JGR Solid Earth

RESEARCH ARTICLE
10.1029/2024JB029557

Physically Structured Variational Inference for Bayesian Full Waveform Inversion

Xuebin Zhao[1] and **Andrew Curtis**[1]

[1]School of Geosciences, University of Edinburgh, Edinburgh, UK

Key Points:

- We propose a new variational inference methodology to construct a Bayesian posterior solution with a desired correlation structure
- The method is efficient in terms of both memory requirements and computation, with some loss of generality in the solution
- We apply the inversion results to two post-inversion problems where the volume of stored CO_2 in a subsurface reservoir is estimated

Correspondence to:
X. Zhao,
xuebin.zhao@ed.ac.uk

Citation:
Zhao, X., & Curtis, A. (2024). Physically structured variational inference for Bayesian full waveform inversion. *Journal of Geophysical Research: Solid Earth*, *129*, e2024JB029557. https://doi.org/10.1029/2024JB029557

Received 21 MAY 2024
Accepted 4 NOV 2024

Author Contributions:
Conceptualization: Xuebin Zhao
Formal analysis: Andrew Curtis
Funding acquisition: Andrew Curtis
Investigation: Xuebin Zhao
Methodology: Xuebin Zhao, Andrew Curtis
Project administration: Andrew Curtis
Software: Xuebin Zhao
Supervision: Andrew Curtis
Validation: Andrew Curtis
Visualization: Xuebin Zhao
Writing – original draft: Xuebin Zhao
Writing – review & editing: Andrew Curtis

Abstract Full waveform inversion (FWI) creates high resolution models of the Earth's subsurface structures from seismic waveform data. Due to the non-linearity and non-uniqueness of FWI problems, finding globally best-fitting model solutions is not necessarily desirable since they fit noise as well as the desired signal in data. Bayesian FWI calculates a so-called posterior probability distribution function, which describes all possible model solutions and their uncertainties. In this paper, we solve Bayesian FWI using variational inference, and propose a new methodology called physically structured variational inference, in which a physics-based structure is imposed on the variational distribution. In a simple example motivated by prior information from imaging inverse problems, we include parameter correlations between pairs of spatial locations within a dominant wavelength of each other, and set other correlations to zero. This makes the method far more efficient compared to other variational methods in terms of both memory requirements and computation, at the cost of some loss of generality in the solution found. We demonstrate the proposed method with a 2D acoustic FWI scenario, and compare the results with those obtained using other methods. This verifies that the method can produce accurate statistical information about the posterior distribution with hugely improved efficiency (in our FWI example, 1 order of magnitude reduction in computation). We further demonstrate that despite the possible reduction in generality of the solution, the posterior uncertainties can be used to solve post-inversion interrogation problems connected to estimating volumes of subsurface reservoirs and of stored CO_2, with minimal bias, creating a highly efficient FWI-based decision-making workflow.

Plain Language Summary This paper introduces a method to assess uncertainties in seismic images of the subsurface at substantially reduced cost, and to use the information within those uncertainties to answer explicit high-level questions about volumes of subsurface reservoirs and of stored CO_2. Computation efficiency is achieved by explicitly imposing specific correlation structures (originating from physical properties of such spatial imaging problems) on subsurface images a priori. This prevents computing power from being used to re-discover such structures each time an imaging process is performed. In our two-dimensional example in which we image using seismic full waveform inversion, computational cost is reduced by an order of magnitude and fully nonlinear uncertainties can be characterized both in subsurface structural parameters, and in answers to high-level questions.

1. Introduction

Seismic full waveform inversion (FWI) is a method that generates models of the subsurface seismic velocity structure of the Earth given recorded seismograms. This is achieved using both kinematic (phase) and dynamic (amplitude) information in the waveforms (Fichtner et al., 2009; French & Romanowicz, 2014; Tarantola, 1984; Virieux & Operto, 2009). Traditionally, FWI problems are solved using gradient-based local optimization methods, where a misfit function between observed and predicted waveform data is minimized iteratively (Plessix, 2006). This process often requires additional regularization a priori, such as the addition of smoothing or damping terms to the objective function, to stabilize the optimization and improve convergence rates (Asnaashari et al., 2013; Sen & Roy, 2003; Zhdanov, 2002). However, these terms may introduce biases to the final inversion results. In addition, it is challenging to find a good approximation to the true Earth structure using these methods due to the strong non-linearity of the forward function and the non-uniqueness of the inverse problem solution (Boyd & Vandenberghe, 2004).

Recently, FWI has been solved probabilistically using a suite of methods collectively referred to as *Bayesian inference*. In Bayesian FWI, *prior* knowledge about Earth model parameters is updated with new information from the observed waveform data to calculate a *posterior* probability distribution function (pdf), according to

Bayes' rule. In principle this distribution incorporates all prior information combined with all information from the data, and expresses the information in terms of constraints on the model parameters. Bayesian inference thus solves the FWI problem by describing all possible model parameter values that fit the data set to within its uncertainty. The range and probability of different possible models can be used to reduce risk during subsequent decision-making when solving real-world interrogation problems (Arnold & Curtis, 2018; Ely et al., 2018; Poliannikov & Malcolm, 2016; Siahkoohi et al., 2022; X. Zhang & Curtis, 2022; X. Zhao et al., 2022). Note that regularization terms used in deterministic inversion can act similarly to prior information used in Bayesian inference, both of which would bias the inversion results if defined incorrectly.

Different kinds of Bayesian inference methods have been employed to perform probabilistic FWI. A direct generalization from deterministic FWI involves approximating the posterior pdf with a Gaussian distribution, centered around an estimated maximum a posteriori (MAP) model obtained using local optimization methods (Bui-Thanh et al., 2013; Fang et al., 2018; Gouveia & Scales, 1998; Zhu et al., 2016), or through local, low rank pdf approximations using a data assimilation technique (Thurin et al., 2019). If both the likelihood function and prior distribution are assumed to be Gaussians, then this MAP velocity model is equivalent to that obtained using l_2 regularized deterministic FWI (W. Wang et al., 2023). While this kind of method appear to produce probabilistic results, the center point of the resulting posterior distribution may be affected by the starting point of the inversion, and the variances of the Gaussians may not fully capture the uncertainty arising from non-linearity of the forward function (Z. Zhao & Sen, 2021).

Fully non-linear Bayesian FWI can be solved using sampling techniques such as Markov chain Monte Carlo (McMC), where random samples are drawn from the posterior distribution. The inversion results are represented by statistics of the sampled models, such as the mean and standard deviation. For complex posterior distributions such as those that are multimodal, some form of posterior histograms or movies of posterior samples are also analyzed. However, due to the typical high dimensionality (number of parameters to be estimated) of FWI problems, direct sampling methods, including the commonly used Metropolis-Hastings Markov-chain Monte Carlo (MH)-McMC (Hastings, 1970; Metropolis et al., 1953; Mosegaard & Tarantola, 1995; Sambridge & Mosegaard, 2002), become impractical. Some studies therefore employ a target-oriented FWI strategy to reduce the dimensionality and thus improve the sampling efficiency (Ely et al., 2018; Fu & Innanen, 2022; Kotsi et al., 2020b), but these generally have to assume that modeled structure and properties outside of the target volume do not adversely affect waveform fits.

Several advanced techniques have also been introduced to improve the sampling efficiency of McMC for Bayesian FWI, including reversible-jump McMC (RJ-McMC) (P. Guo et al., 2020; Ray et al., 2016, 2018; Sambridge et al., 2006; Visser et al., 2019), Hamiltonian Monte Carlo (de Lima, Corso, et al., 2023; de Lima, Ferreira, et al., 2023; Dhabaria & Singh, 2024; Gebraad et al., 2020; Kotsi et al., 2020a; Zunino et al., 2023), informed-proposal Monte Carlo (Khoshkholgh et al., 2021, 2022) and other gradient based random sampling methods (Berti et al., 2023; Biswas & Sen, 2022; Z. Zhao & Sen, 2021). Nevertheless, as with other classes of methods, Monte Carlo sampling is known to become computationally intractable for high-dimensional parameter spaces due to the curse of dimensionality (Curtis & Lomax, 2001).

In contrast to random sampling, variational inference solves Bayesian inversion through optimization. In variational methods, we define a family of known and tractable distributions, referred to as the variational family. From this family, an optimal member is chosen to approximate the true posterior pdf by minimizing the difference between the variational and posterior distributions (Bishop, 2006; Blei et al., 2017; C. Zhang et al., 2018; X. Zhang et al., 2021). Variational inference solves Bayesian problems under an optimization framework, and the optimization result is fully probabilistic. In some classes of problems it can therefore be relatively more efficient and scalable to high dimensional problems with large data sets. Variational inference has been applied to a variety of geophysical inverse problems, including travel time tomography (Levy et al., 2022; X. Zhang & Curtis, 2020a; X. Zhao et al., 2021), seismic migration (Siahkoohi & Herrmann, 2021; Siahkoohi et al., 2020, 2021, 2023), seismic amplitude inversion (Zidan et al., 2022), earthquake hypocenter inversion (Smith et al., 2022) and slip distribution inversion (L. Sun et al., 2023). However, most of these applications have relatively lower dimensionality and weaker non-linearities compared to FWI.

X. Zhang and Curtis (2020b) introduced a variational method called Stein variational gradient descent (SVGD—Q. Liu & Wang, 2016) to transmission FWI where sources emulating earthquakes are located underneath the velocity structure to be imaged, with receivers on the top surface. SVGD was then applied to 2D reflection FWI

with realistic priors (Izzatullah et al., 2023; X. Zhang & Curtis, 2021a), and 3D acoustic FWI using synthetic data (X. Zhang et al., 2023) and field data (Lomas et al., 2023). A stochastic version of SVGD (Gallego & Insua, 2018) was also employed to improve performance for 3D FWI (X. Zhang et al., 2023). X. Zhao and Curtis (2024a) introduced boosting variational inference (BVI—F. Guo et al., 2016; Miller et al., 2017) for 2D acoustic FWI, where a mixture of Gaussian distributions is used to approximate the true posterior distribution, resulting in an analytic expression for the posterior distribution. Bates et al. (2022) performed medical ultrasound tomography of the brain using FWI, where a mean field (diagonal) Gaussian distribution is employed as the variational distribution. Alternatively, W. Wang et al. (2023) improved the resolution of inversion results by decomposing the variational objective function into two terms and re-weighting them, however that method nevertheless tends to underestimate posterior uncertainties. Yin et al. (2024) used conditional normalizing flows to quantify uncertainties in migration-velocity models.

Other than in W. Wang et al. (2023), in the above studies variational methods were applied to improve the efficiency of Bayesian FWI. For 2D FWI, the required number of forward simulations used to estimate means and variances of subsurface parameters was reduced to the order of 100,000 by X. Zhao and Curtis (2024a), marking a significant reduction given that the dimensionality of the FWI problem tackled was higher than 10,000. Unfortunately, despite this improvement, the computational cost of solving the forward function in FWI remains prohibitively expensive for many practitioners. Consequently, performing Bayesian FWI in realistic projects using current variational methods is still impractical, even with advanced forward simulation strategies (Treeby & Cox, 2010; Y. Wang et al., 2019; X. Zhao et al., 2020).

In this paper, we propose an efficient and accurate variational methodology for Bayesian FWI by imposing physics-based structure on the variational family. The new method incorporates expected posterior parameter correlations explicitly. We show that this leads to significantly improved accuracy with nearly the same computational cost compared to several existing variational methods, or put another way, reduced cost for the same accuracy.

This rest of this paper is organized as follows. In Section 2, we first establish the framework of variational full waveform inversion. Then we introduce the concept of Automatic Differentiation Variational Inference (ADVI), and present our new method which we refer to as physically structured variational inference (PSVI). We end this section by providing a workflow to apply PSVI to Bayesian inverse problems. In Section 3, we demonstrate the proposed method with a 2D synthetic FWI example, and compare the inversion results with those obtained using three other variational methods and to those from traditional linearized optimization based deterministic inversion methods. In Section 4, we interpret the inversion results by solving two post-inversion interrogation problems. Finally, we provide a brief discussion of the proposed method and draw conclusions.

2. Methodology

2.1. Variational FWI

FWI uses full waveform data recorded by seismometers to constrain the Earth's interior structure, typically described by a subsurface velocity model. The forward function is defined to predict waveform data that could be recorded at receivers given a subsurface velocity model. This prediction involves solving a wave equation, either in the time or frequency domain, often in two or three dimensions, and potentially adding measurement noise to the data. For simplicity, we assume that the subsurface consists of an acoustic, isotropic, lossless medium with constant density, thereby ignoring exclusively elastic properties including shear waves, attenuation, and anisotropic properties. This simplification allows the scalar acoustic wave equation to be used in forward simulations which reduces computational load.

In deterministic FWI, the following l_2 norm data misfit function is typically defined between simulated waveform data $\mathbf{d}_{syn}$ and observed data $\mathbf{d}_{obs}$:

$$\chi = \frac{1}{2}||\mathbf{d}_{syn} - \mathbf{d}_{obs}||_2^2 \tag{1}$$

This function is minimized iteratively using gradient based optimization methods, where gradients of the misfit function with respect to model parameters are calculated using the adjoint state method (Plessix, 2006).

In Bayesian FWI, information about the velocity model is characterized by a *posterior* probability distribution function (pdf) which describes the uncertainties associated with different potential models given the observed data. This can be calculated using Bayes' rule:

$$p(\mathbf{m}|\mathbf{d}_{obs}) = \frac{p(\mathbf{d}_{obs}|\mathbf{m})\,p(\mathbf{m})}{p(\mathbf{d}_{obs})} \tag{2}$$

where $p(\cdot)$ denotes a probability distribution. Symbol $x|\,y$ indicates conditional dependence between two random variables x and y, and reads as x given y. Term $p(\mathbf{m})$ describes the *prior* information available on the model parameter $\mathbf{m}$, and $p(\mathbf{d}_{obs}|\mathbf{m})$ is the *likelihood*, meaning the probability of the synthetic waveform data $\mathbf{d}_{syn}$ generated by a given model $\mathbf{m}$ through forward simulation matching the observed data $\mathbf{d}_{obs}$. A Gaussian distribution is often used to define the data likelihood function:

$$p(\mathbf{d}_{obs}|\mathbf{m}) \propto \exp\left[-\frac{(\mathbf{d}_{syn}-\mathbf{d}_{obs})^T\Sigma_{\mathbf{d}}^{-1}(\mathbf{d}_{syn}-\mathbf{d}_{obs})}{2}\right] \tag{3}$$

where $\Sigma_{\mathbf{d}}$ is the covariance matrix of the data error. If a diagonal matrix with a fixed value for all diagonal elements is used to define $\Sigma_{\mathbf{d}}$ (i.e., $\Sigma_{\mathbf{d}} = \sigma^2\mathbf{I}$ where σ is a scalar value controlling the magnitude of the data noise and $\mathbf{I}$ is the identity matrix), then the likelihood can be expressed as a function of data misfit χ in Equation 1:

$$p(\mathbf{d}_{obs}|\mathbf{m}) \propto \exp\left[-\frac{\chi}{\sigma^2}\right] \tag{4}$$

and minimizing χ is equivalent to maximizing the likelihood. The denominator $p(\mathbf{d}_{obs})$ in Equation 2 is referred to as the *evidence* and is a normalization constant to ensure that the posterior solution $p(\mathbf{m}|\mathbf{d}_{obs})$ is a valid probability distribution.

Bayesian inversion is often solved by Monte Carlo sampling methods. However, the required number of samples increases exponentially with the dimensionality of the inverse problem (the number of unknown model parameters), due to the curse of dimensionality (Curtis & Lomax, 2001). It is very expensive to obtain statistics of posterior pdf's in FWI using Monte Carlo methods, especially when the Earth model $\mathbf{m}$ contains more than 10,000 parameters, as is standard in spatially 2D FWI problems (Gebraad et al., 2020). What is worse, 3D FWI problems contain a significantly larger number (possibly many more than a million) of model parameters, making it almost impossible to solve using purely random sampling methods.

In this paper, we use variational inference to solve Bayesian FWI problems. In variational methods, a family of distributions (called the variational family) $\mathcal{Q}(\mathbf{m}) = \{q(\mathbf{m})\}$ is defined, from which we select an optimal member to approximate the true (unknown) posterior distribution. The optimal distribution can be found by minimizing the difference (distance) between the posterior and variational distributions. Typically, the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951) is used to measure the distance between two probability distributions, defined as the following expectation term

$$\mathrm{KL}[q(\mathbf{m})||\,p(\mathbf{m}|\mathbf{d}_{obs})] = \mathbb{E}_{q(\mathbf{m})}[\log q(\mathbf{m}) - \log p(\mathbf{m}|\mathbf{d}_{obs})] \tag{5}$$

The KL divergence of two distributions is non-negative, and equals zero only when the two distributions are identical. Substituting Equation 2 into 5, we find that minimizing the $\mathrm{KL}[q(\mathbf{m})||\,p(\mathbf{m}|\mathbf{d}_{obs})]$ is equivalent to maximizing the following evidence lower bound (ELBO) on $\log p(\mathbf{d}_{obs})$ since $\log p(\mathbf{d}_{obs})$ is a constant independent of $q(\mathbf{m})$:

$$\begin{aligned}\mathrm{ELBO}[q(\mathbf{m})] &= \log p(\mathbf{d}_{obs}) - \mathrm{KL}[q(\mathbf{m})||\,p(\mathbf{m}|\mathbf{d}_{obs})] \\ &= \mathbb{E}_{q(\mathbf{m})}[\log p(\mathbf{m},\mathbf{d}_{obs}) - \log q(\mathbf{m})]\end{aligned} \tag{6}$$

In this way, we convert a random sampling problem into a numerical optimization, while the optimization result is still a probability distribution that approximates the true posterior pdf.

A key challenge in variational inference is to choose the variational family $\mathcal{Q}(\mathbf{m})$. This determines both the accuracy and efficiency of the variational methods: increasing the complexity (and hence, expressivity) of $\mathcal{Q}(\mathbf{m})$ increases the approximation accuracy as well as the optimization complexity. Given the expensive nature of forward simulations in FWI, our primary goal is to reduce computational costs (by reducing the number of forward simulations) while maintaining accuracy at an acceptable level. In the following sections we introduce a method called automatic differentiation variational inference (ADVI—Kucukelbir et al., 2017), and propose an alternative effective variational methodology for FWI.

2.2. ADVI

ADVI is a well-established variational method that defines a Gaussian variational distribution $q = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, parametrized by a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$ (Kucukelbir et al., 2017). In addition, since a Gaussian distribution is defined over the space of real numbers and since in most geophysical imaging problems model parameters are bounded by physical constraints (e.g., seismic velocity should be a positive number), an invertible transform (a bijection) is applied to the Gaussian random variables to ensure that the transformed model parameters satisfy their physical constrains. We use the commonly used *logit* functions

$$\begin{aligned} m_i &= f(\theta_i) = a_i + \frac{b_i - a_i}{1 + \exp(-\theta_i)} \\ \theta_i &= f^{-1}(m_i) = \log(m_i - a_i) - \log(b_i - m_i) \end{aligned} \tag{7}$$

where θ_i is a Gaussian random variable defined in an unconstrained space (from minus to plus infinity), and m_i is the converted model parameter bounded by the lower and upper bounds a_i and b_i, respectively. The transformed probability distribution can be calculated through the change of variable formula

$$\log q(\mathbf{m}) = \log q(\boldsymbol{\Theta}) - \log|\det(\partial_{\boldsymbol{\Theta}} f(\boldsymbol{\Theta}))| \tag{8}$$

where $q(\boldsymbol{\Theta}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the Gaussian variational distribution in the unbounded space. Term $|\det(\cdot)|$ calculates the absolute value of the determinant of the Jacobian matrix $\partial_{\boldsymbol{\Theta}} f(\boldsymbol{\Theta})$, which accounts for the volume change in parameter space corresponding to this transform (Kucukelbir et al., 2017; Rezende & Mohamed, 2015). Variable $\boldsymbol{\Theta}$ is optimized by the ADVI algorithm, and the corresponding $q(\mathbf{m})$ is used to approximate the true posterior distribution.

To determine the optimal Gaussian distribution in the unbounded space, we maximize the ELBO$[q(\mathbf{m})]$ in Equation 6 with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ using a gradient based optimization method. The two gradient terms can be calculated by

$$\nabla_{\boldsymbol{\mu}} \text{ELBO} = \mathbb{E}_{\mathcal{N}(\mathbf{0},\mathbf{I})}[\nabla_{\mathbf{m}} \log p(\mathbf{m}, \mathbf{d}_{obs}) \nabla_{\boldsymbol{\mu}} \mathbf{m} - \nabla_{\boldsymbol{\mu}} \log q(\mathbf{m})] \tag{9}$$

$$\nabla_{\boldsymbol{\Sigma}} \text{ELBO} = \mathbb{E}_{\mathcal{N}(\mathbf{0},\mathbf{I})}[\nabla_{\mathbf{m}} \log p(\mathbf{m}, \mathbf{d}_{obs}) \nabla_{\boldsymbol{\Sigma}} \mathbf{m} - \nabla_{\boldsymbol{\Sigma}} \log q(\mathbf{m})] \tag{10}$$

where $\log q(\mathbf{m})$ (and its gradients) can be calculated analytically using Equation 8. The expectations are taken with respect to a standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ rather than the variational distribution $\log q(\mathbf{m})$, by applying a reparametrization trick (Kingma & Welling, 2014). This allows the gradients to be moved inside the expectation terms. If the joint distribution $\log p(\mathbf{m}, \mathbf{d}_{obs})$ has analytically tractable gradients with respect to $\mathbf{m}$, the whole process of computing the two gradients above can be automated, hence the name ADVI (Kucukelbir et al., 2017).

The calculation of Equation 10 involves computing $|\boldsymbol{\Sigma}|$, where $|\cdot|$ denotes the determinant of a matrix. Direct calculation of $|\boldsymbol{\Sigma}|$ has a computational complexity of $O(n^3)$, which becomes prohibitively expensive for high dimensional inference problems such as FWI. Therefore, we often use a Cholesky factorization to parametrize $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^{\mathbf{T}} \tag{11}$$

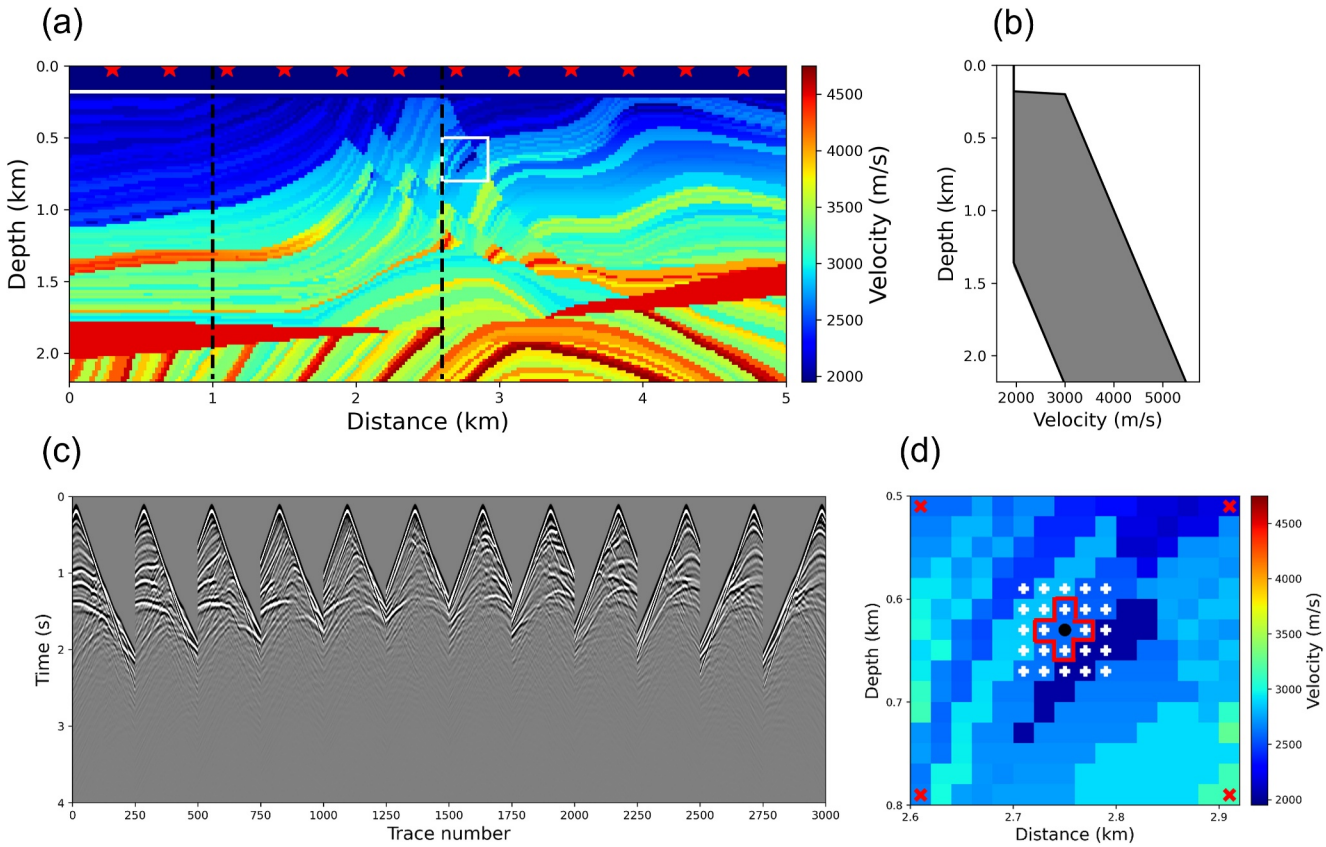

Figure 1. (a) P wave velocity of the Marmousi model used in a 2D acoustic full waveform inversion test. Source locations are indicated by red stars and the receiver line is marked by a white line. Dashed black lines display the locations of two vertical profiles used to compare the posterior marginal probability distributions in Figure 4. (b) Upper and lower bounds of the Uniform prior distribution at different depths. (c) Observed data set which contains 12 common shot gathers. (d) Velocity structure inside the white box in (a), and crosses in cells discussed in the main text.

where $\mathbf{L}$ is a lower triangular matrix. Since $|\mathbf{L}|$ can be calculated easily as the product of its diagonal elements, the determinant $|\mathbf{\Sigma}|$ can be obtained easily by $|\mathbf{\Sigma}| = |\mathbf{L}|^2$. Note that the diagonal elements of $\mathbf{L}$ are associated with the variances of model parameters, and should be non-negative to ensure that $\mathbf{L}$ and $\mathbf{\Sigma}$ are positive semidefinite. The off-diagonal values of $\mathbf{L}$ contain correlation information between model parameters.

For a n-dimensional problem, we need $n(n+1)/2$ parameters to construct a full matrix $\mathbf{L}$, and consequently a full covariance matrix $\mathbf{\Sigma}$. The corresponding method is known as full rank ADVI (Kucukelbir et al., 2017). For example, in Figure 1a, the velocity model comprising 110×250 pixels requires 378,138,750 parameters to describe the full matrix $\mathbf{L}$. This number becomes computationally intractable for large scale 2D and 3D FWI problems.

Alternatively, a mean field approximation is often used to reduce computational complexity, where $\mathbf{L}$ and $\mathbf{\Sigma}$ are parametrized by diagonal matrices. The variational distribution becomes a diagonal Gaussian distribution, which neglects correlations between different model parameters. In this way, the total number of variables that must be optimized is $2n$ (both $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ contain n independent elements), so is doubled compared to a conventional deterministic inversion, a computational overhead that is manageable for most problems. Mean field ADVI has been applied to Bayesian FWI in several studies (Bates et al., 2022; W. Wang et al., 2023; X. Zhang et al., 2023), demonstrating that the method is computationally efficient and is able to provide an accurate mean model of the posterior distribution. However, in problems with significant posterior correlations, it tends to strongly underestimate posterior parameter uncertainties since correlation information is neglected a priori (X. Zhang et al., 2023).

2.3. PSVI

Full rank ADVI and mean field ADVI represent two extreme approaches to construct $\mathbf{L}$: the former aims to optimize all off-diagonal elements of $\mathbf{L}$ to capture the full correlation information of $\mathbf{m}$, whereas the latter sets the off-diagonal elements to zero to reduce computational requirements. In the following, we parametrize $\mathbf{L}$ using a physics-guided structure, which models a subset of its off-diagonal elements.

In most imaging problems, accurate correlation information plays an important role in capturing true structures such as the continuity of properties across neighboring spatial volumes. Since modeling a full covariance matrix (i.e., full rank ADVI) for high dimensional problems is practically intractable, another approach is to model the most important correlations in vector $\mathbf{m}$, guided by physical properties (prior knowledge) of imaging problems. To illustrate, Figure 1d shows a 2D velocity structure discretized using $nx \times nz$ square grid cells in horizontal and vertical directions, with each cell representing a velocity value at the corresponding spatial location. It is often the case that any grid cell, such as the one marked by a black dot in Figure 1d, is strongly correlated with its surrounding cells (e.g., cells marked by white pluses). The magnitude of correlations between this central cell and other cells decreases as the distance between two locations increases. Cells that are far away from the black dot (e.g., cells denoted by red crosses in Figure 1d) are only weakly correlated with the black-dotted cell, so these correlations can safely be ignored. This feature has been observed in many different imaging problems (Ardizzone et al., 2018; Biswas & Sen, 2022; Gebraad et al., 2020); a clear example displaying such correlations in a velocity profile with depth is shown in Figure 6 of X. Zhang and Curtis (2021b), from the results of surface wave dispersion inversion using two independent nonlinear inversion methods (invertible neural networks and Monte Carlo).

This suggests that it might suffice to model correlations only between parameter values that are spatially close to each other, that is, which lie within a dominant wavelength (since dominant wavefield interactions involved in reflection and refraction respond to controlling physical properties averaged spatially over ~1 wavelength), and ignore those that are far away by assuming a particular sparse structure for $\mathbf{L}$. We therefore set off-diagonal elements of $\mathbf{L}$ which represent the main correlations of interest as parameters to be optimized during variational inversion, while imposing all other off-diagonal elements to be zero. Note that we thus impose only a structure on $\mathbf{L}$ rather than placing constraints on the values of its (non-zero) off-diagonal elements: those values are updated freely during inversion.

Recall that the 2D velocity model displayed in Figure 1d has $nx \times nz$ cells with nx and nz being the numbers of cells in horizontal and vertical directions, defined by vector $\mathbf{m}$ in row-major order (i.e., the first nx elements of $\mathbf{m}$ comprise the first row of the 2D image, the second nx elements comprise the second row, and so on). As illustrated in Equation 12 below, the first-order off-diagonal elements (blue ones in Equation 12 that are directly below the diagonal elements) contain correlation information between two horizontally adjacent grid cells, and off-diagonal elements that are nx rows below the main diagonal elements (red ones in Equation 12) describe correlations between two vertically adjacent cells

$$\mathbf{L} = \begin{bmatrix} l_{0,1} & & & & & & \\ l_{1,1} & l_{0,2} & & & & & \\ 0 & l_{1,2} & l_{0,3} & & & & \\ \dots & 0 & l_{1,3} & \dots & & & \\ l_{nx,1} & \dots & 0 & \dots & l_{0,n-2} & & \\ 0 & \dots & \dots & \dots & l_{1,n-2} & l_{0,n-1} & \\ \dots & 0 & l_{nx,n-nx} & \dots & 0 & l_{1,n-1} & l_{0,n} \end{bmatrix} \tag{12}$$

where subscript $n = nx \times nz$ is the dimensionality of $\mathbf{m}$. Note that in Equation 12, the first subscript i indicates a block of off-diagonal elements that are i rows below the main diagonal (i.e., at an offset of i from the main diagonal), and the second subscript j indicates that $l_{i,j}$ is the jth element of that off-diagonal block. This differs from the commonly used indexing scheme in which the two subscripts imply the row and column number of an element. If we set all remaining elements of $\mathbf{L}$ to zero, then covariance matrix $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^{\mathbf{T}}$ mainly has non-zero

entities only at two off-diagonal blocks located 1 and nx rows below and above the main diagonal elements (similar to the red and blue elements in Equation 12). If such a covariance matrix $\mathbf{\Sigma}$ is used, the variational distribution would also capture a specific spatial correlation structure that includes parameter correlations between pairs of adjacent cells in both horizontal and vertical directions. Thus, for the grid cell denoted by the black dot in Figure 1d, we would model correlations between this cell and its four adjacent cells inside the red box in Figure 1d: all other correlations are set to zero.

We can impose any desired correlation structure on $\mathbf{\Sigma}$, by setting the corresponding off-diagonal blocks in $\mathbf{L}$ as unknown control parameters (i.e., parameters that define $\mathbf{L}$) and optimizing them during inversion. The size of the defined correlation template should be relatively small compared to the dimensionality of the problem, so the total number of parameters required to construct $\mathbf{L}$ would also be relatively small compared to that in full rank ADVI. For example, if the white pluses in Figure 1d are used to define a 5×5 correlation kernel then the required number of parameters to construct $\mathbf{\Sigma}$ is smaller than $13n$. Here n is the dimensionality of model vector $\mathbf{m}$, and the number 13 consists of 1 main diagonal block and 12 off-diagonal blocks representing 12 different offsets between cells marked by the white crosses and the central cell in the 5×5 kernel. Since each off-diagonal block contains fewer parameters than the main diagonal block (i.e., the blue and red elements in Equation 12 are fewer than the diagonal elements), the total number of parameters is smaller than $13n$, which is a significant reduction compared to $n(n+1)/2$ parameters used in full rank ADVI.

We implement the aforementioned approach to parametrize the matrix $\mathbf{L}$ and obtain a sparse approximation of the covariance matrix. For the avoidance of doubt, even though a (transformed) Gaussian is used to parametrize the posterior pdf and gradients are used in the optimization, the posterior uncertainties estimated by this method (and by other forms of ADVI) are approximations to the full, non-linearized, posterior distribution. The inversion results from our new method thus effectively and efficiently capture structured full-posterior correlations. Since this structure originated from the inherent physical properties of imaging problems, we name the method as PSVI.

2.4. PSVI Workflow

In this section, we provide details about how to implement and perform Bayesian inversion using PSVI. Similarly to ADVI, in PSVI we update a variational distribution $q(\mathbf{m})$ (defined by a vector $\boldsymbol{\mu}$ and a matrix $\mathbf{L}$) to maximize the ELBO (Equation 6) using gradient based optimization methods. This can be achieved easily using advanced automatic differentiation libraries such as TensorFlow (Abadi et al., 2016) and PyTorch (Paszke et al., 2019), where gradients of the ELBO can be calculated automatically. To explain, we rewrite Equation 6 as:

$$\begin{aligned} \mathrm{ELBO}[q(\mathbf{m};\boldsymbol{\mu},\mathbf{L})] &= \mathbb{E}_{\mathcal{N}(\mathbf{0},\mathbf{I})}[\log p(\mathbf{m},\mathbf{d}_{obs}) + \log q(\mathbf{m})] \\ &\approx \frac{1}{k}\sum_{i=1}^{k}[\log p(\mathbf{m}_i,\mathbf{d}_{obs}) + \log q(\mathbf{m}_i)] \end{aligned} \tag{13}$$

where the first line is simply the reparametrization trick (derivation details can be found in Appendix A of X. Zhao et al. (2021)). The expectation term in the first line is approximated by Monte Carlo integration, resulting in the second line. Since we would normally perform many iterations to maximize the ELBO, the Monte Carlo approximation would converge statistically toward the correct solution, even if only a small number of (k in this case) samples are used in each iteration—perhaps only a single sample (Kucukelbir et al., 2017) is sufficient for some problems. To estimate the ELBO, we draw random samples ϵ_i from a multivariate standard normal distribution, which are first transformed to $\mathbf{\Theta_i}$ distributed according to $\mathcal{N}(\boldsymbol{\mu}, \mathbf{L}\mathbf{L}^{\mathbf{T}})$, then converted to $\mathbf{m}_i$:

$$\mathbf{m}_i = f(\mathbf{\Theta_i}) = f(\boldsymbol{\mu} + \mathbf{L}\epsilon_i) \tag{14}$$

using Equation 7. After that, we can calculate $\log p(\mathbf{m}_i,\mathbf{d}_{obs})$ and $\log q(\mathbf{m}_i)$ easily. Note that all calculations involved in the second line in Equation 13 are analytic except for $\log p(\mathbf{m}_i,\mathbf{d}_{obs})$ (for most geophysical problems $\log p(\mathbf{m}_i,\mathbf{d}_{obs})$ and its gradient with respect to $\mathbf{m}_i$ are calculated using numerical methods, e.g., the adjoint state method) whose gradient information can be registered into the computational graph of the automatic differentiation engine (e.g., an example of how to do this can be found in Paszke et al. (2017)). Therefore, the gradient of the ELBO can be calculated automatically.

This suggests the following algorithm to implement PSVI for Bayesian inversion:

1. Initialization: set $\boldsymbol{\mu} = \mathbf{0}$ and $\mathbf{L} = \mathbf{I}$ ($\mathbf{L}$ is a sparse matrix with specific off-diagonal elements registered as unknown parameters to be optimized);
2. Draw k random samples ϵ_i from $\mathcal{N}(\mathbf{0}, \mathbf{I})$;
3. Calculate the corresponding model samples $\mathbf{m}_i$ using Equation 14;
4. Estimate the ELBO (Equation 13) and its gradient (using automatic differentiation) to update $\boldsymbol{\mu}$ and $\mathbf{L}$;
5. Repeat 2–4 until the estimates of $\boldsymbol{\mu}$ and $\mathbf{L}$ converge to stable values.

Given that the computational cost of updating the variational parameters is negligible in comparison to forward and gradient calculations in FWI, the proposed method is almost as efficient as mean field ADVI.

3. 2D Acoustic FWI Example

3.1. Bayesian FWI

In this section, we test the proposed PSVI algorithm in a 2D acoustic FWI example. The true velocity model, shown in Figure 1a, is obtained by truncating the original Marmousi model (Martin et al., 2006) and down-sampling it into 110 × 250 regular grid cells. The grid cell size is 20 m in both directions. For simplicity, we maintain a constant density. We simulate 12 sources on the surface with a spacing of 400 m (indicated by red stars in Figure 1a). A receiver line containing 250 receivers at an interval of 20 m is placed on the seabed at 200 m depth (white line in Figure 1a). The observed waveform data are generated by solving the 2D acoustic wave equation using a time-domain finite difference method. The simulation length is 4 s with a sample interval of 2 ms. The source function is a Ricker wavelet with a dominant frequency of 10 Hz. Figure 1c displays this observed waveform data set.

We define a Uniform prior distribution for the velocity values in each grid cell. Figure 1b shows the lower and upper bounds of the prior distribution at different depths. We set the velocity in the water layer (down to 200 m depth) to its true value during inversion. The likelihood function is a Gaussian distribution (Equation 3) with a diagonal covariance matrix Σ_d assuming independence among all data points. The data noise is assumed to be 1% of the average over the set of maximum amplitude values, one from each trace. This indicates that the likelihood function is defined using Equation 4 (based on the l_2 misfit function in Equation 1). The same finite difference solver is used to calculate the synthetic waveform data $\mathbf{d}_{syn}$, and the gradient of the data misfit (negative log-likelihood function) with respect to the velocity model is computed using the adjoint-state method (Plessix, 2006). We follow the workflow described in Section 2.4 to perform PSVI, in which Monte Carlo integration is used to estimate the ELBO. The automatic differentiation framework provided by PyTorch is used to build a computational graph, which calculates the ELBO and its gradient with respect to the variational parameters automatically (Paszke et al., 2019). The optimization process is carried out using the Adam algorithm (Kingma & Ba, 2014).

We apply mean field ADVI and PSVI to this Bayesian FWI problem. Considering the dimensionality of this problem ($100 \times 250 = 25{,}000$), full rank ADVI is not performed since constructing a full covariance matrix would be extremely expensive in terms of both memory requirements and computational cost. For mean field ADVI, we use a diagonal Gaussian distribution to approximate the posterior distribution in the unbounded space. For PSVI, a 5×5 correlation kernel is employed to model the main correlations between model parameters, as illustrated by the white pluses in Figure 1d for the central black dotted cell. The choice of this correlation kernel is based on the estimated dominant wavelength of this problem (approximately 200 m in shallow subsurface). In both tests, variational parameters ($\boldsymbol{\mu}$ and $\mathbf{L}$) are updated for 5,000 iterations, with two random samples per iteration used to approximate the ELBO$[q(\mathbf{m})]$ and its gradients with respect to $\boldsymbol{\mu}$ and $\mathbf{L}$ using Equation 13. Figure 2 displays the negative ELBOs for these two tests as a function of iterations, indicating that both algorithms achieve a reasonable level of convergence with nearly the same convergence speed, even though PSVI has far more parameters to optimize.

Figures 3a and 3b depict the inversion results. The mean, standard deviation and the relative error (computed by dividing the absolute error between the true and mean models by the standard deviation model) of the posterior distribution are displayed from top to bottom row. The two mean velocity maps exhibit similar features across most locations, generally resembling the true velocity map in Figure 1a. The inversion results struggle to recover

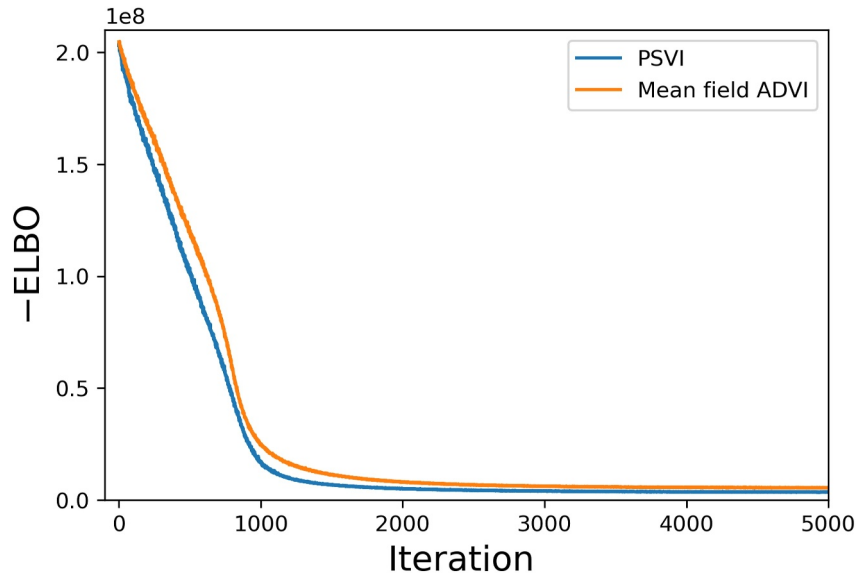

Figure 2. Variation of the negative evidence lower bound with respect to iterations.

some thin layers in the deeper part of the model, potentially due to the relatively low frequency (10 Hz) data used for FWI. Additionally, certain discrepancies are observed between these two maps at specific locations. For example, in the tilt layers annotated by red and black arrows in Figures 3a and 3b, the mean velocity model from mean field ADVI displays discontinuities, while the PSVI results show more continuity, closely resembling the true velocity model. One possible reason for this discrepancy is that accurate correlation information is crucial for recovering the continuity of spatial locations, especially for these thin layers. All correlations between pairs of model parameters are neglected in mean field ADVI, and thus the results may fail to recover the true velocity structures at these locations. By incorporating physically structured correlations between cells within a dominant wavelength, the proposed method improves the inversion accuracy.

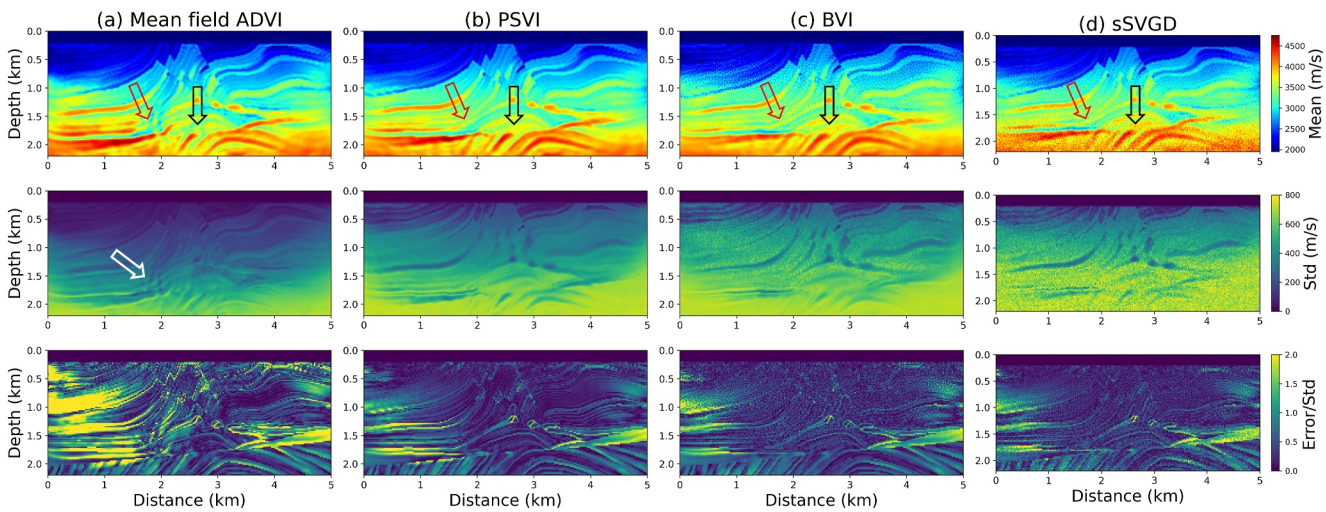

Figure 3. Mean (top row), standard deviation (middle row), and relative error (bottom row) of the posterior distribution obtained using (a) mean field ADVI, (b) PSVI, (c) boosting variational inference, and (d) stochastic Stein variational gradient descent, respectively. The relative error is the absolute error between the mean and true models divided by the corresponding standard deviation. Arrows in (a) highlight regions where mean field Automatic Differentiation Variational Inference (ADVI) fails to recover the correct velocity structures that should be spatially continuous, therefore relative errors from mean field ADVI are large. Physically structured variational inference (PSVI) provides more accurate inversion results that are similar to the true velocity model and to those obtained from two independent inversion methods, since in PSVI the main posterior correlation information is considered.

Both inversion results show increased uncertainties with greater depth, since the sensitivity of observed seismic data decreases at depth, thus deeper parts of the model are less constrained by the data. The standard deviation values obtained from mean field ADVI are generally smaller than those from PSVI, especially in the shallower subsurface above 1.5 km depth. This is because mean field ADVI tends to underestimate posterior uncertainties by neglecting correlations. Similar phenomena have been observed in previous studies (Ely et al., 2018; W. Wang et al., 2023; X. Zhao & Curtis, 2024a). Therefore, the relative errors from mean field ADVI are larger compared to those from the proposed method, especially at locations with a depth of 1 km and a horizontal location between 0 and 1.5 km, where the mean model deviates from the true model by more than 3 standard deviations. This discrepancy suggests a low credibility of the inversion results obtained from mean field ADVI. As marked by a white arrow in Figure 3a, lower uncertainty noise is observed, which correspond to layers that are not continuous in the mean velocity map marked by a red arrow. This feature again proves that mean field ADVI provides biased uncertain estimates. By contrast, such uncertainty structures are not observed in Figure 3b, indicating that PSVI has the capability to correct some biases introduced by mean field ADVI.

To validate the inversion results displayed in Figure 3b, we apply two additional variational methods to this problem: *boosting variational inference* (BVI—F. Guo et al., 2016; Miller et al., 2017) and *stochastic Stein variational gradient descent* (sSVGD—Gallego & Insua, 2018). BVI and sSVGD are two entirely independent Bayesian inference methods which have been applied to acoustic FWI problems, and have shown to provide reasonably correct mean and variance statistics of posterior solutions in two and three dimensional Earth models (X. Zhang et al., 2023; X. Zhao & Curtis, 2024a). In this work, they are used to verify our results. Figures 3c and 3d depict the inversion results obtained using BVI and sSVGD, respectively. They present very similar features compared to those displayed in Figure 3b: the same continuous structures in the deeper part of the model (denoted by red and black arrows) are observed in the mean velocity maps, and similar higher standard deviation values associated with lower relative errors (distributed within 2 standard deviations) are also present.

To further analyze the accuracy of the inversion results, in Figure 4 we compare the posterior marginal distributions obtained from the four tested methods along two vertical profiles at horizontal locations of 1 km (top row) and 2.6 km (bottom row), respectively. The location of these two profiles are displayed by dashed black lines in Figure 1a. The first profile (at a horizontal location of 1 km) is strategically placed in regions where the relative errors from mean field ADVI (Figure 3a) are higher, while the second one (at 2.6 km) is centrally located within the imaging region. Red lines show the true velocity values and black lines show the mean velocity values obtained using different methods. Overall, the marginal distributions in Figure 4a are narrower compared to those in Figure 4b–4d, indicating lower posterior uncertainties akin to Figure 3. In the first row of Figure 4 between depths of 0.7–1 km and 1.3–1.8 km, the true velocity values are excluded from the posterior distribution obtained using mean field ADVI, whereas those values correctly reside within the high probability region of the posterior pdfs obtained using the other three methods. These phenomena again prove that mean field ADVI tends to underestimate the posterior uncertainties and introduce biases into the inversion results. By including the main correlation information between adjacent grid cells, PSVI yields better inversion results that are highly consistent with two entirely independent methods. Therefore, we assert that the posterior standard deviations derived from PSVI are likely to be correct.

Given that PSVI is designed to capture correlations between spatially close grid cells, we compare the posterior correlation coefficients between model parameters estimated using different methods. Figure 5 shows the covariance matrices for velocity values within the white box in Figure 1a, obtained using the above four inversion methods. Mean field ADVI uses a transformed diagonal Gaussian distribution to approximate the posterior pdf and disregards correlations between model parameters, thus the posterior covariance matrix predominantly exhibits strong diagonal values corresponding to the variances of model parameters. By incorporating a specific (desired) correlation structure into the variational distribution, the covariance matrix obtained using PSVI displays off-diagonal values representing correlations between different parameters, which are not observed from the results using mean field ADVI. Due to the use of a 5×5 correlation kernel (as represented by the white pluses in Figure 1d), we only include correlation information between a given grid cell and cells within two layers of cells surrounding it. As a result, Figure 5b displays four off-diagonal blocks (two above and two below the diagonal elements). We observe negative correlations between neighboring cells (in the first off-diagonal block below and above the diagonal values) and positive correlations between every second neighboring cells (found in the second off-diagonal block).

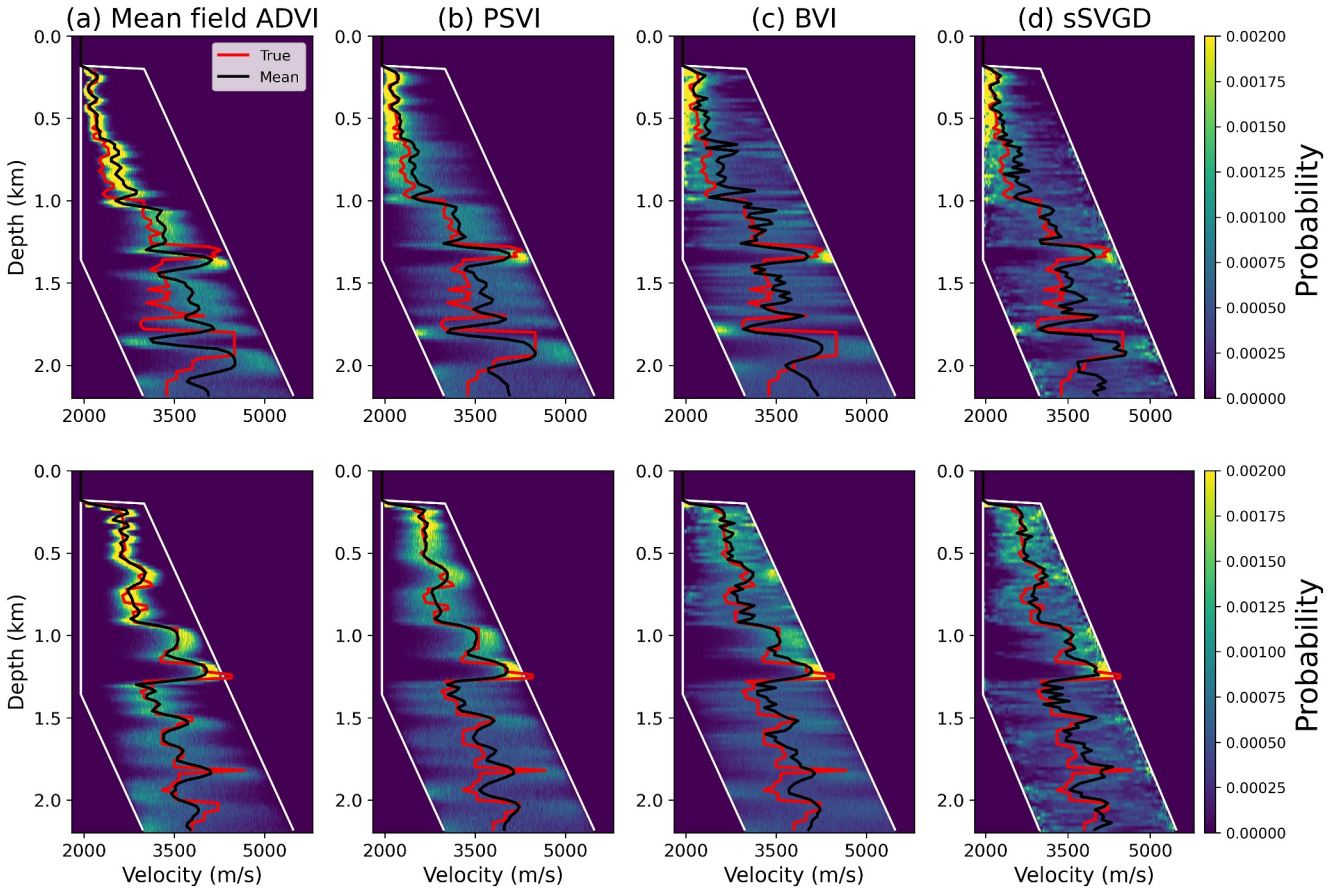

Figure 4. Posterior marginal distributions colored from dark blue (zero probability) to yellow (maximum value of marginal pdf's in each plot), along two vertical profiles at distances of 1 km (top row) and 2.6 km (bottom row) obtained using (a) mean field ADVI, (b) PSVI, (c) boosting variational inference and (d) stochastic Stein variational gradient descent. The locations of these two profiles are represented by black dashed lines in Figure 1a. In each figure, two white lines show the prior bounds, and black and red lines show the mean and true velocity values. Similarly to Figure 3, mean field Automatic Differentiation Variational Inference provides inaccurate inversion results (especially in the top row in which the true velocity values are excluded from the posterior pdf) with underestimated posterior uncertainties, whereas the other three mutually independent methods provide highly consistent marginal pdf's, indicating that those results are likely to be correct.

In Figures 5c and 5d, similar negative off-diagonal correlation blocks are observed in the covariance matrices obtained using BVI and sSVGD. This confirms that in this test we successfully capture the correct correlation information between adjacent cells by using PSVI. While there may be positive correlations with cells two layers apart, these are not visible; this may be because Figures 5c and 5d show a general "speckle" of non-zero background correlation values that are absent in Figure 5b. In PSVI, we construct a sparse covariance matrix with specific non-zero off-diagonal elements, and set all other values to zero. This neglects correlations between locations that are spatially far away from each other. It should be noted that we do not know whether any of the values in Figures 5c and 5d are correct, since they do not match between the two panels. In the next section, we also prove that these non-zero background correlations play a less significant role in a simulation of a real-world decision-making process. So again we suggest that our implementation of PSVI has modeled the most prominent and consistent features of the correlation structure.

Full rank ADVI was not performed in the above example due to the extreme cost of both memory requirements and computation. In Appendix A we run another FWI test with lower dimensionality, making it possible to apply full rank ADVI and to compare the inversion results directly with those obtained using mean field ADVI and PSVI. That example provides another demonstration of what information is lost in PSVI when we model specific correlations only.

Finally, we analyze the efficiency of the proposed method and compare its cost with other methods. As mentioned in Section 2, the number of unknown control parameters that need to be optimized in PSVI is higher than that in

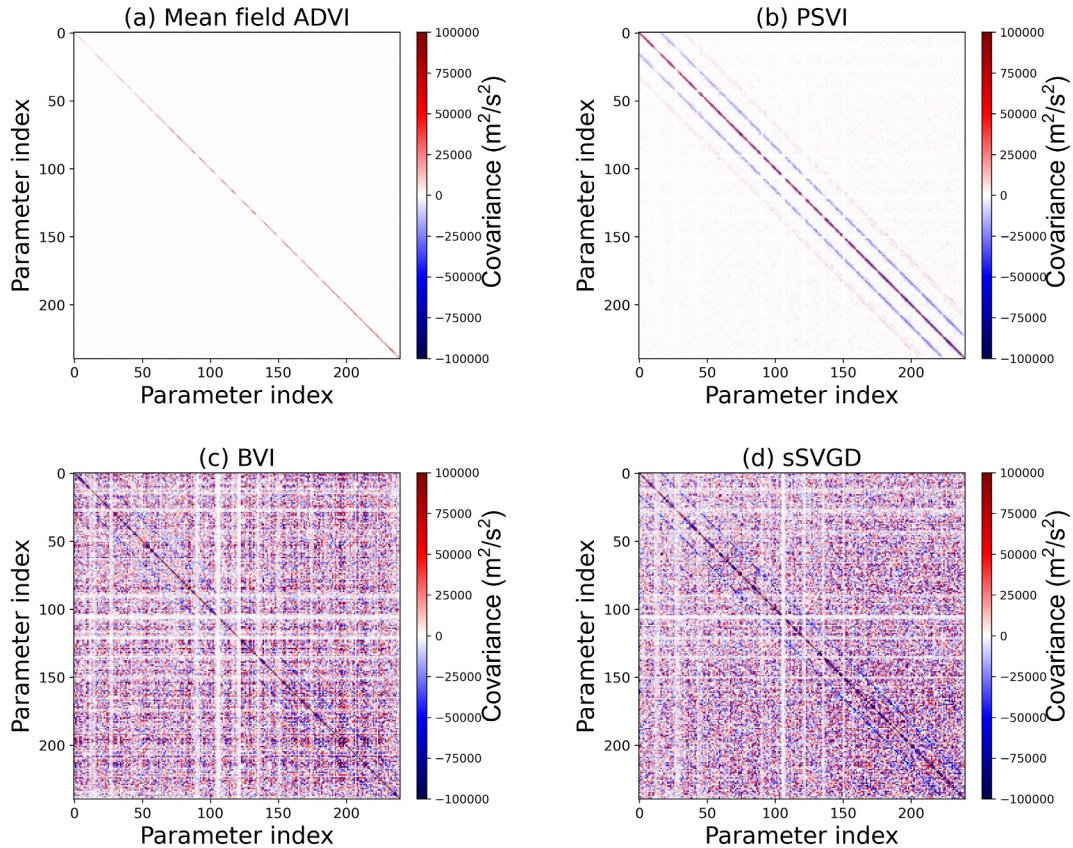

Figure 5. Covariance matrices for velocity values inside the white box in Figure 1a, calculated using the inversion results from (a) mean field ADVI, (b) PSVI, (c) boosting variational inference, and (d) stochastic Stein variational gradient descent.

mean field ADVI but is significantly lower than that in full rank ADVI. In our test, we find that the computational cost for optimizing these variational parameters is much cheaper (almost negligible) compared to the cost of forward and adjoint simulations in FWI. Therefore, the number of simulations serves as a good metric for the overall cost in this example.

Table 1 summarizes the number of simulations used in each method. The same simulation settings are used in mean field ADVI and PSVI (10,000 simulations consisting of 5,000 iterations with 2 samples per iteration). For both BVI and sSVGD, we perform a total of 120,000 forward evaluations. In addition, relatively larger step sizes are used in these two tests to speedup the convergence of BVI and sSVGD. However, they still remain one order of magnitude more computationally expensive than mean field ADVI and PSVI. In addition, Figure 2 shows that mean field ADVI and PSVI present roughly the same convergence rate given the same number of forward simulations. This verifies the statement that PSVI is almost as efficient as mean field ADVI. The latter is known to be a particularly inexpensive (yet biased) method for Bayesian inversion from previous studies (Bates et al., 2022; L. Sun et al., 2023; X. Zhang & Curtis, 2020a; X. Zhao et al., 2021). On the other hand, the PSVI method improves the inversion accuracy and provides similar results compared to two accurate but more computationally demanding methods (BVI and sSVGD). Thus, the proposed method is shown to be an efficient algorithm that has provided reliable uncertainty estimates.

Table 1
Number of Forward and Gradient Evaluations for Mean Field Automatic Differentiation Variational Inference, Physically Structured Variational Inference, Boosting Variational Inference, and Stochastic Stein Variational Gradient Descent

Method	Number of gradient evaluations
Mean field ADVI	10,000
PSVI	10,000
BVI	120,000
sSVGD	120,000

Note. The values represent an indication of the computational cost of each method, as the evaluation of data-model gradients in FWI is by far the most expensive part of each calculation.

3.2. Deterministic FWI

In this section we compare variational Bayesian FWI results to inversion results obtained using traditional deterministic FWI methods. We consider

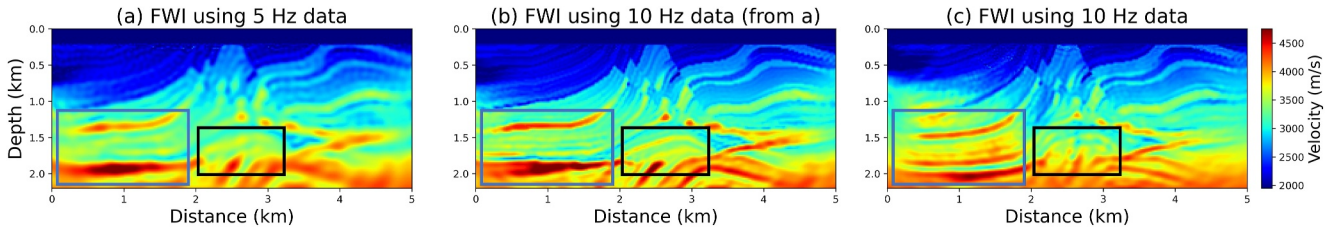

Figure 6. Deterministic full waveform inversion results obtained using waveform data with a dominant frequency of (a) 5 Hz, (b) 10 Hz but starting from the low frequency inversion results displayed in (a), and (c) 10 Hz directly. Blue and black boxes highlight differences between these three inversion results.

the l_2 misfit function in Equation 1 with additional damping and smoothing terms to stabilize the inversion process:

$$\chi_{reg} = \frac{1}{2\sigma^2}\|\mathbf{d}_{syn} - \mathbf{d}_{obs}\|_2^2 + \lambda_1\|\mathbf{m} - \mathbf{m}_0\|_2^2 + \lambda_2\|\mathbf{Dm}\|_2^2 \tag{15}$$

where σ is the magnitude of data noise, λ_1 and λ_2 are two parameters controlling the magnitude of the damping term $\|\mathbf{m} - \mathbf{m}_0\|_2^2$ and the smoothing term $\|\mathbf{Dm}\|_2^2$. Model $\mathbf{m}_0$ is a reference model and matrix $\mathbf{D}$ is a finite-difference derivative operator. Equation 15 is minimized iteratively using optimization algorithms and the optimization result satisfies

$$\mathbf{m}^* = \underset{\mathbf{m}}{\operatorname{argmin}}\, \chi_{reg} \tag{16}$$

The initial model for deterministic FWI is defined to be a laterally-constant velocity model and the velocity value at each depth is equal to the mean value of the Uniform prior distribution displayed in Figure 1b. This model is also used as the reference model $\mathbf{m}_0$. The L-BFGS (D. C. Liu & Nocedal, 1989) algorithm is used to minimize the misfit function in Equation 15. Two regularization factors are selected by trial and error. Figure 6c illustrates the obtained velocity model, which shows inconsistency with the true velocity model especially within the two boxes. This is because we start FWI from a relatively poor initial model and the algorithm finds a locally optimal solution due to the strong nonlinearity of the FWI problems. We perform a second test using a multiscale strategy (Bunks et al., 1995), in which a long wavelength (low wavenumber) velocity model is inverted first using low frequency waveform data with a dominant frequency of 5 Hz. The initial model in this inversion is the same as that used in the previous test. The inversion result is displayed in Figure 6a, where long wavelength components are recovered accurately compared to the true velocity model. We further treat this model as the starting point for the inversion using high frequency waveform data (with a 10 Hz dominant frequency), the same frequency data that were used to obtain the variational Bayesian inversion results displayed in Figure 3 and the deterministic FWI in Figure 6c. The corresponding inversion result is shown in Figure 6b, which resembles both the true velocity model and the inverted mean models from the four variational inversion results.

This test demonstrates that although deterministic FWI is able to provide an accurate velocity model, the method relies strongly on a good initial estimate and reference model (with correct long wavelength components) to avoid finding a locally optimal solution. In addition, the use of regularization terms injects arbitrary prior information into the inversion process that might not be true; their selection is usually ad hoc and requires many additional tests to be perform. Note that incorrect prior information might bias Bayesian inversion as well (X. Zhao & Curtis, 2024b). Nevertheless, regularization terms are not necessary for Bayesian FWI, as displayed in this work in which a non-informative Uniform prior distribution is used. This does not prioritize any model sample a priori. Finally, it is impossible to find all solutions that fit the observed waveform data to within data uncertainties and therefore to assess reasonable uncertainty information within such a deterministic inversion framework, which is particularly important for post inversion decision making processes as demonstrated in the next section. This shows the superiority of the proposed PSVI results and of Bayesian FWI, at additional computational cost.

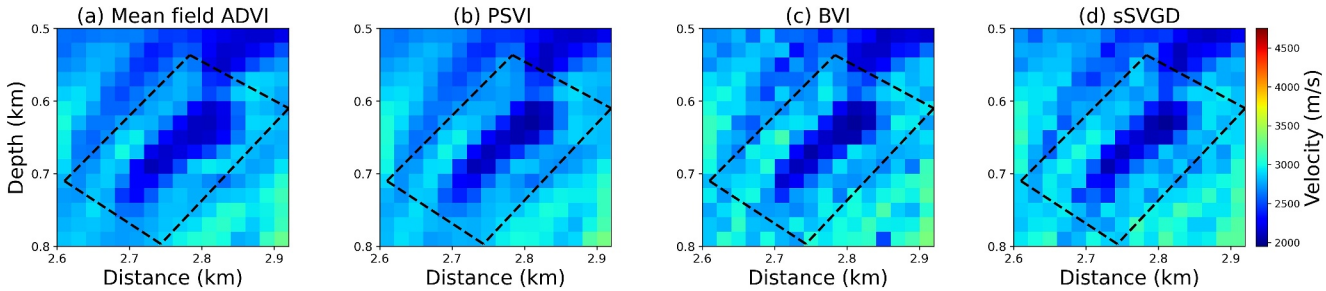

Figure 7. Mean velocity maps inside the white box in Figure 1a (corresponding to the true velocity map displayed in Figure 1d), obtained using (a) mean field ADVI, (b) PSVI, (c) boosting variational inference, and (d) stochastic Stein variational gradient descent. Black dashed boxes show the region where interrogation is performed.

4. Interrogating FWI Results

The objective of scientific investigations is typically to answer some specific and high-level questions. Examples of questions in the field of geophysics can be: *How large is a subsurface structure? Is this a good location for carbon capture and storage* (*CCS*)? Normally these questions are answered in a biased manner, using only one or a low number of possible models, and without evaluating uncertainties in the results. Interrogation theory provides a systematic way to obtain the least-biased answer to these questions (Arnold & Curtis, 2018). In this section, we solve two interrogation problems using the FWI results obtained above, to evaluate the potential practical value of the correlations estimated by PSVI.

Interrogation theory shows that the optimal answer a^* to a specific question Q that has a continuous space of possible answers is expressed by the following expectation term:

$$a^* = \mathbb{E}[T(\mathbf{m}|Q)] = \int_{\mathbf{m}} T(\mathbf{m}|Q)p(\mathbf{m}|\mathbf{d}_{obs})\, d\mathbf{m}, \tag{17}$$

where optimality is defined with respect to a squared utility (Arnold & Curtis, 2018). The expectation is taken with respect to the posterior distribution $p(\mathbf{m}|\mathbf{d}_{obs})$ of model parameter $\mathbf{m}$. Term $T(\mathbf{m}|Q)$ is a target function conditioned on the question Q of interest. It is defined to map the high dimensional model parameter $\mathbf{m}$ into a low dimensional target function value t in a target space $\mathbb{T}$, within which the question Q can be answered directly. In such cases the optimal answer in Equation 17 is simply the expectation or mean of the posterior target function.

4.1. Interrogation for Reservoir Size

Figure 7 shows the inverted mean models of the velocity structure within the white box in Figure 1a, obtained through (a) mean field ADVI, (b) PSVI, (c) BVI, and (d) sSVGD. In each figure, we observe a low velocity body at the center of the model section, outlined by a dashed black box. In this first example, we treat this low velocity zone as a reservoir and use interrogation theory to estimate its size.

Previously, volume-related questions were answered using seismic imaging results obtained from travel time tomographic inversion (X. Zhao et al., 2022) and FWI (X. Zhang & Curtis, 2022; X. Zhao & Curtis, 2024a). Following these studies, we define a target function $T(\mathbf{m}|Q)$ as the area of the largest continuous low velocity body; this function converts a high dimensional velocity model into a scalar value, representing the estimated reservoir area from a given posterior sample. Note that this process involves using a velocity threshold to distinguish between low and high velocities. We use the same data-driven method introduced in X. Zhao et al. (2022) to determine the least biased estimate of this threshold value. This involves selecting some cells that are almost definitely inside the low velocity anomaly, others that are almost definitely outside; we then choose the least-biased threshold value such that the expected probability of interior cells being below that value equals the expected probability of exterior cells being above that value, according to the posterior pdf. We are then able to calculate the target function for every posterior sample.

Figure 8 displays the posterior distributions of the target function (reservoir size) using the four inversion results obtained previously. In this synthetic test, the true reservoir area is precisely known from Figure 1d and is denoted by red lines in Figure 8. The optimal (least-biased) answer estimated from each inversion method corresponds to

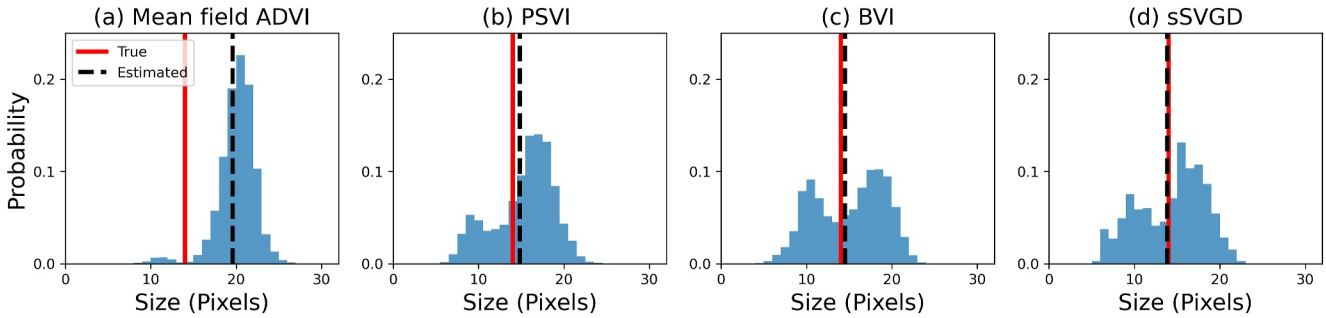

Figure 8. Posterior distributions of the low velocity reservoir size using full waveform inversion results obtained from (a) mean field ADVI, (b) PSVI, (c) boosting variational inference, and (d) stochastic Stein variational gradient descent, respectively. Red lines denote the true reservoir size, and black dashed lines denote the optimal size obtained using interrogation theory.

the mean value of the respective posterior target function (as per Equation 17), and is displayed by a dashed black line in each figure. As discussed in previous sections, mean field ADVI tends to underestimate posterior uncertainties and provides biased inversion results. We see that, the corresponding interrogation results in Figure 8a are also biased: the optimal answer shows a significant error and is far from the true answer, and indeed the true answer is even excluded from the posterior distribution of the estimated volume. By contrast, if we impose physically structured correlation information on model parameter, the optimal answer estimated by PSVI aligns closely with the true answer (Figure 8b). The posterior distribution of the target function also successfully captures bimodal uncertainties, similar to those obtained using BVI and sSVGD.

4.2. Interrogation for CO_2 Storage

In the second example, we apply the inversion results to answer a more realistic and practically interesting question. Assume the low velocity reservoir identified above is used in a CCS project and is injected with CO_2. The injection of CO_2 into a porous rock produces changes in petrophysical parameters of the rock, such as pore fluid phase and water saturation. These changes further result in variations in seismic response of a reservoir, such as seismic velocity. Leveraging the FWI results, we can use these variations to monitor the injected CO_2 in a subsurface CCS project by answering the question: *what is the total volume of* CO_2 *stored in this reservoir?*

For the characterization of changes in seismic velocity due to physical parameters related to CO_2, especially CO_2 saturation (S_{co_2}) in the reservoir, we first represent the P wave velocity v_p of a saturated rock using the bulk modulus K_{sat}, shear modulus G_{sat} and density ρ_{sat} of the rock by

$$v_p = \sqrt{\frac{K_{sat} + 4G_{sat}/3}{\rho_{sat}}} \tag{18}$$

The bulk modulus can be calculated using the Gassmann equation (Gassmann, 1951):

$$K_{sat} = K_d + \frac{\left(1 - \frac{K_d}{K_m}\right)^2}{\frac{\phi}{K_f} + \frac{1-\phi}{K_m} - \frac{K_d}{K_m^2}} \tag{19}$$

where ϕ is the porosity, and K_d, K_m and K_f are the bulk moduli of dry rock, solid matrix and pore fluid. The density of a saturated rock can be calculated as

$$\rho_{sat} = (1 - \phi)\rho_m + \phi\rho_f \tag{20}$$

where ρ_m and ρ_f are the densities of grain matrix and fluid, respectively. The shear modulus G_{sat} is not affected by fluid and only depends on the shear modulus of dry rock G_d

$$G_{sat} = G_d \tag{21}$$

Table 2
Rock Physics Parameters and Their Associated Standard Deviations (Uncertainties) Estimated From the Sleipner Field (Dupuy et al., 2017; Ghosh & Ojha, 2020)

Parameter	K_m (GPa)	K_d (GPa)	K_w (GPa)	K_{co_2} (GPa)	G_m (GPa)	G_d (GPa)	$\rho_m\ (kg/m^3)$	$\rho_w\ (kg/m^3)$	$\rho_{co_2}\ (kg/m^3)$	ϕ (%)
Mean value	39.3	2.56	2.31	0.08	44.8	8.1	2,664	1,030	700	0.3
Uncertainty	1.41	0.08	0.07	0.04	0.81	0.24	3	20	77	0.02

Assuming the reservoir is saturated by two distinct fluids, water and CO_2, the saturation values for water (S_w) and CO_2 (S_{co_2}) are constrained by the relation: $S_w + S_{co_2} = 1$. Then, the bulk modulus and density of fluid can be calculated using the mixing rules

$$\rho_f = S_w\rho_w + S_{co_2}\rho_{co_2} \tag{22}$$

$$K_f = S_w^e K_w + \left(1 - S_w^e\right) K_{co_2} \tag{23}$$

where ρ_w, ρ_{co_2}, K_w and K_{co_2} are the densities and bulk moduli of water and CO_2, and e is an empirical value (Brie et al., 1995). In this example, we use $e = 11$ as suggested by Kim et al. (2013). The injection of CO_2 into a reservoir alters the saturation values S_w and S_{co_2}, changing K_f and ρ_f, and thus also v_p through Equations 18–23. Therefore, we can estimate S_{co_2} using P wave velocity values obtained from FWI.

To simplify the problem, we assume that some of the aforementioned rock physics parameters follow Gaussian distributions. Their means and standard deviations are estimated from the Sleipner field (Dupuy et al., 2017; Ghosh & Ojha, 2020; Strutz & Curtis, 2024), as listed in Table 2. Given these parameters, we build a direct relationship between P wave velocity v_p and CO_2 saturation S_{co_2}. The results are depicted by the joint probability distribution of v_p and S_{co_2} displayed in Figure 9a. The red curve is the reference $v_p - S_{co_2}$ curve obtained using the mean values from Table 2. In Figure 9a, the posterior distribution of CO_2 saturation for any P-wave velocity value can be obtained. For example, Figures 9b and 9c illustrate two such posterior pdfs corresponding to velocity values of 2,045 m/s (solid white line in Figure 9a) and 1,840 m/s (dashed white line). In Figure 9 we observe that seismic velocity is sensitive to small CO_2 saturations (below 0.2) and is insensitive for larger S_{co_2} values (Kim et al., 2013).

In the previous interrogation example, we defined the largest continuous low velocity body as the reservoir of interest for a posterior velocity sample. For each grid cell within the identified reservoir, we substitute its velocity value into Figure 9a to obtain the posterior pdf of CO_2 saturation. Finally, the total (2D) CO_2 volume V_{co_2} stored in the reservoir can be calculated by

$$V_{co_2} = \sum V\phi S_{co_2} \tag{24}$$

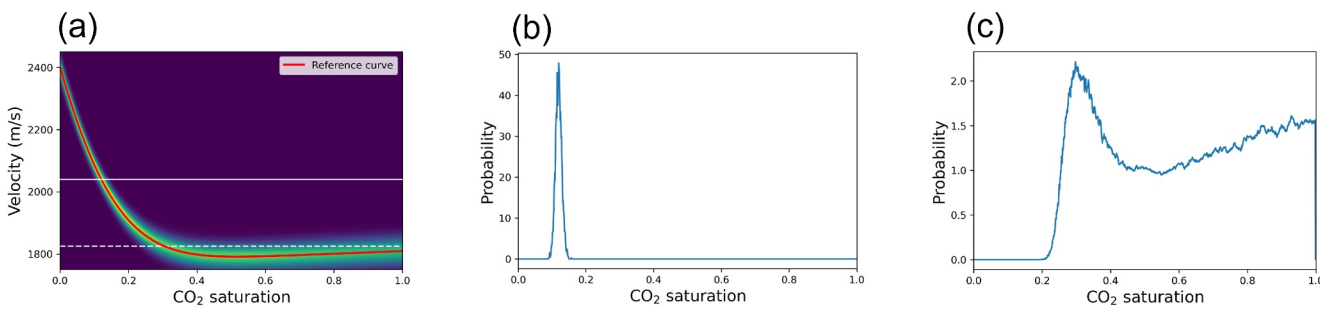

Figure 9. (a) Joint probability distribution of P wave velocity and CO_2 saturation given other parameters listed in Table 2. Red curve shows a one-to-one mapping between v_p and S_{co_2} obtained using the mean values in Table 2, and the color scale from red through green to dark blue represents the probability distribution of velocity, given any value of CO_2 and the Gaussian distributions defined in Table 2. (b) and (c) display the posterior distributions of CO_2 saturation for velocity values of 2,045 and 1,840 m/s, marked by solid and dashed white lines, respectively, in (a).

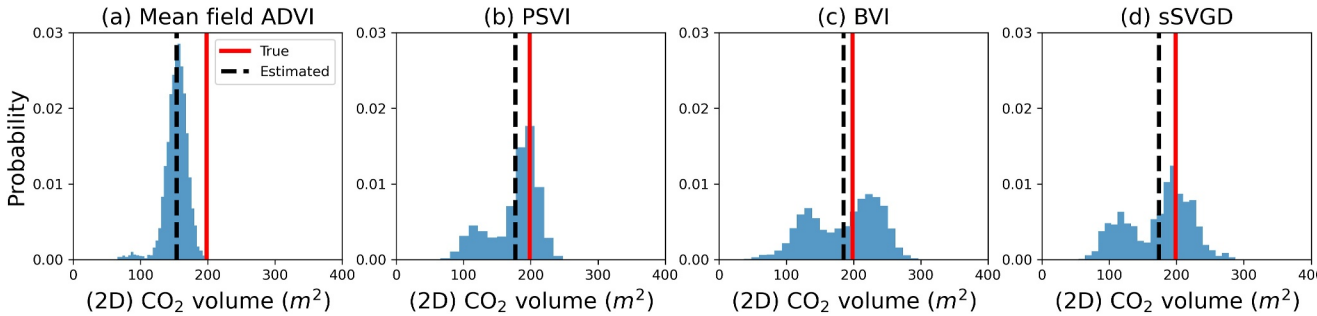

Figure 10. Posterior distributions of the (2D) CO_2 volume stored in the low velocity reservoir, calculated using (a) mean field ADVI, (b) PSVI, (c) boosting variational inference and (d) stochastic Stein variational gradient descent. Red lines denote the true CO_2 volume, and black dashed lines denote the least-biased CO_2 volume estimated using interrogation theory.

where V is the (2D) volume (i.e., area) of each grid cell in FWI, and the summation is taken over all grid cells within the reservoir. This defines the target function for this interrogation problem.

Figure 10 displays the posterior distributions of the estimated (2D) CO_2 volume obtained using different inversion methods. Similarly to the example estimating reservoir size displayed in Figure 8, mean field ADVI provides rather biased interrogation results since it tends to underestimate posterior uncertainties. In contrast, the other three methods provide similar (and so probably reasonably correct) posterior distributions with two distinct modes. The three estimated answers are close to the true value, which lies inside the high probability region of the posterior distributions.

In Figures 8 and 10, posterior histograms from mean field ADVI are hugely biased. In Figure 8, interrogation results from PSVI and BVI are similar and are slightly different from those from sSVGD; whereas in Figure 10, results from BVI and sSVGD are similar and are slightly different from the results obtained using PSVI. Theoretically, both BVI and sSVGD are able to approximate any posterior distribution up to any level of accuracy, even if the posterior distribution is multimodal which is possibly true for FWI problems. On the other hand, PSVI employs a transformed Gaussian distribution with a particular correlation structure, thus is less accurate for highly nonlinear problems with complex posterior pdf's (note that in algorithm development we often have to accept such a loss of generality in the solution when we wish to gain efficiency). Nevertheless, these three methods provide approximately consistent interrogation results, and are significantly better than those using mean field ADVI. We can therefore conclude from Figures 8 and 10 that uncertainty information provided by PSVI can be used to answer real-world questions correctly. Moreover, the non-zero background correlations ignored by PSVI (displayed in Figures 5c and 5d) are shown to be less important for post-inversion decision-making.

5. Discussion

PSVI can be considered as an intermediate approach between mean field ADVI and full rank ADVI (Kucukelbir et al., 2017). Mean field ADVI neglects all correlations to reduce computations and thus strongly underestimates posterior uncertainties. Full rank ADVI includes full correlation information between model parameters but is computationally intractable for high dimensional problems such as 2D or 3D FWI. PSVI, with its ability to capture structured correlations, strikes a balance between efficiency and accuracy. In the context of Bayesian FWI, where problems are often high dimensional and non-linear, PSVI offers improved inversion results while maintaining a computational cost comparable to mean field ADVI. For inverse problems with lower dimensionality such that modeling a full covariance matrix is affordable, full rank ADVI could be a more suitable choice. However, in three of these methods a transformed Gaussian distribution is used to approximate the posterior distribution. When dealing with inverse problems with strong multimodality, these Gaussian-based methods are not suitable as a Gaussian distribution might provide incorrect posterior uncertainties. For example, imagine the true posterior pdf is a mixture of two Gaussian components, using one single (transformed) Gaussian distribution cannot achieve an accurate approximation (an example is provided in Figure 1 in X. Zhao and Curtis (2024a)). X. Zhao et al. (2021) also showed that full rank ADVI provides biased inversion results for a travel time tomography problem. Nevertheless, in this work we introduce PSVI specifically for Bayesian FWI

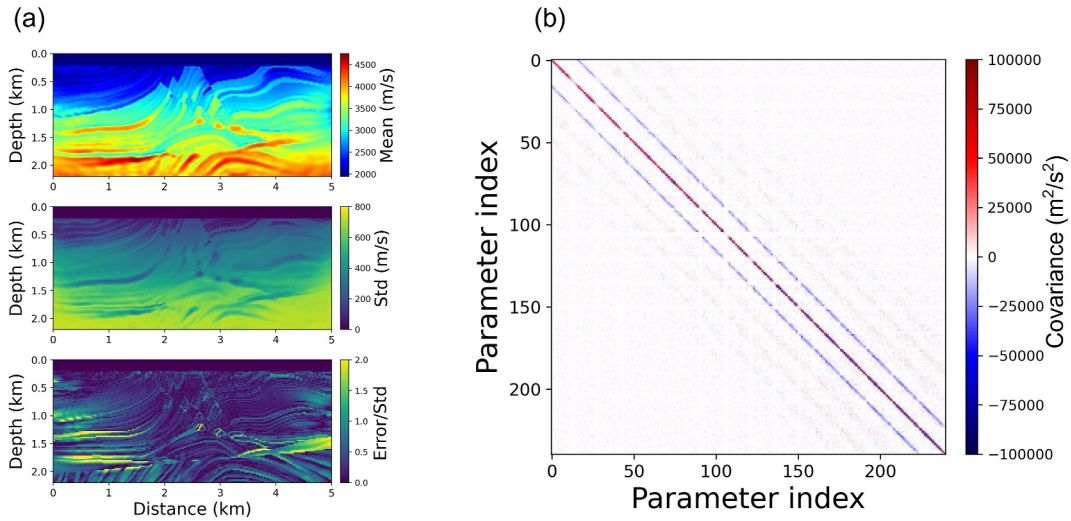

Figure 11. Inversion results obtained from physically structured variational inference using an 11 × 11 correlation kernel. (a) Mean, standard deviation and relative error maps. (b) Covariance matrix inside the white box in Figure 1a.

since it is so expensive that we almost cannot afford to use other methods, especially for 3D problems. Note that such a trade-off between efficiency and accuracy exists in all computational based problems.

As computational resources permit, it is advisable to use other variational methods such as normalizing flows (Rezende & Mohamed, 2015), BVI (F. Guo et al., 2016; Miller et al., 2017) or deterministic or stochastic SVGD (Gallego & Insua, 2018; Q. Liu & Wang, 2016) for strongly multimodal problems. These methods have shown effectiveness in solving multimodal problems, albeit at the cost of a larger number of forward simulations. The No Free Lunch theorem (Wolpert & Macready, 1997) can be paraphrased as: no method is better than any other method when averaged across all problems. There is therefore no possibility to find a "best" method in general. Nevertheless, individual classes of problems may have more or less efficient algorithms, so having a variety of methods allows for tailored decisions to be based on the nature of the problem to be addressed.

In the 2D FWI example, we use a 5 × 5 correlation kernel as displayed in Figure 1d. To investigate the impact of the correlation kernel size on inversion results, we conduct an additional test using an 11 × 11 kernel. The mean, standard deviation and relative error maps of the obtained posterior distribution are displayed in Figure 11a, which reveal nearly identical features, such as the continuous layers discussed previously, when compared to those obtained using the 5 × 5 correlation kernel (Figure 3b). Figure 11b displays the posterior covariance matrix, which as expected presents more non-zero off-diagonal covariance blocks than the 5 × 5 kernel (Figure 5b). The covariance magnitudes decay from the main diagonal block, and become relatively small from the second off-diagonal block. However, modeling these additional covariances requires more parameters to construct the matrix **L**. In addition, from Figures 5c and 5d, the covariance matrices calculated using BVI and sSVGD exhibit only one prominent off-diagonal block, probably because the non-linearity of FWI makes it challenging to capture a broader correlation structure with embedding prior knowledge of the type of structure sought. Therefore, we conclude that the 5 × 5 correlation kernel used above is a reasonable choice that trades off both accuracy and efficiency.

In real applications, if other prior knowledge about the subsurface structure is available (e.g., from seismic travel time tomography), we can design specific correlation kernels to capture target-oriented correlation information. Furthermore, the underlying principles of PSVI can be adapted to address temporal problems such as time-lapse (4D) seismic monitoring in which we might expect spatial regularity in the location of injected fluids, or in earthquake forecasting where correlations between seismic events over time might be captured effectively.

PSVI is not merely an extension of mean field ADVI as proposed by Kucukelbir et al. (2017). In fact it can be used to extend a variety of variational methods to enhance their accuracy and efficiency. For example, in BVI the physically structured approach in PSVI can replace diagonal Gaussians in modeling the Gaussian component distributions used in X. Zhao and Curtis (2024a). This substitution is likely to improve the accuracy of each

component while maintaining similar computational efficiency, potentially leading to a reduction in the required number of components and overall computational cost for BVI.

Similarly to BVI, PSVI produces an analytic posterior expression. Therefore, saving and loading inversion results, generating new posterior samples, and sharing the posterior distribution with others post inversion is simple (Scheiter et al., 2022). In addition, such an analytic expression can be used to vary prior information in Bayesian inference without requiring any further forward function evaluations using a variational prior replacement methodology (X. Zhao & Curtis, 2024b). The proposed method can also be extended to other general Gaussian-based methods such as Gaussian processes (Blatter et al., 2021; Ray, 2021; Ray & Myer, 2019; Valentine & Sambridge, 2020a, 2020b) and mixture density networks (Bishop, 1994; Bloem et al., 2023; Devilee et al., 1999; Earp & Curtis, 2020; Hansen & Finlay, 2022; Meier et al., 2007; Shahraeeni & Curtis, 2011; Shahraeeni et al., 2012), to capture desired correlation structures. Interestingly, special neural network structures are designed for the same purpose, such as the *coupling layer* (Dinh et al., 2015, 2017; Durkan et al., 2019; Y. Sun & Williamson, 2024; X. Zhang & Curtis, 2021b; X. Zhao et al., 2021) and the *autoregressive layer* (De Cao et al., 2019; Huang et al., 2018; Kingma et al., 2016; Levy et al., 2022; Papamakarios et al., 2017). However, they often come with a higher number of parameters to be optimized, making PSVI an attractive and practical choice.

Considering that solving the forward function in 2D FWI is not hugely expensive, we use a relatively smaller step size and more iterations during variational inversion to ensure that the optimization process has converged stably. Figure 2 illustrates that the negative ELBOs stop decreasing after 2,500–3,000 iterations, indicating that the full 5,000 iterations used here might be redundant. For higher dimensional problems such as 3D FWI, we can potentially use larger step sizes with fewer iterations, thereby optimizing the balance between computational resources and convergence speed.

Similarly to PSVI, Zunino and Mosegaard (2019) introduced an efficient method to solve large linear (or weakly non-linear, so linearizable) inverse problems, in which an approximate covariance matrix is constructed to reduce memory requirement and computational cost for calculating a full covariance matrix. Specifically, Zunino and Mosegaard (2019) used Kronecker products to decompose the covariance matrix, assuming that the forward problem is linear (linearizable) and that the posterior covariance matrix is separable (e.g., along different spatial directions). While conceptually this shares similarities with PSVI, direct combination of PSVI and the Kronecker product decomposition (Zunino & Mosegaard, 2019) in the context of FWI problems is not straightforward since (a) the forward problem (i.e., solving a wave equation) is a strongly non-linear operator, and (b) neither the forward operator nor the posterior covariance matrix are easily separable by Kronecker products. A combination of these two ideas might be possible in other applications, for example, in large scale 3D surface wave dispersion inversions, in which modeling a full covariance matrix is impractical and the forward operator is weakly non-linear and separable in horizontal and depth directions.

The two interrogation examples presented here underscore the significance of estimating accurate uncertainties, even if that demands a substantial increase in computational input. Biased uncertainty information (such as that provided by mean field ADVI) leads to incorrect answers about Earth properties. Therefore, while obtaining an accurate mean velocity model in Bayesian inversion, or just the best-fit model in deterministic inversion, may appear useful, they are far from sufficient for an unbiased and quantitative interpretation of the true Earth. The pursuit of not only precision in mean velocity models but also robust and reliable uncertainty estimates is important for a comprehensive understanding of subsurface structures.

In the first interrogation example, we estimated the size of a subsurface reservoir, where we use relative velocity values and classify them as either low or high based on a velocity threshold value (X. Zhao et al., 2022). In the second example, we take the absolute velocity values and convert them into CO_2 saturation estimates using a non-linear rock physics relationship. If the inversion is performed with higher frequency data, the inverted velocity values would be better constrained and become more accurate. Consequently, the posterior distribution of the estimated CO_2 volume can be improved. In this synthetic example we calculate the true CO_2 volume using exactly the same rock physics model as that used for interrogation, and obtain accurate interrogation results, proving the effectiveness of interrogation theory for answering real-world scientific questions and assessing the corresponding uncertainties. In real problems, more complicated and realistic situations should be considered to make the interrogation results practically reasonable and more accurate, such as interaction between rock and fluid, existence of multiple types of fluids, attenuation and anisotropy of subsurface media, potential chemical reaction between different fluids with rock, etc. All these factors will help us make a more accurate transform

from seismic velocity values to fluid properties, thereby improving interrogation accuracy. It is therefore promising that in future, 3D Bayesian FWI together with more advanced reservoir simulation and rock physics inversion techniques, can facilitate more sophisticated and realistic interrogation applications in subsurface CCS, or other subsurface projects. This comprehensive approach, enriched with full uncertainty assessments, could significantly contribute to our understanding and improve decision-making in the context of such endeavors.

6. Conclusion

In this work, we propose PSVI to perform 2D Bayesian FWI, in which a physical structure is imposed on the uncertainties in variational distributions based on prior information about imaging problem solutions. In our application, correlations between specific pairs of spatial locations are parametrized and inferred during inversion. Thus, we are able to capture the main correlations with a desired structure in a computationally efficient manner. We apply the proposed method together with three other variational methods: mean field automatic differentiation variational inference (ADVI), BVI and sSVGD, to a synthetic FWI example. This demonstrates that PSVI yields accurate first-order statistical information, including the mean and standard deviation maps as well as the marginal distributions, which are all consistent with those obtained using BVI and sSVGD. It also provides other second-order statistical information, specifically the posterior covariances. In addition, the obtained full uncertainty information is verified through the application of the inversion results to two post-inversion interrogation problems: one estimating a subsurface reservoir size and another estimating CO_2 volume in a CCS project. In our examples, PSVI exhibits nearly the same computational efficiency as mean field ADVI while enhancing the inversion accuracy significantly. This opens the possibility that 3D probabilistic FWI with full uncertainty estimation can be performed both efficiently and accurately.

Appendix A: Comparison Between ADVI and PSVI

In this Appendix, we compare the performance of mean field ADVI, PSVI and full rank ADVI, by considering another FWI test with a lower dimensionality. Figure A1 shows the true velocity model used in this test (obtained by downsampling and truncating the velocity model in Figure 1a used in the main text) with a grid size of 45 × 90 in horizontal and vertical directions. Similarly, source and receiver locations are marked by red stars and a white line. The remaining parameters used in this test are exactly the same as those used in the main text. To further reduce the dimensionality of the problem such that full rank ADVI is affordable (i.e., modeling a full covariance matrix is possible in terms of memory requirement and computational cost), we consider only a subset of the full imaging region inside the dashed red box in Figure A1, within which velocity values are inverted. This region

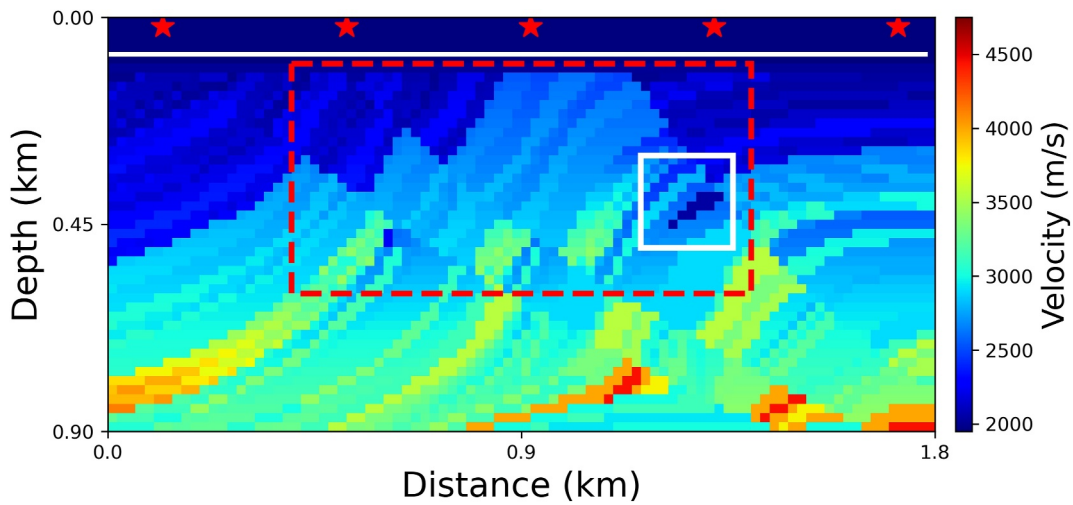

Figure A1. True velocity model used to compare mean field Automatic Differentiation Variational Inference (ADVI), physically structured variational inference and full rank ADVI in Appendix A. Source locations are indicated by red stars and the receiver line is marked by a white line. Red dashed box indicates the study region within which velocity values are inverted. White box displays regions within which the posterior covariance matrices obtained from different methods are compared in Figure A3.

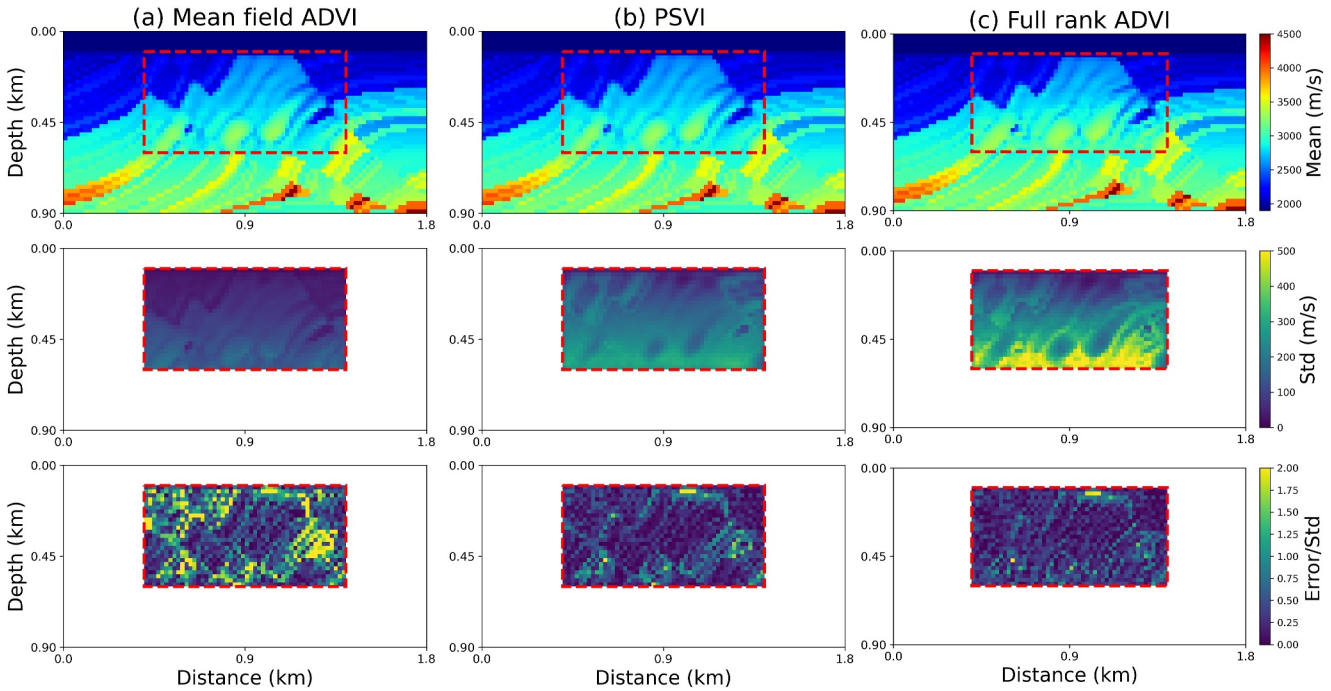

Figure A2. Mean (top row), standard deviation (middle row) and relative error (bottom row) of the posterior distribution obtained using (a) mean field ADVI, (b) physically structured variational inference and (c) full rank Automatic Differentiation Variational Inference, respectively. The relative error is the absolute error between the mean and true models divided by the corresponding standard deviation.

contains 25 × 50 grid cells in both directions. All other velocity values are set to be their true values during inversion.

Figure A2 displays the mean, standard deviation and the relative error maps of the posterior pdf's obtained using (a) mean field ADVI, (b) PSVI, and (c) full rank ADVI, respectively. The three mean velocity maps generally resemble each other and the true velocity model. As expected, mean field ADVI underestimates posterior uncertainties due to ignoring all correlations between model parameters, and thus the lowest standard deviations and the largest relative errors are present; results from PSVI and full rank ADVI show higher standard deviations and smaller relative errors, since we consider either dominant or all correlations of the posterior distribution. We observe that higher standard deviation values are present at the bottom of the imaging region in Figure A2c compared to those in Figure A2b, which shows that there is some information loss between PSVI and full rank ADVI. Another main factor for this is that running full rank ADVI with this dimensionality might be numerically unstable: recall from Equation 14 that, optimizing a full matrix $\mathbf{L}$ is equivalent to training a neural network with one fully connected layer and n neurons (n being the dimensionality of the problem and $n = 1250$ in this example). This is difficult to train and the results would possibly be overfit.

In Figure A3, we compare the posterior covariance matrices inside the white box in Figure A1, obtained from the three inversion results. No posterior correlation is observed from mean field ADVI, whereas main and full correlations are estimated using PSVI and full rank ADVI respectively, and are approximately consistent with each other to within the limitations of the PSVI parametrization. Additional off-diagonal correlations are observed in full rank ADVI indicating that a larger set of correlations could perhaps be included in the PSVI parametrization. Nevertheless, Figures A2 and A3 demonstrate that PSVI captures dominant correlations that are important to describe accurate posterior uncertainties and ignores those that are less important, to reduce computational cost.

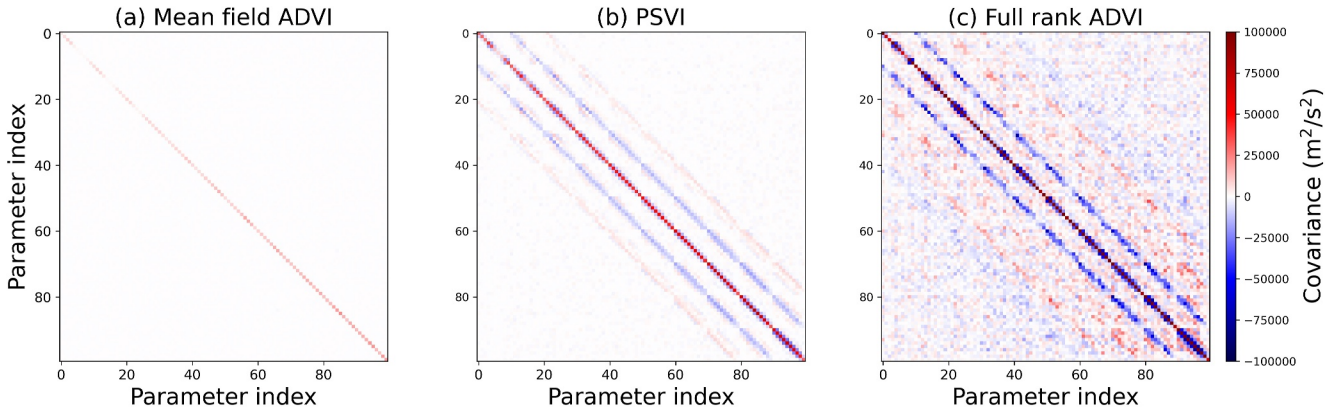

Figure A3. Covariance matrices for velocity values inside the white box in Figure A1, calculated using the inversion results from (a) mean field ADVI, (b) physically structured variational inference and (c) full rank Automatic Differentiation Variational Inference.

Data Availability Statement

Software used to perform variational inference can be found at Pyro website (Bingham et al., 2018) and in X. Zhang and Curtis (2023). Software used to perform Automatic Differentiation can be found at PyTorch website (Paszke et al., 2019).

Acknowledgments
The authors thank Edinburgh Imaging Project (EIP - https://blogs.ed.ac.uk/imaging/) sponsors (BP and TotalEnergies) for supporting this research. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising from this submission.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). TensorFlow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Ardizzone, L., Kruse, J., Wirkert, S., Rahner, D., Pellegrini, E. W., Klessen, R. S., et al. (2018). Analyzing inverse problems with invertible neural networks. *arXiv preprint arXiv:1808.04730*.

Arnold, R., & Curtis, A. (2018). Interrogation theory. *Geophysical Journal International*, *214*(3), 1830–1846. https://doi.org/10.1093/gji/ggy248

Asnaashari, A., Brossier, R., Garambois, S., Audebert, F., Thore, P., & Virieux, J. (2013). Regularized seismic full waveform inversion with prior model information. *Geophysics*, *78*(2), R25–R36. https://doi.org/10.1190/geo2012-0104.1

Bates, O., Guasch, L., Strong, G., Robins, T. C., Calderon-Agudo, O., Cueto, C., et al. (2022). A probabilistic approach to tomography and adjoint state methods, with an application to full waveform inversion in medical ultrasound. *Inverse Problems*, *38*(4), 045008. https://doi.org/10.1088/1361-6420/ac55ee

Berti, S., Aleardi, M., & Stucchi, E. (2023). A computationally efficient Bayesian approach to full-waveform inversion. *Geophysical Prospecting*, *72*(2), 580–603. https://doi.org/10.1111/1365-2478.13437

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., et al. (2018). Pyro: Deep universal probabilistic programming [Software]. *Journal of Machine Learning Research*. Retrieved from https://pyro.ai/

Bishop, C. M. (1994). *Mixture density networks*. Aston University.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Biswas, R., & Sen, M. K. (2022). Transdimensional 2D full-waveform inversion and uncertainty estimation. *arXiv preprint arXiv:2201.09334*.

Blatter, D., Ray, A., & Key, K. (2021). Two-dimensional Bayesian inversion of magnetotelluric data using trans-dimensional Gaussian processes. *Geophysical Journal International*, *226*(1), 548–563. https://doi.org/10.1093/gji/ggab110

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. https://doi.org/10.1080/01621459.2017.1285773

Bloem, H., Curtis, A., & Tetzlaff, D. (2023). Introducing conceptual geological information into Bayesian tomographic imaging. *Basin Research*, *36*(1), e12811. https://doi.org/10.1111/bre.12811

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Brie, A., Pampuri, F., Marsala, A., & Meazza, O. (1995). Shear sonic interpretation in gas-bearing sands. In *SPE Annual Technical Conference and Exhibition* (p. 701).

Bui-Thanh, T., Ghattas, O., Martin, J., & 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*(6), A2494–A2523. https://doi.org/10.1137/12089586x

Bunks, C., Saleck, F. M., Zaleski, S., & Chavent, G. (1995). Multiscale seismic waveform inversion. *Geophysics*, *60*(5), 1457–1473. https://doi.org/10.1190/1.1443880

Curtis, A., & Lomax, A. (2001). Prior information, sampling distributions, and the curse of dimensionality. *Geophysics*, *66*(2), 372–378. https://doi.org/10.1190/1.1444928

De Cao, N., Titov, I., & Aziz, W. (2019). Block neural autoregressive flow. *arXiv preprint arXiv:1904.04676*.

de Lima, P. D. S., Corso, G., Ferreira, M. S., & de Araújo, J. M. (2023). Acoustic full waveform inversion with Hamiltonian Monte Carlo method. *Physica A: Statistical Mechanics and its Applications*, *617*, 128618. https://doi.org/10.1016/j.physa.2023.128618

de Lima, P. D. S., Ferreira, M. S., Corso, G., & de Araújo, J. M. (2023). Bayesian time-lapse full waveform inversion using Hamiltonian Monte Carlo. *arXiv preprint arXiv:2311.02999*.

Devilee, R., Curtis, A., & Roy-Chowdhury, K. (1999). An efficient, probabilistic neural network approach to solving inverse problems: Inverting surface wave velocities for Eurasian crustal thickness. *Journal of Geophysical Research*, *104*(B12), 28841–28857. https://doi.org/10.1029/1999jb900273

Dhabaria, N., & Singh, S. C. (2024). Hamiltonian Monte Carlo based elastic full-waveform inversion of wide-angle seismic data. *Geophysical Journal International*, ggae112.

Dinh, L., Krueger, D., & Bengio, Y. (2015). NICE: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*.

Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2017). Density estimation using real NVP. *arXiv preprint arXiv:1605.08803*.

Dupuy, B., Ghaderi, A., Querendez, E., & Mezyk, M. (2017). Constrained AVO for CO_2 storage monitoring at Sleipner. *Energy Procedia*, *114*, 3927–3936. https://doi.org/10.1016/j.egypro.2017.03.1524

Durkan, C., Bekasov, A., Murray, I., & Papamakarios, G. (2019). Neural spline flows. In *Advances in neural information processing systems* (pp. 7509–7520).

Earp, S., & Curtis, A. (2020). Probabilistic neural network-based 2D travel-time tomography. *Neural Computing and Applications*, *32*(22), 17077–17095. https://doi.org/10.1007/s00521-020-04921-8

Ely, G., Malcolm, A., & Poliannikov, O. V. (2018). Assessing uncertainties in velocity models and images with a fast nonlinear uncertainty quantification method. *Geophysics*, *83*(2), R63–R75. https://doi.org/10.1190/geo2017-0321.1

Fang, Z., Da Silva, C., Kuske, R., & Herrmann, F. J. (2018). Uncertainty quantification for inverse problems with weak partial-differential-equation constraints. *Geophysics*, *83*(6), R629–R647. https://doi.org/10.1190/geo2017-0824.1

Fichtner, A., Kennett, B. L., Igel, H., & Bunge, H.-P. (2009). Full seismic waveform tomography for upper-mantle structure in the Australasian region using adjoint methods. *Geophysical Journal International*, *179*(3), 1703–1725. https://doi.org/10.1111/j.1365-246x.2009.04368.x

French, S., & Romanowicz, B. A. (2014). Whole-mantle radially anisotropic shear velocity structure from spectral-element waveform tomography. *Geophysical Journal International*, *199*(3), 1303–1327. https://doi.org/10.1093/gji/ggu334

Fu, X., & Innanen, K. A. (2022). A time-domain multisource Bayesian/Markov chain Monte Carlo formulation of time-lapse seismic waveform inversion. *Geophysics*, *87*(4), R349–R361. https://doi.org/10.1190/geo2021-0443.1

Gallego, V., & Insua, D. R. (2018). Stochastic gradient MCMC with repulsive forces. *arXiv preprint arXiv:1812.00071*.

Gassmann, F. (1951). Elastic waves through a packing of spheres. *Geophysics*, *16*(4), 673–685. https://doi.org/10.1190/1.1437718

Gebraad, L., Boehm, C., & Fichtner, A. (2020). Bayesian elastic full-waveform inversion using Hamiltonian Monte Carlo. *Journal of Geophysical Research: Solid Earth*, *125*(3), e2019JB018428. https://doi.org/10.1029/2019jb018428

Ghosh, R., & Ojha, M. (2020). Prediction of elastic properties within CO_2 plume at Sleipner field using AVS inversion modified for thin-layer reflections guided by uncertainty estimation. *Journal of Geophysical Research: Solid Earth*, *125*(11), e2020JB019782. https://doi.org/10.1029/2020jb019782

Gouveia, W. P., & Scales, J. A. (1998). Bayesian seismic waveform inversion: Parameter estimation and uncertainty analysis. *Journal of Geophysical Research*, *103*(B2), 2759–2779. https://doi.org/10.1029/97jb02933

Guo, F., Wang, X., Fan, K., Broderick, T., & Dunson, D. B. (2016). Boosting variational inference. *Advances in Neural Information Processing Systems*.

Guo, P., Visser, G., & Saygin, E. (2020). Bayesian trans-dimensional full waveform inversion: Synthetic and field data application. *Geophysical Journal International*, *222*(1), 610–627. https://doi.org/10.1093/gji/ggaa201

Hansen, T. M., & Finlay, C. C. (2022). Use of machine learning to estimate statistics of the posterior distribution in probabilistic inverse problems—An application to airborne EM data. *Journal of Geophysical Research: Solid Earth*, *127*(11), e2022JB024703. https://doi.org/10.1029/2022jb024703

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109. https://doi.org/10.2307/2334940

Huang, C.-W., Krueger, D., Lacoste, A., & Courville, A. (2018). Neural autoregressive flows. *arXiv preprint arXiv:1804.00779*.

Izzatullah, M., Ravasi, M., & Alkhalifah, T. (2023). Physics reliable frugal uncertainty analysis for full waveform inversion. *arXiv preprint arXiv:2305.07921*.

Khoshkholgh, S., Zunino, A., & Mosegaard, K. (2021). Informed proposal Monte Carlo. *Geophysical Journal International*, *226*(2), 1239–1248. https://doi.org/10.1093/gji/ggab173

Khoshkholgh, S., Zunino, A., & Mosegaard, K. (2022). Full-waveform inversion by informed-proposal Monte Carlo. *Geophysical Journal International*, *230*(3), 1824–1833. https://doi.org/10.1093/gji/ggac150

Kim, J., Nam, M. J., & Matsuoka, T. (2013). Estimation of CO_2 saturation during both CO_2 drainage and imbibition processes based on both seismic velocity and electrical resistivity measurements. *Geophysical Journal International*, *195*(1), 292–300. https://doi.org/10.1093/gji/ggt232

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., & Welling, M. (2016). Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems* (pp. 4743–4751).

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kotsi, M., Malcolm, A., & Ely, G. (2020a). Time-lapse full-waveform inversion using Hamiltonian Monte Carlo: A proof of concept. In *SEG Technical Program Expanded Abstracts 2020* (pp. 845–849). Society of Exploration Geophysicists.

Kotsi, M., Malcolm, A., & Ely, G. (2020b). Uncertainty quantification in time-lapse seismic imaging: A full-waveform approach. *Geophysical Journal International*, *222*(2), 1245–1263. https://doi.org/10.1093/gji/ggaa245

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research*, *18*(1), 430–474.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Levy, S., Laloy, E., & Linde, N. (2022). Variational Bayesian inference with complex geostatistical priors using inverse autoregressive flows. *Computers & Geosciences*, *171*, 105263. https://doi.org/10.1016/j.cageo.2022.105263

Liu, D. C., & Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, *45*(1), 503–528. https://doi.org/10.1007/bf01589116

Liu, Q., & Wang, D. (2016). Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in neural information processing systems* (pp. 2378–2386).

Lomas, A., Luo, S., Irakarama, M., Johnston, R., Vyas, M., & Shen, X. (2023). 3D probabilistic full waveform inversion: Application to Gulf of Mexico field data. In *84th EAGE Annual Conference & Exhibition* (Vol. 2023, pp. 1–5). https://doi.org/10.3997/2214-4609.202310720

Martin, G. S., Wiley, R., & Marfurt, K. J. (2006). Marmousi2: An elastic upgrade for Marmousi. *The Leading Edge*, *25*(2), 156–166. https://doi.org/10.1190/1.2172306

Meier, U., Curtis, A., & Trampert, J. (2007). Fully nonlinear inversion of fundamental mode surface waves for a global crustal model. *Geophysical Research Letters*, *34*(16), L16304. https://doi.org/10.1029/2007gl030989

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092. https://doi.org/10.1063/1.1699114

Miller, A. C., Foti, N. J., & Adams, R. P. (2017). Variational boosting: Iteratively refining posterior approximations. In *International conference on machine learning* (pp. 2420–2429).

Mosegaard, K., & Tarantola, A. (1995). Monte Carlo sampling of solutions to inverse problems. *Journal of Geophysical Research*, *100*(B7), 12431–12447. https://doi.org/10.1029/94jb03097

Papamakarios, G., Pavlakou, T., & Murray, I. (2017). Masked autoregressive flow for density estimation. In *Advances in neural information processing systems* (pp. 2338–2347).

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). Automatic differentiation in PyTorch. Retrieved from https://pytorch.org/docs/stable/notes/extending.html#extending-torch-autograd

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). PyTorch: An imperative style, high-performance deep learning library [Software]. *Advances in Neural Information Processing Systems*, *32*. Retrieved from https://pytorch.org/

Plessix, R.-E. (2006). A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. *Geophysical Journal International*, *167*(2), 495–503. https://doi.org/10.1111/j.1365-246x.2006.02978.x

Poliannikov, O. V., & Malcolm, A. E. (2016). The effect of velocity uncertainty on migrated reflectors: Improvements from relative-depth imaging. *Geophysics*, *81*(1), S21–S29. https://doi.org/10.1190/geo2014-0604.1

Ray, A. (2021). Bayesian inversion using nested trans-dimensional Gaussian processes. *Geophysical Journal International*, *226*(1), 302–326. https://doi.org/10.1093/gji/ggab114

Ray, A., Kaplan, S., Washbourne, J., & Albertin, U. (2018). Low frequency full waveform seismic inversion within a tree based Bayesian framework. *Geophysical Journal International*, *212*(1), 522–542. https://doi.org/10.1093/gji/ggx428

Ray, A., & Myer, D. (2019). Bayesian geophysical inversion with trans-dimensional Gaussian process machine learning. *Geophysical Journal International*, *217*(3), 1706–1726. https://doi.org/10.1093/gji/ggz111

Ray, A., Sekar, A., Hoversten, G. M., & Albertin, U. (2016). Frequency domain full waveform elastic inversion of marine seismic data from the Alba field using a Bayesian trans-dimensional algorithm. *Geophysical Journal International*, *205*(2), 915–937. https://doi.org/10.1093/gji/ggw061

Rezende, D. J., & Mohamed, S. (2015). Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*.

Sambridge, M., Gallagher, K., Jackson, A., & Rickwood, P. (2006). Trans-dimensional inverse problems, model comparison and the evidence. *Geophysical Journal International*, *167*(2), 528–542. https://doi.org/10.1111/j.1365-246x.2006.03155.x

Sambridge, M., & Mosegaard, K. (2002). Monte Carlo methods in geophysical inverse problems. *Reviews of Geophysics*, *40*(3), 3-1–3-29. https://doi.org/10.1029/2000rg000089

Scheiter, M., Valentine, A., & Sambridge, M. (2022). Upscaling and downscaling Monte Carlo ensembles with generative models. *Geophysical Journal International*, *230*(2), 916–931. https://doi.org/10.1093/gji/ggac100

Sen, M. K., & Roy, I. G. (2003). Computation of differential seismograms and iteration adaptive regularization in prestack waveform inversion. *Geophysics*, *68*(6), 2026–2039. https://doi.org/10.1190/1.1635056

Shahraeeni, M. S., & Curtis, A. (2011). Fast probabilistic nonlinear petrophysical inversion. *Geophysics*, *76*(2), E45–E58. https://doi.org/10.1190/1.3540628

Shahraeeni, M. S., Curtis, A., & Chao, G. (2012). Fast probabilistic petrophysical mapping of reservoirs from 3D seismic data. *Geophysics*, *77*(3), O1–O19. https://doi.org/10.1190/geo2011-0340.1

Siahkoohi, A., & Herrmann, F. J. (2021). Learning by example: Fast reliability-aware seismic imaging with normalizing flows. In *First International Meeting for Applied Geoscience & Energy* (pp. 1580–1585).

Siahkoohi, A., Rizzuti, G., & Herrmann, F. J. (2022). Deep Bayesian inference for seismic imaging with tasks. *Geophysics*, *87*(5), S281–S302. https://doi.org/10.1190/geo2021-0666.1

Siahkoohi, A., Rizzuti, G., Louboutin, M., Witte, P. A., & Herrmann, F. J. (2021). Preconditioned training of normalizing flows for variational inference in inverse problems. *arXiv preprint arXiv:2101.03709*.

Siahkoohi, A., Rizzuti, G., Orozco, R., & Herrmann, F. J. (2023). Reliable amortized variational inference with physics-based latent distribution correction. *Geophysics*, *88*(3), 1–137. https://doi.org/10.1190/geo2022-0472.1

Siahkoohi, A., Rizzuti, G., Witte, P. A., & Herrmann, F. J. (2020). Faster uncertainty quantification for inverse problems with conditional normalizing flows. *arXiv preprint arXiv:2007.07985*.

Smith, J. D., Ross, Z. E., Azizzadenesheli, K., & Muir, J. B. (2022). HypoSVI: Hypocentre inversion with Stein variational inference and physics informed neural networks. *Geophysical Journal International*, *228*(1), 698–710. https://doi.org/10.1093/gji/ggab309

Strutz, D., & Curtis, A. (2024). Variational Bayesian experimental design for geophysical applications: Seismic source location, amplitude versus offset inversion, and estimating CO_2 saturations in a subsurface reservoir. *Geophysical Journal International*, *236*(3), 1309–1331. https://doi.org/10.1093/gji/ggad492

Sun, L., Wang, L., Xu, G., & Wu, Q. (2023). A new method of variational Bayesian slip distribution inversion. *Journal of Geodesy*, *97*(1), 10. https://doi.org/10.1007/s00190-023-01701-9

Sun, Y., & Williamson, P. (2024). Invertible neural networks for uncertainty quantification in refraction tomography. *The Leading Edge*, *43*(6), 358–366. https://doi.org/10.1190/tle43060358.1

Tarantola, A. (1984). Inversion of seismic reflection data in the acoustic approximation. *Geophysics*, *49*(8), 1259–1266. https://doi.org/10.1190/1.1441754

Thurin, J., Brossier, R., & Métivier, L. (2019). Ensemble-based uncertainty estimation in full waveform inversion. *Geophysical Journal International*, *219*(3), 1613–1635. https://doi.org/10.1093/gji/ggz384

Treeby, B. E., & Cox, B. T. (2010). k-wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields. *Journal of Biomedical Optics*, *15*(2), 021314. https://doi.org/10.1117/1.3360308

Valentine, A. P., & Sambridge, M. (2020a). Gaussian process models—I. A framework for probabilistic continuous inverse theory. *Geophysical Journal International*, *220*(3), 1632–1647. https://doi.org/10.1093/gji/ggz520

Valentine, A. P., & Sambridge, M. (2020b). Gaussian process models—II. Lessons for discrete inversion. *Geophysical Journal International*, *220*(3), 1648–1656.

Virieux, J., & Operto, S. (2009). An overview of full-waveform inversion in exploration geophysics. *Geophysics*, *74*(6), WCC1–WCC26. https://doi.org/10.1190/1.3238367

Visser, G., Guo, P., & Saygin, E. (2019). Bayesian transdimensional seismic full-waveform inversion with a dipping layer parameterization. *Geophysics*, *84*(6), R845–R858. https://doi.org/10.1190/geo2018-0785.1

Wang, W., McMechan, G. A., & Ma, J. (2023). Re-weighted variational full waveform inversions. *Geophysics*, *88*(4), 1–61. https://doi.org/10.1190/geo2021-0766.1

Wang, Y., Zhou, H., Zhao, X., Zhang, Q., Zhao, P., Yu, X., & Chen, Y. (2019). Cu *Q*-RTM: A CUDA-based code package for stable and efficient *Q*-compensated reverse time migration. *Geophysics*, *84*(1), F1–F15. https://doi.org/10.1190/geo2017-0624.1

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82. https://doi.org/10.1109/4235.585893

Yin, Z., Orozco, R., Louboutin, M., & Herrmann, F. J. (2024). Wise: Full-waveform variational inference via subsurface extensions. *Geophysics*, *89*(4), 1–31. https://doi.org/10.1190/geo2023-0744.1

Zhang, C., Bütepage, J., Kjellström, H., & Mandt, S. (2018). Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(8), 2008–2026. https://doi.org/10.1109/tpami.2018.2889774

Zhang, X., & Curtis, A. (2020a). Seismic tomography using variational inference methods. *Journal of Geophysical Research: Solid Earth*, *125*(4), e2019JB018589. https://doi.org/10.1029/2019jb018589

Zhang, X., & Curtis, A. (2020b). Variational full-waveform inversion. *Geophysical Journal International*, *222*(1), 406–411. https://doi.org/10.1093/gji/ggaa170

Zhang, X., & Curtis, A. (2021a). Bayesian full-waveform inversion with realistic priors. *Geophysics*, *86*(5), 1–20. https://doi.org/10.1190/geo2021-0118.1

Zhang, X., & Curtis, A. (2021b). Bayesian geophysical inversion using invertible neural networks. *Journal of Geophysical Research: Solid Earth*, *126*(7), e2021JB022320. https://doi.org/10.1029/2021jb022320

Zhang, X., & Curtis, A. (2022). Interrogating probabilistic inversion results for subsurface structural information. *Geophysical Journal International*, *229*(2), 750–757. https://doi.org/10.1093/gji/ggab496

Zhang, X., & Curtis, A. (2023). VIP – Variational inversion package with example implementations of Bayesian tomographic imaging [Software]. *arXiv preprint arXiv:2310.13325*. Retrieved from https://github.com/xin2zhang/VIP

Zhang, X., Lomas, A., Zhou, M., Zheng, Y., & Curtis, A. (2023). 3-D Bayesian variational full waveform inversion. *Geophysical Journal International*, *234*(1), 546–561. https://doi.org/10.1093/gji/ggad057

Zhang, X., Nawaz, A., Zhao, X., & Curtis, A. (2021). An introduction to variational inference in geophysical inverse problems. *Advances in Geophysics*, *62*, 73–140. https://doi.org/10.1016/bs.agph.2021.06.003

Zhao, X., & Curtis, A. (2024a). Bayesian inversion, uncertainty analysis and interrogation using boosting variational inference. *Journal of Geophysical Research: Solid Earth*, *129*(1), e2023JB027789. https://doi.org/10.1029/2023jb027789

Zhao, X., & Curtis, A. (2024b). Variational prior replacement in Bayesian inference and inversion. *Geophysical Journal International*, *239*(2), 1236–1256. https://doi.org/10.1093/gji/ggae334

Zhao, X., Curtis, A., & Zhang, X. (2021). Bayesian seismic tomography using normalizing flows. *Geophysical Journal International*, *228*(1), 213–239. https://doi.org/10.1093/gji/ggab298

Zhao, X., Curtis, A., & Zhang, X. (2022). Interrogating subsurface structures using probabilistic tomography: An example assessing the volume of Irish Sea basins. *Journal of Geophysical Research: Solid Earth*, *127*(4), e2022JB024098. https://doi.org/10.1029/2022jb024098

Zhao, X., Zhou, H., Chen, H., & Wang, Y. (2020). Domain decomposition for large-scale viscoacoustic wave simulation using localized pseudo-spectral method. *IEEE Transactions on Geoscience and Remote Sensing*, 1–14.

Zhao, Z., & Sen, M. K. (2021). A gradient-based Markov chain Monte Carlo method for full-waveform inversion and uncertainty analysis. *Geophysics*, *86*(1), R15–R30. https://doi.org/10.1190/geo2019-0585.1

Zhdanov, M. S. (2002). *Geophysical inverse theory and regularization problems* (Vol. 36). Elsevier.

Zhu, H., Li, S., Fomel, S., Stadler, G., & Ghattas, O. (2016). A Bayesian approach to estimate uncertainty for full-waveform inversion using a priori information from depth migration. *Geophysics*, *81*(5), R307–R323. https://doi.org/10.1190/geo2015-0641.1

Zidan, A., Li, Y., & Cheng, A. (2022). Regularized seismic amplitude inversion via variational inference. *Geophysical Prospecting*, *70*(9), 1507–1527. https://doi.org/10.1111/1365-2478.13248

Zunino, A., Gebraad, L., Ghirotto, A., & Fichtner, A. (2023). HMCLab: A framework for solving diverse geophysical inverse problems using the Hamiltonian Monte Carlo method. *Geophysical Journal International*, *235*(3), 2979–2991. https://doi.org/10.1093/gji/ggad403

Zunino, A., & Mosegaard, K. (2019). An efficient method to solve large linearizable inverse problems under Gaussian and separability assumptions. *Computers & Geosciences*, *122*, 77–86. https://doi.org/10.1016/j.cageo.2018.09.005