}_k^T}{\mathbf{y}_k^T\mathbf{s}_k} - \frac{\mathbf{B}_k\mathbf{y}_k\Delta\mathbf{y}_k^T\mathbf{B}_k}{\mathbf{g}_k^T\mathbf{B}_k\mathbf{y}_k}\right) \tag{12}$$

The Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm is an iterative method for solving unconstrained nonlinear optimization problems (Fletcher, 2013). Like the related DFP, BFGS determines the descent direction by preconditioning the gradient with curvature information. It does so by gradually improving an approximation to the Hessian matrix of the loss function, obtained only from gradient evaluations (or approximate gradient evaluations) via a generalized secant method. The inverse Hessian approximation given by using the BFGS approximation is formulated as follows:

$$\mathbf{B}_{k+1} = \left(\mathbf{I} - \frac{\mathbf{s}_k\mathbf{y}_k^T}{\mathbf{y}_k^T\mathbf{s}_k}\right)\mathbf{B}_k\left(\mathbf{I} - \frac{\mathbf{y}_k\mathbf{s}_k^T}{\mathbf{y}_k^T\mathbf{s}_k}\right) + \frac{\mathbf{s}_k\mathbf{s}_k^T}{\mathbf{y}_k^T\mathbf{s}_k}, \tag{13}$$

where $\mathbf{I}$ is the identity matrix. In the following articles, we will both use the SRVM method and the BFGS method to perform the inverse Hessian approximation in numerical testing.

Randomized SVD for the approximation inverse Hessian

Equation 11, 12, 13 can all be considered as the low-rank approximation of the inverse Hessian. Randomized SVD, (Liberty et al., 2007; Halko et al., 2011), provides an efficient way to factorize large matrices into their corresponding eigenvalues and eigenvectors, especially for matrices with low-rank property.

Given a symmetric $\mathrm{M} \times \mathrm{M}$ matrix $\mathbf{Z}$ and a set of $\mathrm{M} \times \mathrm{N_r}$ random vectors $\mathbf{X}$, where M is the model size and $\mathrm{N_r}$ estimated rank order of matrix $\mathbf{Z}$. As shown in algorithm 1, the single-pass randomized SVD algorithm processes the matrix $\mathbf{Z}$ as follows:

Algorithm 1 Single-pass randomised SVD algorithm
1: $\mathbf{Y} = \mathbf{Z}\mathbf{X}$
2: $\mathbf{Q}\mathbf{R} = \mathbf{Y}$
3: $\mathbf{A}\left(\mathbf{Q}^T\mathbf{X}\right) = \mathbf{Q}^T\mathbf{Y}$
4: $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T = \mathbf{A}$
5: $\mathbf{V} = \mathbf{Q}\mathbf{U}$
6: $\mathbf{Z} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$

With the use of the randomized SVD algorithm, the inverse Hessian can be decomposed as:

$$\mathbf{H}^{-1} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T, \tag{14}$$

where the $\mathbf{V}$ is the eigenvector matrix vector with the dimension of $\mathrm{M} \times \mathrm{N_r}$, and $\mathbf{\Lambda}$ being the eigenvalues matrix with $\mathrm{N_r}$ diagonal elements.

Practical assessment of the posterior covariance

The equation 6 provides an elegant perspective on the objective function, connecting the least-squares inversion with the Bayesian inference. However, for practice inversion method, equation 6 is not always explicitly used, and it can be defined as:

$$\begin{aligned} f(\mathbf{m}) =& \frac{1}{2}(\mathbf{d} - \mathbf{g}(\mathbf{m}))^{\mathrm{T}}\mathbf{C}_d^{-1}(\mathbf{d} - \mathbf{g}(\mathbf{m})) + \\ & \epsilon\delta\mathbf{m}^{\mathrm{T}}\mathbf{C}_m^{-1}\delta\mathbf{m} + \eta\delta\mathbf{m}^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathbf{D}\delta\mathbf{m}, \end{aligned} \tag{15}$$

where the $\delta\mathbf{m} = \mathbf{m} - \mathbf{m}_0$ is the model perturbation. The $\mathbf{C}_d$ is the data covariance matrix and the $\mathbf{C}_m$ is the prior model covariance matrix, with $\mathbf{m}_0$ being the prior mean model for a Gaussian distributed prior $\mathbf{m}_{prior}$. $\mathbf{D}$ is the smoothing operation. With $\partial f(\mathbf{m})/\partial\mathbf{m} = 0$ the above equation becomes:

$$\bar{\mathbf{m}} = \mathbf{m}_0 + (\mathbf{G}^T\mathbf{C}_d^{-1}\mathbf{G} + \epsilon\mathbf{C}_d^{-1})^{-1} + \eta\delta\mathbf{D}^T\mathbf{D})^{-1}\mathbf{G}^T\mathbf{C}_d^{-1}(\mathbf{d} - \mathbf{G}\mathbf{m}_0), \tag{16}$$

In which $\bar{\mathbf{m}}$ is the inverted model (also known as the MAP model), and $\mathbf{G}$ is the Fréchet derivative, and the $\mathbf{D}$ is acting as the smoothing operator. So, we can rewrite the practical form of the posterior covariance matrix as follows:

$$\mathbf{C}_M = (\mathbf{G}^T\mathbf{C}_d^{-1}\mathbf{G} + \epsilon\mathbf{C}_m^{-1})^{-1} = (\mathbf{H}_d + \epsilon\mathbf{C}_m^{-1})^{-1} \tag{17}$$

The above equation differs from equation 9 by adding a pre-factor ϵ. According to Liu and Peter (2019), after rewriting equation 12, the model covariance matrix can be approximated as:

$$\mathbf{C}_M \approx \mathbf{H}^{-1}\mathbf{C}_m \tag{18}$$

We will also use this equation to evaluate the uncertainty of the inversion results for the elastic FWI.

After obtaining a low-rank SVD approximation to the posterior covariance $\mathbf{C}_m$, we can draw and compare a Gaussian random sampling on $\mathbf{C}_m$, and $\mathbf{C}_M$, and the sampling of the prior and the posterior distribution can be expressed, respectively as

$$\mathbf{m}_{prior} = \mathbf{m}_0 + \mathbf{C}_m^{1/2}\mathbf{n}, \tag{19}$$

$$\mathbf{m}_{post} = \bar{\mathbf{m}} + \mathbf{C}_M^{1/2}\mathbf{n}, \tag{20}$$

$\mathbf{n}$ being the 2-D Gaussian random sampler of zero mean and unit variance

$$\mathbf{C}_M^{1/2} = \mathbf{H}^{-1/2}\mathbf{C}_m^{1/2} = \mathbf{V}\Lambda^{-1/2}\mathbf{V}^T\mathbf{C}_m^{1/2}, \tag{21}$$

where $\mathbf{V}$ and Λ are the eigenvectors and eigenvalues of $\mathbf{H}^{-1}$, respectively. This way, we can assess the prior and posterior model uncertainties through visual comparisons of the random samplings on the prior and posterior distributions.

Isotropic elastic FWI elastic gradient update

The gradient for the FWI parameters can be calculated by using the zero-lag correlation between the observed data and the synthetic data (Tarantola, 1984; Köhn et al., 2012). For the elastic isotropic FWI, parameterized by λ,μ, and ρ, the gradients for these three parameters can be calculated with the flowing equation:

$$\begin{aligned}
\frac{\partial E}{\partial \lambda} &= -\sum_{\text{sources}} \int \mathrm{d}t \left(\frac{\partial u_x}{\partial x} + \frac{\partial u_z}{\partial z}\right)\left(\frac{\partial \Psi_x}{\partial x} + \frac{\partial \Psi_z}{\partial z}\right), \\
\frac{\partial E}{\partial \mu} &= -\sum_{\text{sources}} \int \mathrm{d}t \left(\frac{\partial u_x}{\partial z} + \frac{\partial u_z}{\partial x}\right)\left(\frac{\partial \Psi_x}{\partial z} + \frac{\partial \Psi_z}{\partial x}\right) + 2\left(\frac{\partial u_x}{\partial x}\frac{\partial \Psi_x}{\partial x} + \frac{\partial u_z}{\partial z}\frac{\partial \Psi_z}{\partial z}\right), \\
\frac{\partial E}{\partial \rho} &= -\sum_{\text{sources}} \int \mathrm{d}t \left(\frac{\partial^2 u_x}{\partial t^2}\Psi_x + \frac{\partial^2 u_z}{\partial t^2}\Psi_z\right).
\end{aligned} \tag{22}$$

While the term $\mathbf{u}$ denotes the forward-modeled wavefield for the actual model parameters, the wavefield Ψ is generated by propagating the residual data. The gradients for V_p, V_s , and ρ can be written as:

$$\begin{aligned}
\left(\frac{\partial E}{\partial V_p}\right) &= 2\rho V_p \left(\frac{\partial E}{\partial \lambda}\right) \\
\left(\frac{\partial E}{\partial V_s}\right) &= -4\rho V_s \left(\frac{\partial E}{\partial \lambda}\right) + 2\rho V_s \left(\frac{\partial E}{\partial \mu}\right) \\
\left(\frac{\partial E}{\partial \rho_{\text{vel}}}\right) &= \left(V_p^2 - 2V_s^2\right)\left(\frac{\partial E}{\partial \lambda}\right) + V_s^2 \left(\frac{\partial E}{\partial \mu}\right) + \left(\frac{\partial E}{\partial \rho}\right)
\end{aligned} \tag{23}$$

BNN-based Implicit elastic FWI

In this section, we use a probabilistic interpretation of neural network learning by implementing a coordinate-based Bayesian neural network (BNN), denoted as $N_{\text{bnn}}(\mathcal{C};\theta)$. In

each BNN layer, denoted as b_l, where l is the index for the number of layers in the BNN, for simplicity in this study, we assume each element of the weight and bias vectors to follow the Gaussian distribution, and be characterized with the corresponding mean value and the standard deviations. During the forward calculation in the neural network, realizations of the weights and biases are drawn randomly from their current probability distributions. For example, assume that w_i, and the bias b_i are the i^{th} values in weight vector $\mathbf{w}_l$ and bias vector $\mathbf{b}_l$ in the l^{th} layer. w_i is defined as:

$$w_i = w_{\mu_i} + \log(1 + e^{\delta_{w_i}})w_{\epsilon_i}, \tag{24}$$

and b_i is defined as:

$$b_i = b_{\mu_i} + \log(1 + e^{\delta_{b_i}})b_{\epsilon_i}, \tag{25}$$

where w_{μ_i} and b_{μ_i} are the mean values of w_i and b_i probability distribution respectively, and w_{ϵ_i} and b_{ϵ_i} are a random values generated with the Gaussian distribution, i.e., $w_{\epsilon_i} \sim N(0,1)$, $b_{\epsilon_i} \sim N(0,1)$. w_{μ_i} and b_{μ_i} are the values which determine the mean of the probability distribution for parameter w_i and b_i respectively. δ_{w_i} and δ_{b_i} influence the standard deviation for the parameter w_i and b_i. Each value in the weight vector $\mathbf{w}$ and bias vector $\mathbf{b}$ is obtained through the above process. We regard the w_{μ_i}, δ_{w_i} , b_{μ_i} and δ_{b_i} as the training parameters, since these values directly determine the shape of the Gaussian distribution of the parameters. For simplicity, the training parameters are all noted as θ. After all the training parameters in each layer are realized, the output of the BNN is obtained with the following equation:

$$N_{\mathrm{bnn}}(\mathbf{c};\theta) = (b_n \circ b_{n-1} \circ \cdots \circ b_1)\mathrm{m}_{\mathrm{std}} + \mathrm{m}_{\mathrm{mean}}. \tag{26}$$

The training of the BNN is based on the methodology of the variational inference. In variational inference, we seek to find the parameters θ of a distribution on the weights $p(\mathbf{w}|\theta)$ that minimizes the Kullback-Leibler (KL) divergence with the true posterior $p(\mathbf{w}|D)$ on the weights.

The BNN implicit FWI formulation involves the following training process. First, we obtain the weights and bias as described above. Second. we use the weights to calculate the outputs of the network, which are the elastic property models V_p,V_s, and ρ. Third, use the same recurrent neural network (RNN) as discussed previously to carry out forward modeling via $F(\hat{\mathbf{u}}, N_{\mathrm{bnn}}(\mathbf{c};\theta)) = \mathbf{f}$ and obtain the synthetic data $\hat{\mathbf{d}}_{\mathrm{syn}}$. Fourth, we calculate the $\log p(w_i|\theta)$, $\log p(w_i)$, and $\log p(D_j|\mathbf{w})$. For the FWI problem, the term $\log p(D_j|\mathbf{w})$ can be denoted as $\log p(d^{\mathrm{obs}}_{s,t,r}|\mathbf{w})$, where the $d^{\mathrm{obs}}_{s,t,r}$ represents the observed data at time t and receiver r for one shot index s. As the weights $\mathbf{w}$ directly determines the geophysics parameters $\mathbf{m}$, and the synthetic data vector $d^{\mathrm{syn}}_{s,t,r}$ is also calculated according to the wave equation with $\mathbf{m}$, we can conclude that $p(d^{\mathrm{obs}}_{s,t,r}|\mathbf{w}) \propto p(d^{\mathrm{obs}}_{s,t,r}|d^{\mathrm{syn}}_{s,t,r})$. Thus, for the implicit FWI problem, we can use $\log p(d^{\mathrm{obs}}_{s,r,t}|d^{\mathrm{syn}}_{s,r,t})$ to replace $\log p(d^{\mathrm{obs}}_{s,r,t}|\mathbf{w})$. Thus, the loss function we use for BNN IFWI is defined as:

$$\begin{aligned}\phi(\theta) \propto \phi(\theta)_{\mathrm{EIFWI}} =& \lambda \sum_{i=1}^{n} [\log p(w_i|\theta) - \log p(w_i)] - \\ & \sum_{s=1}^{N_s}\sum_{t=1}^{N_t}\sum_{r=1}^{N_r} \log p(d^{\mathrm{obs}}_{s,r,t}|d^{\mathrm{syn}}_{s,r,t}),\end{aligned} \tag{27}$$

where the N_s, N_t, and N_r are the maximum number of the shots, receiving time and the receivers. Fifth, we use a gradient-based method to update the weights: $\theta = \theta - \alpha \triangle_\theta \phi_{\text{EIFWI}}$, where the t is the iteration time, and $\triangle_\theta \phi_{\text{EIFWI}}$ is the partial derivative of the loss with respect to the parameter θ. λ is the value that controls the contribution of the $(\log p(w|\theta) - \log p(w))$ term, the network complexity term, in the loss function. α is the step length for the gradient based optimization method. The calculate of the gradients are also implicated with the automatic differential method.

NUMERICAL TEST

In this section, we will use part of the full Marmousi model as the V_p and V_s models to perform the inversion. The size of the model is 100×200. The grid length we use here is $dx = dz = 20$. The inversions are all carried out in the time domain. We assume that all the sources are well known, and we use Ricker's wavelet as the source, with the main frequency 10Hz. All the sources and receivers are located on the model's surface, with a shot interval of 600m and a receiver interval of 20m. We use the Wolfe condition to calculate the step (Nocedal and Wright, 1999).

Figure 1 (a) and (b) are the true V_p, V_s models respectively. Figure 1 (c) and (d) are inversion results for V_p and V_s after 500 iterations after using the BFGS method. We can see that the inversion is successful. Most of the structures of the elastic model have been correctly updated, especially for the anomaly located in the center of the models. The deeper part of the model is less updated, and due to that, we have limited acquisition illumination. If the acquisition system is cross-well, which means that we have shots on one side of the model and receivers on the other side of the model, we could have better illumination for the deeper part of the model. Figures 1 (e) and (f) illustrate the uncertainty analysis of the FWI for the V_p, and V_s. Figures 1 (e) and (f) are the standard deviation of the posterior sampling for parameter V_p, and V_s using equation 21. Figure 1 (g) and (h) are the absolute model errors for the V_p, V_s respectively, which is the difference between the true models and MAP models. Our assumption for a successful uncertainty quantification is that the uncertainty should match well with the absolute model error. From the comparison between the third and the fourth rows, we can see that the errors of the center anomaly align well with the standard deviation uncertainty indicating the correct quantification of the uncertainty for the shallower part of the elastic model. However, the model errors for the deeper parts of the model are poorly reflected in the standard deviation plots. This may be because the inverse approximation Hessian is constructed with the gradients and the model updates of the FWI. If the model updates and the gradients have small updates in the deeper part of the model, then the uncertainty quantification according to such a Hessian also has few information uncertainty updates for the deeper part of the model. A reasonable guess for the uncertainty for the deeper part of the model should be larger than the shallower part.

To improve the uncertainty quantification of the uncertainty analysis for FWI, we formulate the diagonal of the initial Hessian by using the following formula:

$$\mathbf{B}^p_{diag}[ix, z] = \exp\left[-\frac{1}{2}\left(a\frac{z - nz - \Delta l}{\Delta l/2}\right)^2\right], z \in (0, nz) \tag{28}$$

$\mathbf{B}^p_{diag}[ix, z]$ represent one column of the guessing of the initial Hessian for parameter $p \in$

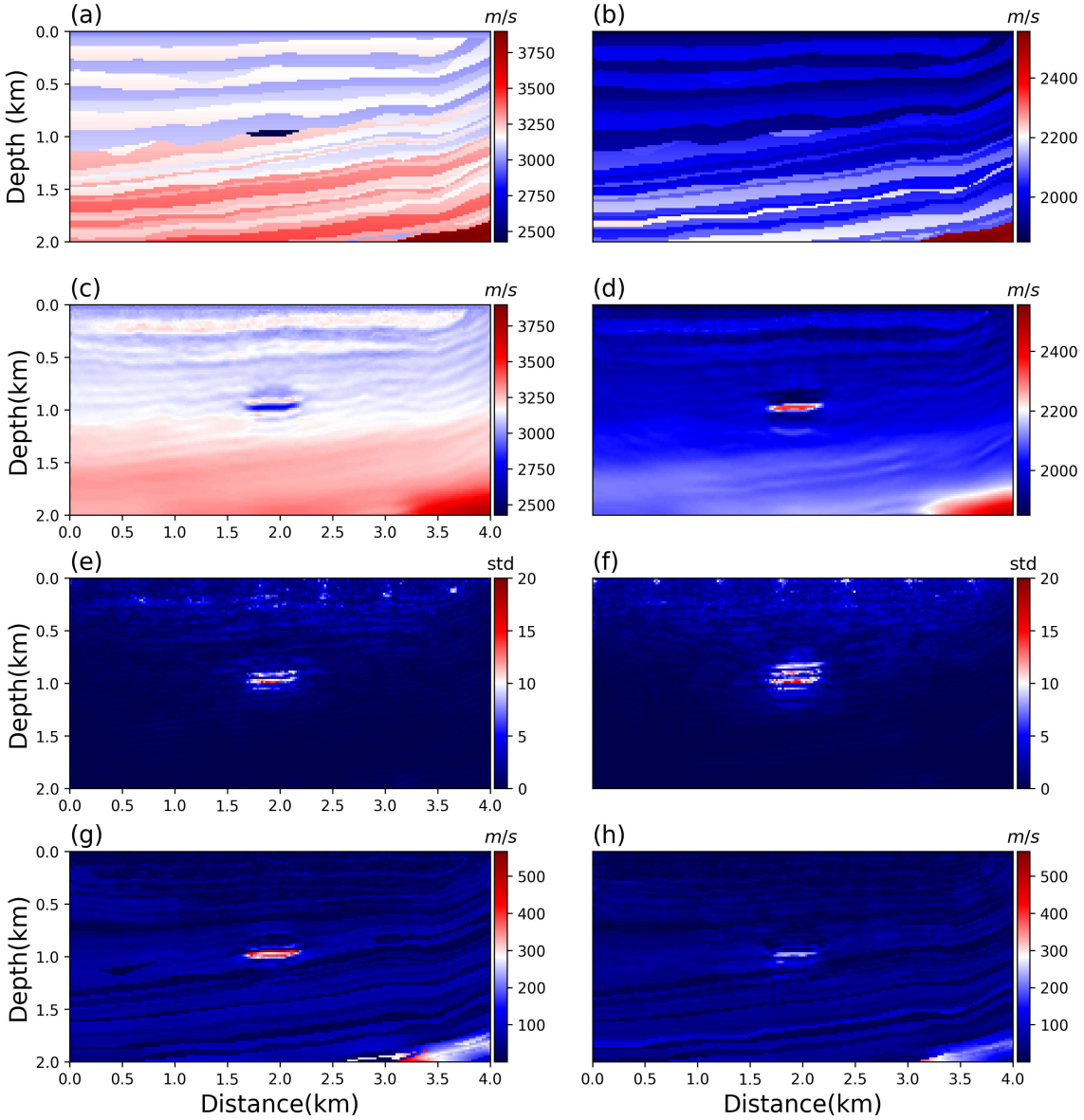

FIG. 1. FWI uncertainty analysis with $\mathbf{B}_0 = \mathbf{I}$. (a) true V_p. (b) true V_s. (c) V_p inversion result. (d) V_s inversion result (e) V_p standard deviation. (f) V_s standard deviation. (e) Absolute V_p model error. (f) Absolute V_s model error.

(V_p, V_s). After calculating each column of the $\mathbf{B}^p_{diag}$, we need to diagonal the $\mathbf{B}^p_{diag}$ and put it into the main diagonal term of the initial Hessian $\mathbf{B}_{k0}$ for the corresponding p parameter. a is the hyper-parameter of this method. In this test, we choose it to be $a = 12$ and $\Delta l = nx$. Figure 2 shows the inversion results and the uncertainty quantification by using this initial guessing of the Hessian. Figure 2 (c) and (d) are inversion results by utilizing the Hessian with the initial guessing with equation 28. With the same inversion strategy but with different initial Hessian Guessing, we can see that the deeper part of the models is better recovered, while the shallower part of the models remains relatively the same. Figure 2 (e) and (f) shows the standard deviation of the V_p, and V_s uncertainty quantification using the initial Hessian suggested. We observe the better characterization of the uncertainties

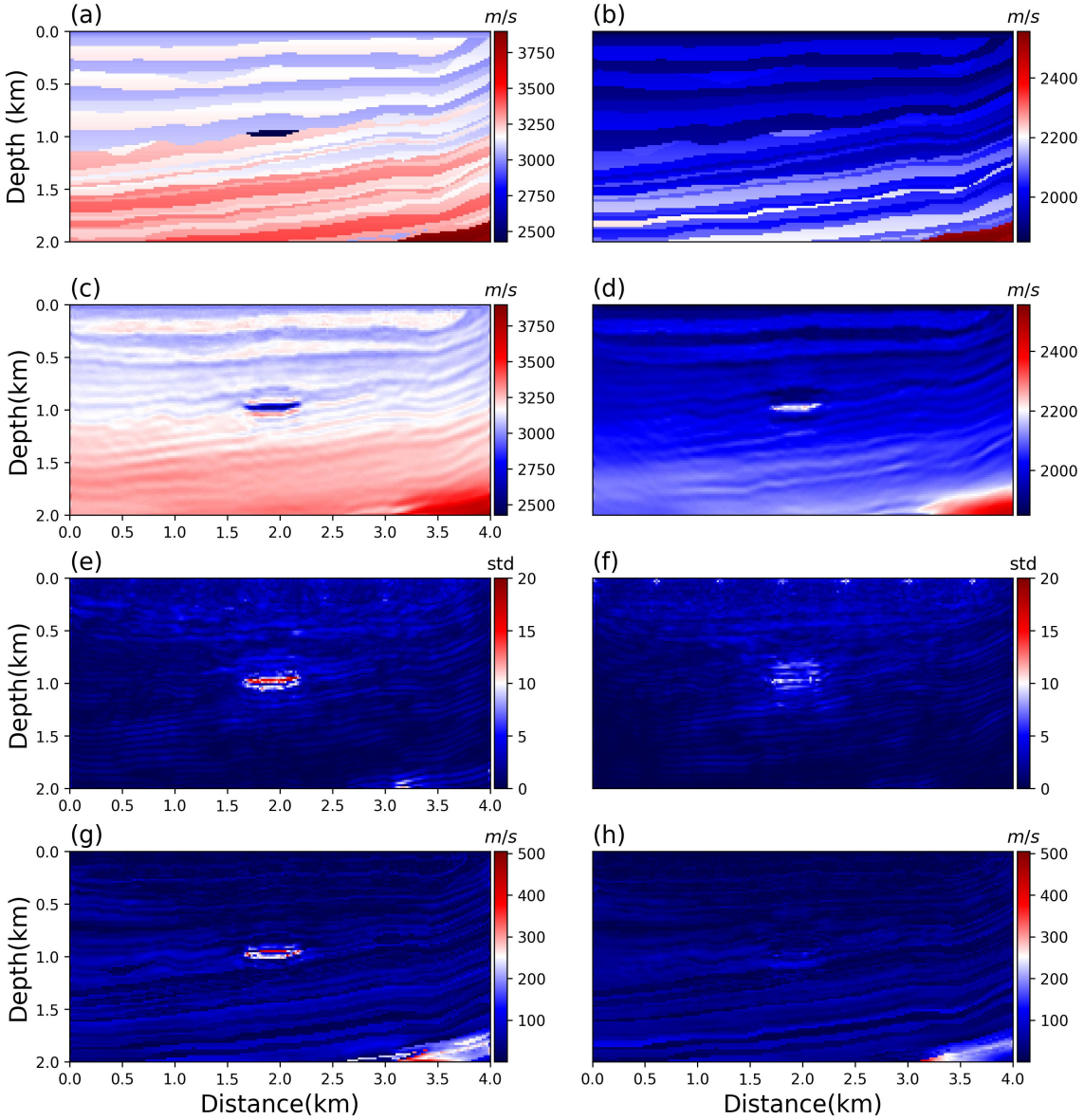

FIG. 2. FWI uncertainty analysis with $\mathbf{B}_0$ formulated with equation 28. (a) true V_p. (b) true V_s. (c) V_p inversion result. (d) V_s inversion result (e) V_p standard deviation. (f) V_s standard deviation. (e) Absolute V_p model error. (f) Absolute V_s model error.

for the high-velocity structure on the right corner of the V_p model. Also, the acquisition footprints in Figure 1 are also mitigated. The deeper part of the uncertainties for the V_s model is also better revealed.

BNN FWI inversion results

We perform the forward calculation using the well-trained BNN 1000 times and obtain a set of the 1000 models for all the V_p, V_s, and ρ models. As the weights are drawn from the well-trained posterior probability distribution function, each forward calculation gives relatively different velocity models. Then we calculate the mean and the standard deviation

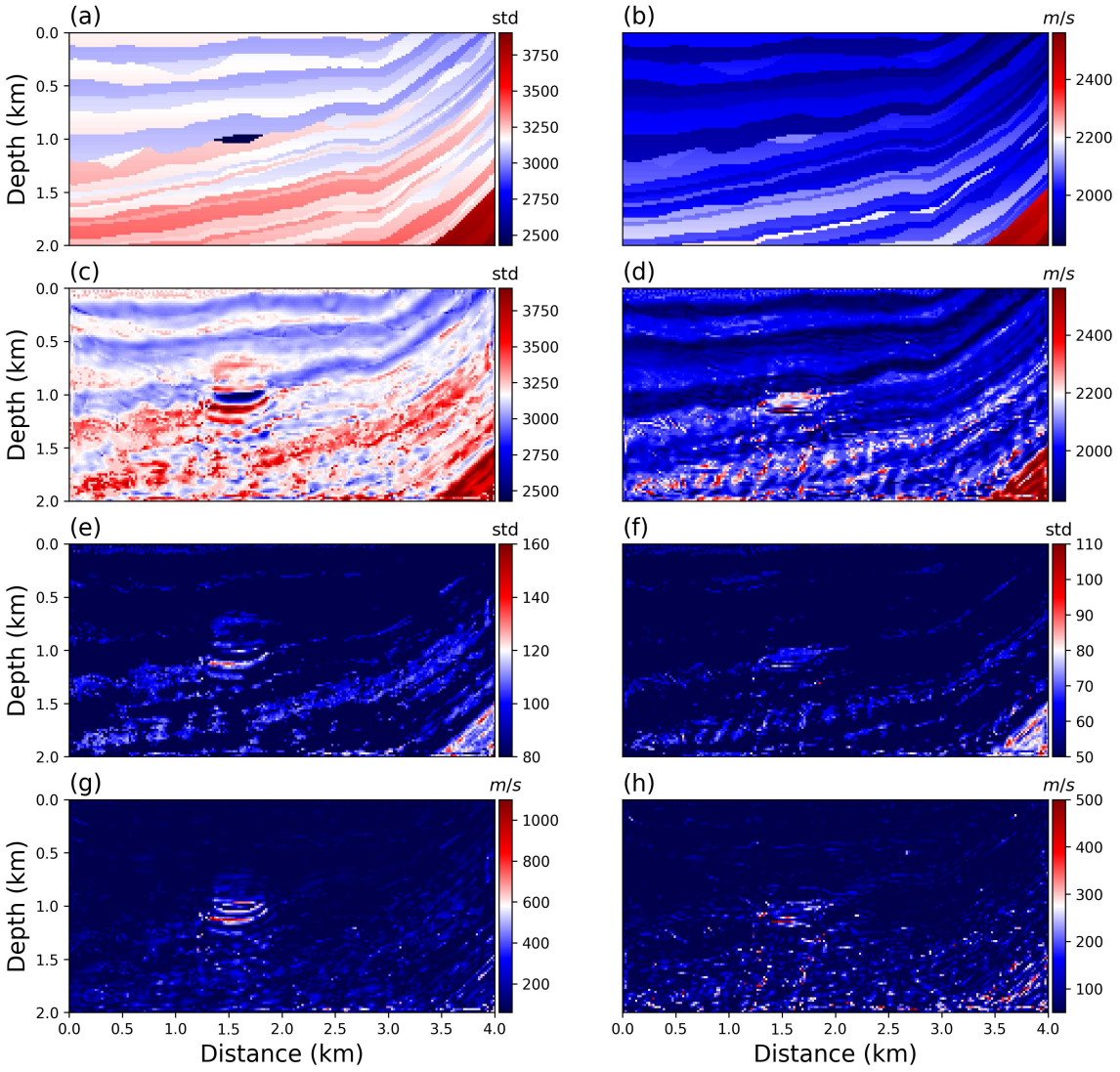

FIG. 3. FWI uncertainty analysis with incline Guessing. (a) true V_p. (b) true V_s. (c) V_p BNN FWI result. (d) V_s BNN result (e) V_p BNN FWI standard deviation. (f) V_s BNN FWI standard deviation. (e) Absolute V_p model error. (f) Absolute V_s model error.

of the velocity models to give the uncertainty analysis of these velocities generated with BNN. If the velocity model agrees with each other at a certain location of the model, the uncertainty for this point is low (and vice-versa). We will compare the standard deviation of the prediction sets for the elastic models with the absolute model error. If the standard deviations match well with the model error, then we consider that this is a valid uncertainty quantification. The third row of Figure 3 illustrates the standard deviation, and the last row demonstrates the absolute model uncertainty. We can see that the standard deviations estimated by the BNN match well with the absolute model error. We can see that most of the prediction errors are positioned in the center anomaly of the model and the deeper layers, and these model errors are all correctly reflected on the standard deviation plots. For instance, the largest model error for V_p in Figure 3 (g) is located below the anomaly, and we can also observe that the standard deviation below the anomaly is large. We can also clearly see more noise is presented on these model errors, especially for the deeper part of

the model, indicating that we should have high uncertainty in these areas, indicating the successful uncertainty quantification of the FWI.

Both the uncertainties given by the inverse approximation Hessian and the BNN have the ability to characterize the inverse results for the model uncertainty of FWI. BNN method have good ability to characterize overall uncertainty of the elastic model, but noise can be observed in the MAP model and the uncertainty quantification model. Inverse Hessian approximation method is noise free, but have some limitation on estimating the uncertainty of the model where it have weekly illuminated, and a modification of the initial guessing could help to release this issue.

CONCLUSION

We compare the uncertainty analysis given by the inverse Hessian approximation method in conventional FWI with the uncertainty analysis given by the BNN. Both of the method have successfully capture the main characterization of model error, though with difference uncertainty pattern. In the uncertainty analysis, using the inverse Hessian approximation method, we improve the uncertainty analysis by changing the initialization of the inverse Hessian. The deeper part of the model have better uncertainty analysis by utilization such a method.

ACKNOWLEDGMENT

We thank the sponsors of CREWES for continued support. This work was funded by CREWES industrial sponsors and NSERC (Natural Science and Engineering Research Council of Canada) through the grant CRDPJ 543578-19. The first author is also supported by the Chine Scholarship Council (CSC).

REFERENCES

An, M., 2012, A simple method for determining the spatial resolution of a general inverse problem: Geophysical Journal International, **191**, No. 2, 849–864.

Bui-Thanh, T., Ghattas, O., Martin, J., and Stadler, G., 2013, A computational framework for infinite-dimensional bayesian inverse problems part i: The linearized case, with application to global seismic inversion: SIAM Journal on Scientific Computing, **35**, No. 6, A2494–A2523.

Chen, K., and Sacchi, M. D., 2020, Time-domain elastic gauss–newton full-waveform inversion: a matrix-free approach: Geophysical Journal International, **223**, No. 2, 1007–1039.

Evensen, G., 1994, Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics: Journal of Geophysical Research: Oceans, **99**, No. C5, 10,143–10,162.

Fichtner, A., and Leeuwen, T. v., 2015, Resolution analysis by random probing: Journal of Geophysical Research: Solid Earth, **120**, No. 8, 5549–5573.

Fichtner, A., and Trampert, J., 2011, Hessian kernels of seismic data functionals based upon adjoint techniques: Geophysical Journal International, **185**, No. 2, 775–798.

Fletcher, R., 2013, Practical methods of optimization: John Wiley & Sons.

Fletcher, R., and Powell, M. J., 1963, A rapidly convergent descent method for minimization: The computer journal, **6**, No. 2, 163–168.

Gauss, C. F., 1877, Theoria motus corporum coelestium in sectionibus conicis solem ambientium, vol. 7: FA Perthes.

Gouveia, W. P., and Scales, J. A., 1998, Bayesian seismic waveform inversion: Parameter estimation and uncertainty analysis: Journal of Geophysical Research: Solid Earth, **103**, No. B2, 2759–2779.

Halko, N., Martinsson, P.-G., and Tropp, J. A., 2011, Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions: SIAM review, **53**, No. 2, 217–288.

Jordan, M., 2015, Estimation of spatial uncertainties in tomographic images, *in* 77th EAGE Conference and Exhibition-Workshops, vol. 2015, European Association of Geoscientists & Engineers, 1–4.

Kalman, R. E., 1960, A new approach to linear filtering and prediction problems.

Köhn, D., De Nil, D., Kurzmann, A., Przebindowska, A., and Bohlen, T., 2012, On the influence of model parametrization in elastic full waveform tomography: Geophysical Journal International, **191**, No. 1, 325–345.

Lailly, P., 1983, The seismic inverse problem as a sequence of before stack migration: Conference on inverse scattering: Theory and applications: Society for industrial and applied mathematics, *in* Expanded Abstracts.

Liberty, E., Woolfe, F., Martinsson, P.-G., Rokhlin, V., and Tygert, M., 2007, Randomized algorithms for the low-rank approximation of matrices: Proceedings of the National Academy of Sciences, **104**, No. 51, 20,167–20,172.

Liu, Q., and Peter, D., 2019, Square-root variable metric based elastic full-waveform inversion—part 2: Uncertainty estimation: Geophysical Journal International, **218**, No. 2, 1100–1120.

Nocedal, J., and Wright, S. J., 1999, Numerical optimization: Springer.

Pan, W., Innanen, K. A., Margrave, G. F., Fehler, M. C., Fang, X., and Li, J., 2016, Estimation of elastic constants for hti media using gauss-newton and full-newton multiparameter full-waveform inversion: Geophysics, **81**, No. 5, R275–R291.

Petra, N., Martin, J., Stadler, G., and Ghattas, O., 2014, A computational framework for infinite-dimensional bayesian inverse problems, part ii: Stochastic newton mcmc with application to ice sheet flow inverse problems: SIAM Journal on Scientific Computing, **36**, No. 4, A1525–A1555.

Pratt, R. G., 1999, Seismic waveform inversion in the frequency domain, part 1: Theory and verification in a physical scale model: Geophysics, **64**, No. 3, 888–901.

Rawlinson, N., Fichtner, A., Sambridge, M., and Young, M. K., 2014, Seismic tomography and the assessment of uncertainty: Advances in geophysics, **55**, 1–76.

Tarantola, A., 1984, Inversion of seismic reflection data in the acoustic approximation: Geophysics, **49**, No. 8, 1259–1266.

Tarantola, A., 1986, A strategy for nonlinear elastic inversion of seismic reflection data: Geophysics, **51**, No. 10, 1893–1903.

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

Trampert, J., Fichtner, A., and Ritsema, J., 2013, Resolution tests revisited: the power of random numbers: Geophysical Journal International, **192**, No. 2, 676–680.

Tromp, J., Tape, C., and Liu, Q., 2005, Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels: Geophysical Journal International, **160**, No. 1, 195–216.

Virieux, J., 1986, P-sv wave propagation in heterogeneous media: Velocity-stress finite-difference method: Geophysics, **51**, No. 4, 889–901.

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

Zhang, J., and McMechan, G. A., 1995, Estimation of resolution and covariance for large matrix inversions: Geophysical Journal International, **121**, No. 2, 409–426.

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