Geophysical Journal International

Geophys. J. Int. (2023) **234,** 546–561 https://doi.org/10.1093/gji/ggad057
Advance Access publication 2023 February 08
GJI Seismology

3-D Bayesian variational full waveform inversion

Xin Zhang,[1,*] Angus Lomas,[2] Muhong Zhou,[2] York Zheng[2] and Andrew Curtis[1]

[1] *School of GeoSciences, University of Edinburgh, Edinburgh EH8 9XP, United Kingdom. E-mail:* X.Zhang2@ed.ac.uk
[2] *International Centre for Business and Technology, BP p.l.c., Sunbury, London, United Kingdom*

Accepted 2023 February 7. Received 2023 January 18; in original form 2022 October 12

SUMMARY

Seismic full-waveform inversion (FWI) provides high resolution images of the subsurface by exploiting information in the recorded seismic waveforms. This is achieved by solving a highly non-linear and non-unique inverse problem. Bayesian inference is therefore used to quantify uncertainties in the solution. Variational inference is a method that provides probabilistic, Bayesian solutions efficiently using optimization. The method has been applied to 2-D FWI problems to produce full Bayesian posterior distributions. However, due to higher dimensionality and more expensive computational cost, the performance of the method in 3-D FWI problems remains unknown. We apply three variational inference methods to 3-D FWI and analyse their performance. Specifically, we apply automatic differential variational inference (ADVI), Stein variational gradient descent (SVGD) and stochastic SVGD (sSVGD), to a 3-D FWI problem and compare their results and computational cost. The results show that ADVI is the most computationally efficient method but systematically underestimates the uncertainty. The method can therefore be used to provide relatively rapid but approximate insights into the subsurface together with a lower bound estimate of the uncertainty. SVGD demands the highest computational cost, and still produces biased results. In contrast, by including a randomized term in the SVGD dynamics, sSVGD becomes a Markov chain Monte Carlo method and provides the most accurate results at intermediate computational cost. We thus conclude that 3-D variational FWI is practically applicable, at least in small problems, and can be used to image the Earth's interior and to provide reasonable uncertainty estimates on those images.

Key words: Inverse theory; Probability distributions; Waveform inversion.

1 INTRODUCTION

A wide variety of academic studies and practical applications require that we interrogate the Earth's subsurface for answers to scientific questions. A common approach is to image subsurface properties in three dimensions using data recorded on the Earth's surface, and to interpret those images to address questions of interest. In order to provide well-justified and robust answers to such interrogation problems, it is necessary to assess the uncertainty in property estimates (Arnold & Curtis 2018).

Seismic full-waveform inversion (FWI) uses full seismic recordings to characterize properties of the Earth's interior, and can provide high resolution images of the subsurface (Tarantola 1984, 1988; Gauthier *et al.* 1986; Pratt 1999; Tromp *et al.* 2005; Fichtner *et al.* 2006; Plessix 2006). The method has been applied at industrial scale (Virieux & Operto 2009; Prieux *et al.* 2013; Warner *et al.* 2013), regional scale (Chen *et al.* 2007; Fichtner *et al.* 2009; Tape *et al.* 2009; Chen *et al.* 2015) and global scale (French & Romanowicz 2014; Bozdağ *et al.* 2016; Fichtner *et al.* 2018a; Lei *et al.* 2020).

Due to the non-linearity of relationships between model parameters and seismic waveforms, insufficient data coverage and noise in the data, FWI always has non-unique solutions and infinitely many sets of model parameters fit the data to within their uncertainty. It is therefore important to quantify uncertainties in the solution in order to better assess the reliability of inverted models (Tarantola 2005).

FWI problems are traditionally solved using optimization methods in which one seeks an optimal set of parameter values by minimizing the difference or misfit between observed data and model-predicted data. The strong non-linearity and non-uniqueness of the problem implies that a good starting model is required to avoid convergence to incorrect solutions (generally alternative modes or stationary points of the misfit function). Such models are not always available in practice. To alleviate this requirement a range of misfit functions that may reduce multimodality have been proposed (Luo & Schuster 1991; Gee & Jordan 1992; Fichtner *et al.* 2008; Brossier *et al.* 2010; Van Leeuwen & Mulder 2010; Bozdağ *et al.* 2011; Métivier *et al.* 2016; Warner & Guasch 2016;

[*] Now at: School of Engineering and Technology, China University of Geosciences, 100083 Beijing, China

Yuan *et al.* 2020; Sambridge *et al.* 2022). Nevertheless, none of the standard methods of solution using any of these misfit functions has been shown to allow accurate estimates of uncertainty to be made in realistic FWI problems.

Bayesian inference provides a different way to solve inverse problems and quantify uncertainties. The method uses Bayes' theorem to update a *prior* probability density function (pdf) with new information from the data to construct a so-called *posterior* probability density function. The prior pdf describes information available about the parameters of interest prior to the inversion (independently of the current data set), while the posterior pdf describes the resultant state of information after combining information in the prior pdf with information in the current data. In principle, Bayesian inference provides accurate estimates of uncertainty.

Markov chain Monte Carlo (McMC) is one method to characterize the posterior pdf which has been used widely in many fields. In McMC one constructs a set (chain) of successive samples generated from the posterior pdf by taking a structured random walk in parameter space (e.g. Brooks *et al.* 2011); those samples can thereafter be used to infer the values of useful statistics of that pdf (mean, standard deviation, etc.). The Metropolis–Hastings algorithm is one such method (Metropolis & Ulam 1949; Hastings 1970) and has been applied to many applications in geophysics, including gravity inversion (Mosegaard & Tarantola 1995; Bosch *et al.* 2006; Rossi 2017), vertical seismic profile inversion (Malinverno *et al.* 2000), surface wave dispersion inversion (Bodin *et al.* 2012; Shen *et al.* 2012; Young *et al.* 2013; Galetti *et al.* 2017; Zhang *et al.* 2018b), electrical resistivity inversion (Malinverno 2002; Galetti & Curtis 2018), electromagnetic inversion (Minsley 2011; Ray *et al.* 2013; Blatter *et al.* 2019), traveltime tomography (Bodin & Sambridge 2009; Galetti *et al.* 2015, 2017) and more recently FWI (Ray *et al.* 2017; Sen & Biswas 2017; Guo *et al.* 2020). However, the basic Metropolis–Hastings algorithm becomes computationally intractable in high dimensional space if the chain is attracted to individual misfit minima rather than exploring all possible such minima. To reduce this issue, more advanced McMC methods have been introduced to geophysics, such as Hamiltonian Monte Carlo (Duane *et al.* 1987; Fichtner *et al.* 2018b; Gebraad *et al.* 2020; Kotsi *et al.* 2020), stochastic Newton McMC (Martin *et al.* 2012; Zhao & Sen 2019), Langevin Monte Carlo (Roberts *et al.* 1996; Siahkoohi *et al.* 2020a) and parallel tempering (Hukushima & Nemoto 1996; Dosso *et al.* 2012; Sambridge 2013). However, the above studies mainly address 1-D or 2-D problems because of the high computational expense of moving to 3-D. Some studies have applied McMC methods to 3-D inverse problems including body wave traveltime tomography (Piana Agostinetti *et al.* 2015; Hawkins & Sambridge 2015; Burdick & Lekić 2017; Zhang *et al.* 2020b) and surface wave dispersion inversion (Zhang *et al.* 2018b, 2020a; Ryberg *et al.* 2022), but they require enormous computational cost even for small data sets. Thus, McMC methods are generally considered to be intractable for large data sets and high dimensionality, such as occurs in 3-D FWI problems.

Variational inference solves Bayesian inference problems in a different way: within a predefined family of (simplified) pdfs, the method seeks an optimal approximation to the posterior pdf by minimizing the difference between the approximating pdf and the posterior pdf. A typical metric used to measure this difference is the Kullback–Leibler (KL) divergence (Kullback & Leibler 1951). The method therefore solves an optimization problem rather than a stochastic sampling process as in McMC methods. As a result, in some classes of problems variational inference may be computationally more efficient than McMC methods and provide better scaling to higher dimensionality (Bishop 2006; Blei *et al.* 2017; Zhang *et al.* 2018a). The method can be applied to larger data sets by dividing the data set into small minibatches and using stochastic and distributed optimization methods (Robbins & Monro 1951; Kubrusly & Gravier 1973). In addition, the method can usually be parallelized at the individual sample level which makes the method even more efficient in real time by taking advantage of modern high performance computational facilities. By contrast, McMC methods cannot be parallelized at the sample level since each sample depends on the previous sample, and cannot use minibatches as these break the detailed balance condition that is required by common McMC methods (O'Hagan & Forster 2004).

In variational inference the choice of variational family is important as it determines the accuracy of the approximation and the complexity of the optimization problem. A good choice should be rich enough to approximate complex distributions and simple enough such that the optimization problem remains solvable. Different choices of variational families lead to a variety of specific methods. For example, a common choice is to use a *mean-field* approximation in which the parameters are assumed to be mutually independent (Bishop 2006; Blei *et al.* 2017). In geophysics the method has been applied to invert for geological facies distributions using seismic data (Nawaz & Curtis 2018, 2019; Nawaz *et al.* 2020). While often leading to highly efficient algorithms, this method usually requires bespoke mathematical derivations which restricts its applicability to a limited range of problems. Based on a Gaussian variational family, Kucukelbir *et al.* (2017) proposed a method called automatic differential variational inference (ADVI), which can be applied easily to general problems. For example, the method has been used to solve seismic travel time tomography (Zhang & Curtis 2020a) and earthquake slip inversion problems (Zhang & Chen 2022). A similar method has also been used to solve FWI problems in medical ultrasound (Bates *et al.* 2022).

By exploiting the properties of probability transformations, another set of methods has been proposed in which one optimizes a series of invertible transforms which convert a simple initial distribution to an arbitrary distribution that can approximate the posterior distribution (Rezende & Mohamed 2015; Tran *et al.* 2015; Liu & Wang 2016). Normalizing flow variational inference is one such method which applies a series of invertible and differential transforms (called flows) to an initial distribution; those flows are then optimized to produce an improved approximation to the posterior pdf (Rezende & Mohamed 2015). Normalizing flows have been demonstrated to be an efficient method in geophysical applications such as seismic tomography (Zhao *et al.* 2021) and image denoising (Siahkoohi *et al.* 2020b). However, the method becomes inefficient in very high dimensional space because of the computational cost required by large and flexible forms of flows. Stein variational gradient descent (SVGD) provides an alternative method that uses a set of particles (models) to represent the probability distribution. Those particles are iteratively updated by minimizing the KL-divergence so that in their final state their density approximates the posterior pdf (Liu & Wang 2016). The method has been applied to a range of geophysical applications, including seismic travel time tomography (Zhang & Curtis 2020a), earthquake location (Smith *et al.* 2022), hydrogeological inversion (Ramgraber *et al.* 2021) and 2-D FWI (Zhang & Curtis 2020b, 2021). However, none of these studies are comparable to a typical 3-D FWI problem in terms of dimensionality and computational cost, so the property of the method in 3-D FWI remains unknown.

In this study we explore the properties and efficiency of variational inference methods in 3-D FWI problems, including ADVI and SVGD. In addition, to reduce possible deficiency of SVGD in higher dimensionality (Ba *et al.* 2021) we introduce another method called stochastic SVGD (sSVGD: Gallego & Insua 2018) and compare the method with ADVI and SVGD. In Section 2, we first describe the basic concept of variational inference and then the ADVI, SVGD and sSVGD methods. In Section 3, we apply the suite of methods to a 3-D FWI problem and compare their results and computational costs. The aim of this study is to explore performance of those methods, to assess the computational requirements and to provide useful information for practitioners. Our results demonstrate that the 3-D variational FWI is practically feasible, at least for small problems, and so can be applied to image the Earth's subsurface and to provide uncertainty estimates on the results.

2 METHODS

2.1 Variational inference

Bayesian inference is the process of constructing a posterior probability density function $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$ of model parameters $\mathbf{m}$ given the observed data $\mathbf{d}_{\text{obs}}$, by updating a prior pdf with new information contained in the data. According to Bayes' theorem,

$$p(\mathbf{m}|\mathbf{d}_{\text{obs}}) = \frac{p(\mathbf{d}_{\text{obs}}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d}_{\text{obs}})}, \tag{1}$$

where $p(\mathbf{d}_{\text{obs}}|\mathbf{m})$ is the *likelihood* which describes the probability of observing data $\mathbf{d}_{\text{obs}}$ if model $\mathbf{m}$ was true, $p(\mathbf{m})$ represents the prior pdf which describes information that is known independently of the data, and $p(\mathbf{d}_{\text{obs}})$ is a normalization factor called the *evidence*.

Variational inference solves the above Bayesian inference problem using optimization. The method seeks an optimal approximation $q^*(\mathbf{m})$ to the posterior pdf $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$ within a predefined family of known probability distributions $Q = \{q(\mathbf{m})\}$ by minimizing the KL divergence between $q(\mathbf{m})$ and $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$:

$$q^*(\mathbf{m}) = \underset{q \in Q}{\arg\min}\, \mathrm{KL}[q(\mathbf{m})||p(\mathbf{m}|\mathbf{d}_{\text{obs}})]. \tag{2}$$

The KL divergence measures the difference between two probability distributions and can be expressed as:

$$\mathrm{KL}[q(\mathbf{m})||p(\mathbf{m}|\mathbf{d}_{\text{obs}})] = \mathrm{E}_q[\log q(\mathbf{m})] - \mathrm{E}_q[\log p(\mathbf{m}|\mathbf{d}_{\text{obs}})], \tag{3}$$

where the expectation is taken with respect to the distribution $q(\mathbf{m})$. The KL divergence is non-negative and only equals zero when $q(\mathbf{m}) = p(\mathbf{m}|\mathbf{d}_{\text{obs}})$ (Kullback & Leibler 1951). Expanding the posterior pdf using eq. (1), the KL divergence becomes:

$$\begin{aligned}\mathrm{KL}[q(\mathbf{m})||p(\mathbf{m}|\mathbf{d}_{\text{obs}})] = {} & \mathrm{E}_q[\log q(\mathbf{m})] - \mathrm{E}_q[\log p(\mathbf{m}, \mathbf{d}_{\text{obs}})] \\ & + \log p(\mathbf{d}_{\text{obs}}).\end{aligned} \tag{4}$$

The evidence term $\log p(\mathbf{d}_{\text{obs}})$ is computationally intractable because it requires evaluation of a high dimensional integral for which the computation scales exponentially with the number of parameters. We therefore rearrange eq. (4) to obtain the evidence lower bound (ELBO):

$$\begin{aligned}\mathrm{ELBO}[q] &= \log p(\mathbf{d}_{\text{obs}}) - \mathrm{KL}[q(\mathbf{m})||p(\mathbf{m}|\mathbf{d}_{\text{obs}})] \\ &= \mathrm{E}_q[\log p(\mathbf{m}, \mathbf{d}_{\text{obs}})] - \mathrm{E}_q[\log q(\mathbf{m})].\end{aligned} \tag{5}$$

Since the KL divergence is non-negative, the above equation defines a lower bound for the evidence $\log p(\mathbf{d}_{\text{obs}})$. In addition because the evidence $\log p(\mathbf{d}_{\text{obs}})$ is a constant for a given problem, minimizing the KL-divergence is equivalent to maximizing the ELBO. Consequently, variational inference in eq. (2) can also be expressed as:

$$q^*(\mathbf{m}) = \underset{q \in Q}{\arg\max}\, \mathrm{ELBO}[q(\mathbf{m})]. \tag{6}$$

In variational inference, the choice of the variational family Q is important because it determines the accuracy of the approximation as well as the complexity of the optimization problem. Different methods can be developed depending on different choices of the family. In the following sections we describe a set of different methods: ADVI, SVGD and sSVGD and compare these methods in the application of 3-D FWI.

2.2 Automatic differential variational inference

ADVI is a variational method based on a Gaussian variational family (Kucukelbir *et al.* 2017). Gaussians are defined on the entire set of real numbers and in reality model parameters often have hard constrains (for example, seismic velocity is greater than zero), so in ADVI we first transform those constrained parameters into an unconstrained space using an invertible transform $T : \theta = T(\mathbf{m})$. In this space the joint probability $p(\mathbf{m}, \mathbf{d}_{\text{obs}})$ becomes:

$$p(\theta, \mathbf{d}_{\text{obs}}) = p(\mathbf{m}, \mathbf{d}_{\text{obs}})|det\mathbf{J}_{T^{-1}}(\theta)|, \tag{7}$$

where $\mathbf{J}_{T^{-1}}(\theta)$ is the Jacobian matrix of the inverse of T and $|\cdot|$ denotes absolute value. Define a Gaussian variational family

$$q(\theta; \boldsymbol{\zeta}) = N(\theta|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{8}$$

where $\boldsymbol{\zeta}$ represents variational parameters, that is the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. Although a full covariance matrix can be used for small size problems, it becomes computationally intractable for very high dimensional space (as in 3-D FWI). We therefore use a factorized (mean-field) Gaussian variational approximation:

$$q(\theta; \boldsymbol{\zeta}) = N(\theta|\boldsymbol{\mu}, \mathrm{diag}(\exp(\boldsymbol{\omega})^2)) \tag{9}$$

where we have reparametrized the standard deviation using $\boldsymbol{\sigma} = \exp(\boldsymbol{\omega})$ to ensure that each parameter of $\boldsymbol{\sigma}$ is positive. Note that because we neglect the correlation information between different parameters, the approximation obtained by minimizing the KL divergence systematically underestimates the marginal variance as illustrated in Fig. 1(a) (Bishop 2006).

With the above definition the variational problem in eq. (6) can be written as:

$$\begin{aligned}\boldsymbol{\zeta}^* &= \underset{\boldsymbol{\zeta}}{\arg\max}\, \mathrm{ELBO}[q(\theta; \boldsymbol{\zeta})] \\ &= \underset{\boldsymbol{\zeta}}{\arg\max}\, \mathrm{E}_q\left[\log p(T^{-1}(\theta), \mathbf{d}_{\text{obs}}) + \log|det\mathbf{J}_{T^{-1}}(\theta)|\right] \\ &\quad - \mathrm{E}_q[\log q(\theta)]\,.\end{aligned} \tag{10}$$

This optimization problem can be solved by using gradient ascent methods. As shown in Kucukelbir *et al.* (2017) the gradients of the ELBO with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\omega}$ are:

$$\begin{aligned}\nabla_{\boldsymbol{\mu}}\mathrm{ELBO} = {} & \mathrm{E}_{N(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}\left[\nabla_{\mathbf{m}}\log p(\mathbf{m}, \mathbf{d}_{\text{obs}})\nabla_{\boldsymbol{\theta}}T^{-1}(\theta)\right. \\ & \left. + \nabla_{\theta}\log|det\mathbf{J}_{T^{-1}}(\theta)|\right]\end{aligned} \tag{11}$$

$$\begin{aligned}\nabla_{\boldsymbol{\omega}}\mathrm{ELBO} = {} & \mathrm{E}_{N(\boldsymbol{\eta}|\mathbf{0},\mathbf{I})}\left[(\nabla_{\mathbf{m}}\log p(\mathbf{m}, \mathbf{d}_{\text{obs}})\nabla_{\theta}T^{-1}(\theta)\right. \\ & \left. + \nabla_{\theta}\log|det\mathbf{J}_{T^{-1}}(\theta)|)\boldsymbol{\eta}^{\mathrm{T}}\mathrm{diag}(\exp(\boldsymbol{\omega}))\right] + \mathbf{1},\end{aligned} \tag{12}$$

where $\boldsymbol{\eta}$ is a random variable generated from a standard Normal distribution $N(\mathbf{0}, \mathbf{I})$. The expectations in the above equations can

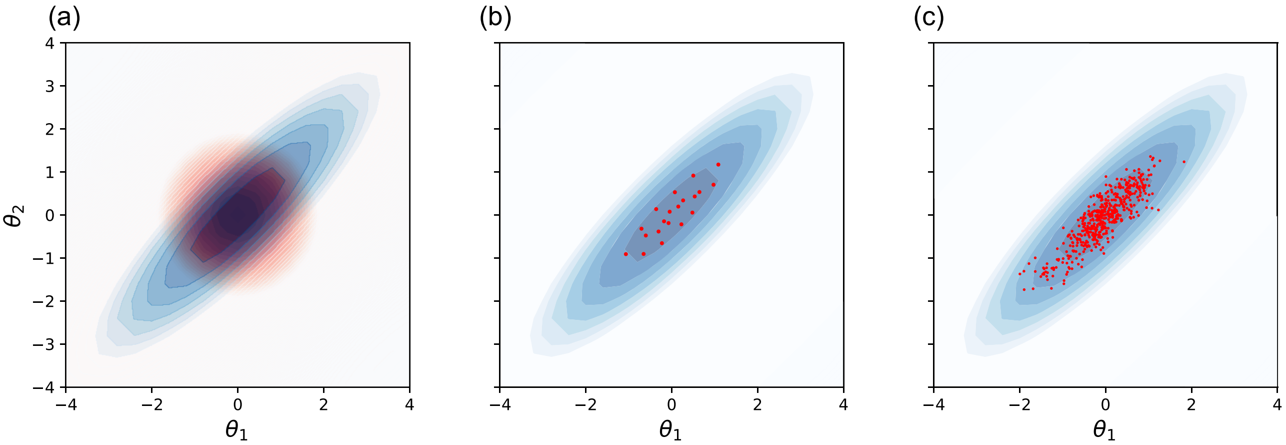

Figure 1. (a) The posterior distribution (red) obtained using ADVI with a mean-field approximation, and the samples obtained using (b) SVGD and (c) sSVGD in the case of a bivariate Gaussian distribution (blue). For both SVGD and sSVGD 20 particles are used.

be estimated by Monte Carlo (MC) integration, which in practice only requires a low number of samples because the optimization is usually performed over many iterations so that statistically the gradients will lead to convergence towards the correct solution (Kucukelbir *et al.* 2017). The variational problem in eq. (10) can therefore be solved by using gradient ascent methods. The final approximation $q^*(\mathbf{m})$ is obtained by transforming $q^*(\boldsymbol{\theta})$ back to the original space. For the transform T, we use a commonly used logarithmic transform (Team *et al.* 2016; Zhang & Curtis 2020a)

$$\begin{aligned}\theta_i &= T(m_i) = \log(m_i - a_i) - \log(b_i - m_i)\\ m_i &= T^{-1}(\theta_i) = a_i + \frac{(b_i - a_i)}{1 + exp(-\theta_i)},\end{aligned} \tag{13}$$

where m_i represents ith parameter in the original constrained space, θ_i is the transformed variable in the unconstrained space, a_i and b_i are the lower and upper bound on m_i, respectively. Although ADVI can generate biased results as we discussed above, it has been demonstrated to be a computationally efficient method compared to SVGD (Zhang & Curtis 2020a; Zhao *et al.* 2021). For this reason, we explore its properties in 3-D FWI problems.

2.3 Stein variational gradient descent

SVGD is a variational method which uses a set of samples (called particles) whose density represents the approximation pdf q. The method iteratively updates those particles by minimizing the KL divergence so that the final set of particles are distributed according to the posterior distribution (Liu & Wang 2016). Since the distribution of a set of particles is in principle entirely flexible, this method can provide more accurate results than ADVI (Zhang & Curtis 2020a). Define the set of particles as $\{\mathbf{m}_i\}$ where $\mathbf{m}_i$ is a d-dimensional parameter vector. SVGD uses a smooth transform $T(\mathbf{m}_i) = \mathbf{m}_i + \epsilon\boldsymbol{\phi}(\mathbf{m}_i)$ to update each particle, where $\boldsymbol{\phi} = [\phi_1, ..., \phi_d]$ is a smooth vector function that describes the perturbation direction and ϵ is the magnitude of the perturbation. Assume T is invertible and define $q_T(\mathbf{m})$ as the transformed probability distribution of pdf $q(\mathbf{m})$. The gradient of the KL-divergence between q_T and the posterior pdf p with respect to ϵ can be computed as (Liu & Wang 2016):

$$\nabla_\epsilon \mathrm{KL}[q_T||p]\,|_{\epsilon=0} = -\mathrm{E}_q\left[trace\left(A_p\boldsymbol{\phi}(\mathbf{m})\right)\right], \tag{14}$$

where A_p is the Stein operator defined by $A_p\boldsymbol{\phi}(\mathbf{m}) = \nabla_{\mathbf{m}}\log p(\mathbf{m})\boldsymbol{\phi}(\mathbf{m})^T + \nabla_{\mathbf{m}}\boldsymbol{\phi}(\mathbf{m})$. Eq. (14) ensures that by maximizing the right-hand expectation we obtain the steepest descent direction of the KL-divergence, and consequently the KL divergence can be minimized by iteratively stepping a small distance in that direction.

The optimal $\boldsymbol{\phi}^*$ that maximize the expectation in eq. (14) can be found by using kernel functions. Say $x, y \in X$ and define a mapping ψ from X to a space where an inner product $\langle , \rangle$ is defined (called a Hilbert space); a *kernel* is a function that satisfies $k(x, y) = \langle\psi(x), \psi(y)\rangle$. Assume a kernel $k(\mathbf{m}', \mathbf{m})$, the optimal $\boldsymbol{\phi}^*$ can be expressed as (Liu & Wang 2016):

$$\boldsymbol{\phi}^* \propto \mathrm{E}_{\{\mathbf{m}'\sim q\}}[A_p k(\mathbf{m}', \mathbf{m})]. \tag{15}$$

Since we use particles $\{\mathbf{m}_i\}$ to represent q, the expectation can be approximated using the particles mean. The KL divergence can therefore be minimized by iteratively applying the transform $T(\mathbf{m}) = \mathbf{m} + \epsilon\boldsymbol{\phi}^*(\mathbf{m})$ to a set of initial particles $\{\mathbf{m}_i^0\}$:

$$\begin{aligned}\boldsymbol{\phi}_l^*(\mathbf{m}) &= \frac{1}{n}\sum_{j=1}^{n}\left[k(\mathbf{m}_j^l, \mathbf{m})\nabla_{\mathbf{m}_j^l}\log p(\mathbf{m}_j^l|\mathbf{d}_{\mathrm{obs}}) + \nabla_{\mathbf{m}_j^l}k(\mathbf{m}_j^l, \mathbf{m})\right]\\ \mathbf{m}_i^{l+1} &= \mathbf{m}_i^l + \epsilon^l\boldsymbol{\phi}_l^*(\mathbf{m}_i^l)\end{aligned} \tag{16}$$

where l denotes the lth iterations, n is the number of particles and ϵ^l is the step size. If the step size $\{\epsilon^l\}$ is sufficiently small then the transform T is invertible, and the process converges to the posterior pdf asymptotically as the number of particles tends to infinity.

For the kernel function we use a commonly used radial basis function (RBF)

$$k(\mathbf{m}, \mathbf{m}') = \exp\left[-\frac{\|\mathbf{m} - \mathbf{m}'\|^2}{2h^2}\right], \tag{17}$$

where h is a scale factor which intuitively controls the interaction intensity between different particles based on their distances apart. As suggested by several studies (Liu & Wang 2016; Zhang & Curtis 2020a, b), we choose h to be $\tilde{d}/\sqrt{2\log n}$ where $\tilde{d}$ is the median of pairwise distances between all particles. This choice ensures that the contribution from each particle $\mathbf{m}_i$'s own gradient is balanced by the influence from all other particles as $\sum_{j\neq i} k(\mathbf{m}_i, \mathbf{m}_j) \approx n\exp(-\frac{1}{2h^2}\tilde{d}^2) = 1$. Note that for the RBF kernel, the second term of $\boldsymbol{\phi}^*$ in eq. (16) becomes $\sum_j \frac{\mathbf{m}-\mathbf{m}_j}{\sigma^2}k(\mathbf{m}_j, \mathbf{m})$ which drives the particle $\mathbf{m}$ away from its neighbouring particles when the kernel takes high values. This second term therefore acts as a repulsive force which prevents the particles from collapsing to a single mode, whereas the first term consists of kernel weighted gradients which drives the particles towards high probability areas. An example of the particles obtained using SVGD in the case of a bivariate Gaussian distribution is shown in Fig. 1(b).

In Geophysics, SVGD has been demonstrated to be an efficient method for a rang of applications (Zhang & Curtis 2020a, b, 2021; Ramgraber *et al.* 2021; Zhao *et al.* 2021; Smith *et al.* 2022; Ahmed *et al.* 2022). In this study, we explore its applicability in 3-D FWI. As in previous studies (Zhang & Curtis 2020b; Zhang *et al.* 2021), in order to handle hard constraints of seismic velocity, we transform seismic velocity into an unconstrained space using eq. (13) and perform SVGD in that space. The final seismic velocities are obtained by transforming the particles back to the original space.

2.4 Stochastic SVGD

Although SVGD has been applied to many different applications (Gong *et al.* 2019; Zhang & Curtis 2020a, b; Pinder *et al.* 2020), the method can provide biased results and is known to underestimate variance for high dimensional problems because of the finite number of particles and the practical limitation of computational cost (Ba *et al.* 2021). In order to further improve accuracy of the method, efforts have been made to bridge the gap between variational inference and McMC methods. sSVGD is one such algorithm which turns SVGD into a Markov chain by adding a Gaussian noise term to the dynamics (Gallego & Insua 2018). By doing this one can start collecting many samples that represent the posterior pdf after a burn-in period instead of having to use a large number of particle from the beginning. In addition, the method guarantees asymptotic convergence to the posterior pdf as the number of iterations tends to infinity, which standard SVGD with a finite number of particles cannot achieve.

To introduce the sSVGD algorithm, we start from a stochastic differential equation (SDE):

$$\mathrm{d}\mathbf{z} = \mathbf{f}(\mathbf{z})\mathrm{d}t + \sqrt{2\mathbf{D}(\mathbf{z})}\mathrm{d}\mathbf{W}(t), \tag{18}$$

where $\mathbf{f}(\mathbf{z})$ is called the *drift*, $\mathbf{W}(t)$ is a Wiener process, and $\mathbf{D}(\mathbf{z})$ is a positive semidefinite diffusion matrix. Generally all continuous Markov processes can be expressed as a SDE of the above form. If we denote the posterior distribution as $p(\mathbf{z})$, Ma *et al.* (2015) proposed a SDE that converges to the distribution $p(\mathbf{z})$

$$\mathbf{f}(\mathbf{z}) = [\mathbf{D}(\mathbf{z}) + \mathbf{Q}(\mathbf{z})]\,\nabla\log p(\mathbf{z}) + \Gamma(\mathbf{z}), \tag{19}$$

where $\mathbf{Q}(\mathbf{z})$ is a skew-symmetric curl matrix, and $\Gamma_i(\mathbf{z}) = \sum_{j=1}^{d}\frac{\partial}{\partial \mathbf{z}_j}(\mathbf{D}_{ij}(\mathbf{z}) + \mathbf{Q}_{ij}(\mathbf{z}))$ is a correction term which amends the bias.

If we discretize eq. (18) with eq. (19) using the Euler–Maruyama discretization, we obtain a practical algorithm:

$$\begin{aligned}\mathbf{z}_{t+1} = \mathbf{z}_t + \epsilon_t\,[(\mathbf{D}\,(\mathbf{z}_t) + \mathbf{Q}(\mathbf{z}_t))\,\nabla\log p(\mathbf{z}_t) + \Gamma(\mathbf{z}_t)]\\ +N(\mathbf{0}, 2\epsilon_t\mathbf{D}(\mathbf{z}_t)),\end{aligned} \tag{20}$$

where $N(\mathbf{0}, 2\epsilon_t\mathbf{D}(\mathbf{z}_t))$ represents a Gaussian distribution. The gradient $\nabla\log p(\mathbf{z}_t)$ can be computed using full data, or Uniformly randomly selected minibatch data subsets which results in a stochastic gradient. In either case the above process converges to the posterior distribution asymptotically as $\epsilon_t \to 0$ and $t \to \infty$ (Ma *et al.* 2015). Matrix $\mathbf{D}(\mathbf{z})$ and $\mathbf{Q}(\mathbf{z})$ can be adjusted to obtain faster convergence to the posterior distribution. For example, by setting $\mathbf{D} = \mathbf{I}$ and $\mathbf{Q} = \mathbf{0}$ one obtains the stochastic gradient Langevin dynamics algorithm (Welling & Teh 2011). If we augment the state space $\mathbf{z}$ with a moment term $\mathbf{x}$ to obtain an augmented space $\bar{\mathbf{z}} = (\mathbf{z}, \mathbf{x})$, and set $\mathbf{D} = \mathbf{0}$ and $\mathbf{Q} = \begin{pmatrix}\mathbf{0} & -\mathbf{I}\\ \mathbf{I} & \mathbf{0}\end{pmatrix}$, the stochastic Hamiltonian Monte Carlo (HMC) method can be derived (Chen *et al.* 2014).

For the set of particles $\{\mathbf{m}_i\}$ defined in the above section we can construct an augmented space $\mathbf{z} = (\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_n) \in R^{nd}$ by concatenating n particles, and use eq. (20) to obtain a valid sampler that runs multiple (n) interacted chains:

$$\begin{aligned}\mathbf{z}_{t+1} = \mathbf{z}_t + \epsilon_t[(\mathbf{D}(\mathbf{z}_t) + \mathbf{Q}(\mathbf{z}_t))\nabla\log p(\mathbf{z}_t) + \Gamma(\mathbf{z}_t)]\\ +N(\mathbf{0}, 2\epsilon_t\mathbf{D}(\mathbf{z}_t)),\end{aligned} \tag{21}$$

where $\mathbf{D}, \mathbf{Q} \in R^{nd\times nd}$ and $\nabla\log p, \Gamma \in R^{nd}$. Define a matrix $\mathbf{K}$

$$\mathbf{K} = \frac{1}{n}\begin{bmatrix} k(\mathbf{m}_1, \mathbf{m}_1)\mathbf{I}_{d\times d} & \dots & k(\mathbf{m}_1, \mathbf{m}_n)\mathbf{I}_{d\times d}\\ \vdots & \ddots & \vdots \\ k(\mathbf{m}_n, \mathbf{m}_1)\mathbf{I}_{d\times d} & \dots & k(\mathbf{m}_n, \mathbf{m}_n)\mathbf{I}_{d\times d}\end{bmatrix}, \tag{22}$$

where $k(\mathbf{m}_i, \mathbf{m}_j)$ is a kernel function and $\mathbf{I}_{d\times d}$ is an identity matrix. According to the definition of kernel functions, the matrix $\mathbf{K}$ is positive definite (Gallego & Insua 2018). The standard SVGD algorithm in eq. (16) can now be expressed in matrix form as

$$\mathbf{z}_{t+1} = \mathbf{z}_t + \epsilon_t[\mathbf{K}\nabla\log p(\mathbf{z}_t) + \nabla\cdot\mathbf{K}] \tag{23}$$

which shows that SVGD can be regarded as a special case of eq. (21) with $\mathbf{D}_\mathbf{K} = \mathbf{K}$, $\mathbf{Q}_\mathbf{K} = \mathbf{0}$ and no noise term. By including the noise term, we construct a stochastic gradient McMC method with SVGD gradients, which we call stochastic SVGD:

$$\mathbf{z}_{t+1} = \mathbf{z}_t + \epsilon_t[\mathbf{K}\nabla\log p(\mathbf{z}_t) + \nabla\cdot\mathbf{K}] + N(\mathbf{0}, 2\epsilon_t\mathbf{K}). \tag{24}$$

According to the discussion above, this process converges to the posterior distribution $p(\mathbf{z}) = \prod_{i=1}^{n} p(\mathbf{m}_i|\mathbf{d}_{\mathrm{obs}})$ asymptotically. Note that when the number of particles is large enough, the noise term would be tiny according to eq. (22). Consequently in such case the method produces the same results as standard SVGD.

In order to use eq. (24) to sample the posterior distribution, we need to draw samples from the Gaussian distribution $N(\mathbf{0}, 2\epsilon_t\mathbf{K})$. This requires computing the lower triangular Cholesky decomposition of the $nd \times nd$ matrix $\mathbf{K}$, which can be computationally expensive. To compute the noise term efficiently, we define a block-diagonal matrix $\mathbf{D}_\mathbf{K}$

$$\mathbf{D}_\mathbf{K} = \frac{1}{n}\begin{bmatrix}\overline{\mathbf{K}} & & \\ & \ddots & \\ & & \overline{\mathbf{K}}\end{bmatrix}, \tag{25}$$

where $\overline{\mathbf{K}}$ is a $n \times n$ matrix with $\overline{\mathbf{K}}_{ij} = k(\mathbf{m}_i, \mathbf{m}_j)$. Note that with this definition, $\mathbf{D}_\mathbf{K}$ can be constructed from $\mathbf{K}$ using $\mathbf{D}_\mathbf{K} = \mathbf{PKP}^{\mathrm{T}}$

where $\mathbf{P}$ is a permutation matrix

$$\mathbf{P} = \left[\begin{array}{cccccccccc}
1 & & & & & & & & & \\
 & & & 1 & & & & & & \\
 & & & & & & \ddots & & & \\
 & & & & & & & 1 & & \\
\hline
 & 1 & & & & & & & & \\
 & & & & 1 & & & & & \\
 & & & & & & \ddots & & & \\
 & & & & & & & & 1 & \\
\hline
 & \ddots & & & \ddots & & \ddots & & \ddots & \\
\hline
 & & 1 & & & & & & & \\
 & & & & & 1 & & & & \\
 & & & & & & \ddots & & & \\
 & & & & & & & & & 1
\end{array}\right]. \tag{26}$$

The action of this permutation matrix on a vector $\mathbf{z}$ rearranges the order of the vector from the basis where the particles are listed sequentially to that where the first coordinates of all particles are listed, then the second, etc. The noise term $\boldsymbol{\eta}$ can therefore be generated using

$$\begin{aligned}
\boldsymbol{\eta} &\sim N(\mathbf{0}, 2\epsilon_t \mathbf{K}) \\
&\sim \sqrt{2\epsilon_t}\mathbf{P}^{\mathrm{T}}\mathbf{P}N(\mathbf{0}, \mathbf{K}) \\
&\sim \sqrt{2\epsilon_t}\mathbf{P}^{\mathrm{T}}N(\mathbf{0}, \mathbf{D_K}) \\
&\sim \sqrt{2\epsilon_t}\mathbf{P}^{\mathrm{T}}\mathbf{L_{D_K}}N(\mathbf{0}, \mathbf{I}),
\end{aligned} \tag{27}$$

where $\mathbf{L_{D_K}}$ is the lower triangular Cholesky decomposition of matrix $\mathbf{D_K}$. Given that $\mathbf{D_K}$ is a block-diagonal matrix, decomposition $\mathbf{L_{D_K}}$ can be calculated easily as we only need to calculate the lower triangular Cholesky decomposition of matrix $\overline{\mathbf{K}}$. Since in practice the number of particles n is usually modest, evaluating the noise term is computationally negligible. We can now use eq. (24) to generate samples from the posterior distribution. An example of the samples obtained using sSVGD in the case of a bivariate Gaussian distribution is shown in Fig. 1c.

3 RESULTS

We apply the above suite of methods to an acoustic 3-D FWI problem. The true model is chosen to be a part of the 3-D overthrust model (Fig. 2a, Aminzadeh 1997), which is discretized using a regular 101 × 101 × 63 grid of cells with 50 m spacing. We deploy 81 sources (red dots in Fig. 2a) and 10 201 receivers (yellow dots in Fig. 2a) at the surface with regular spacings of 500 and 50 m, respectively. The waveform data are calculated using the time-domain finite difference method with a 2–10 Hz Ormsby wavelet (Ryan 1994). Gradients of the likelihood function with respect to velocities are computed using the adjoint method (Tarantola 1988; Tromp *et al.* 2005; Fichtner *et al.* 2006; Plessix 2006; Liu & Gu 2012).

We represent available prior information by a Uniform distribution over an interval width of 2.5 km s^{-1} at each depth (Fig. 2b). Fig. 3 shows a set of cross sections ($Y = 1$ km, 2.5 km and 4 km) of the true model and an example model generated from the prior distribution. For the likelihood function we assume that a Gaussian distribution with a diagonal covariance matrix can be used to represent uncertainties on the waveform data:

$$p(\mathbf{d}_{\mathrm{obs}}|\mathbf{m}) \propto \exp\left[-\frac{1}{2}\sum_i \left(\frac{d_i^{\mathrm{obs}} - d_i(\mathbf{m})}{\sigma_i}\right)^2\right], \tag{28}$$

where i denotes the index of time samples and σ_i is the standard deviation of that data point. In this study we set σ_i to be 2 percent of the median of the maximum amplitude of each seismic trace.

For ADVI we set the initial Gaussian distribution in the unconstrained space to be a standard Normal distribution $N(\boldsymbol{\theta}|\mathbf{0}, \mathbf{I})$, and update the distribution using the ADAM algorithm (Kingma & Ba 2014) for 1000 iterations after which point the average misfit across Monte Carlo samples ceases to decrease. To reduce the computational cost, we compute the gradients in eqs (11) and (12) using minibatch data from 36 sources which are randomly selected from the total of 81 sources. At each iteration the gradients are calculated using four Monte Carlo samples. The final Gaussian distribution is transformed back to the original space, from which we generate 2000 samples to visualize the results.

For SVGD we generate 400 particles from the prior distribution (an example is shown in Fig. 3), and transform them to an unconstrained space using eq. (13). Those particles are then updated using eq. (15) for 1000 iterations after which point the average misfit across particles ceases to decrease. Similarly to above the gradients in eq. (15) are calculated using minibatch data from 36 sources. The final particles are transformed back to the original space.

For sSVGD we start from 20 particles that are generated from the prior distribution, and transform them to the unconstrained space as in SVGD. Those particles are then updated (sampled) using eq. (24) for 4000 iterations with a burn-in period of 2000. To reduce the memory and storage cost, we only retain every fourth sample after the burn-in period. This results in a total of 10 000 samples, which are transformed back to the original space to calculate statistics of the estimated posterior pdf. At each iteration the gradients are also calculated using minibatch data from 36 sources.

3.1 Model comparison

Fig. 4 shows the mean, standard deviation and the relative error computed using $|\mathbf{m}^{mean} - \mathbf{m}^{true}|/\boldsymbol{\sigma}$ where $\boldsymbol{\sigma}$ is the standard deviation, obtained using ADVI, displayed on the same cross sections as in Fig. 3. In the shallow part (depth $Z < 1.5$ km) the mean model shows similar structure to the true model. For example, overthrusted high velocity structures can be observed clearly in the mean model. Over the same depth range the standard deviation model shows similar features to the mean model. A similar phenomenon has been observed in a range of previous studies (Gebraad *et al.* 2020; Zhang & Curtis 2020b, 2021). At greater depths $Z > 1.5$ km the mean model deviates from the true model. This is probably because of the lower sensitivity caused by the short source–receiver offset offered by our acquisition geometry. This is also supported by high uncertainties across the same area. The relative error shows that the deviation of the mean model from the true model is larger than three standard deviations at depth and on both sides, which suggests that the uncertainty is clearly underestimated there. This underestimation is likely caused by the mean-field approximation we have used in ADVI (see Fig. 1a).

Fig. 5 shows the results obtained using SVGD. Overall the results show similar mean and standard deviation structures to those obtained using ADVI. For example, the mean model shows similar features to the true model in the shallow parts, and deviates from the true model at greater depths. The standard deviation also shows similar features to the mean model across the shallow part and higher uncertainties at greater depths. Note that the magnitude of the standard deviations is generally higher than those obtained using ADVI, which again shows the limitation of the mean-field

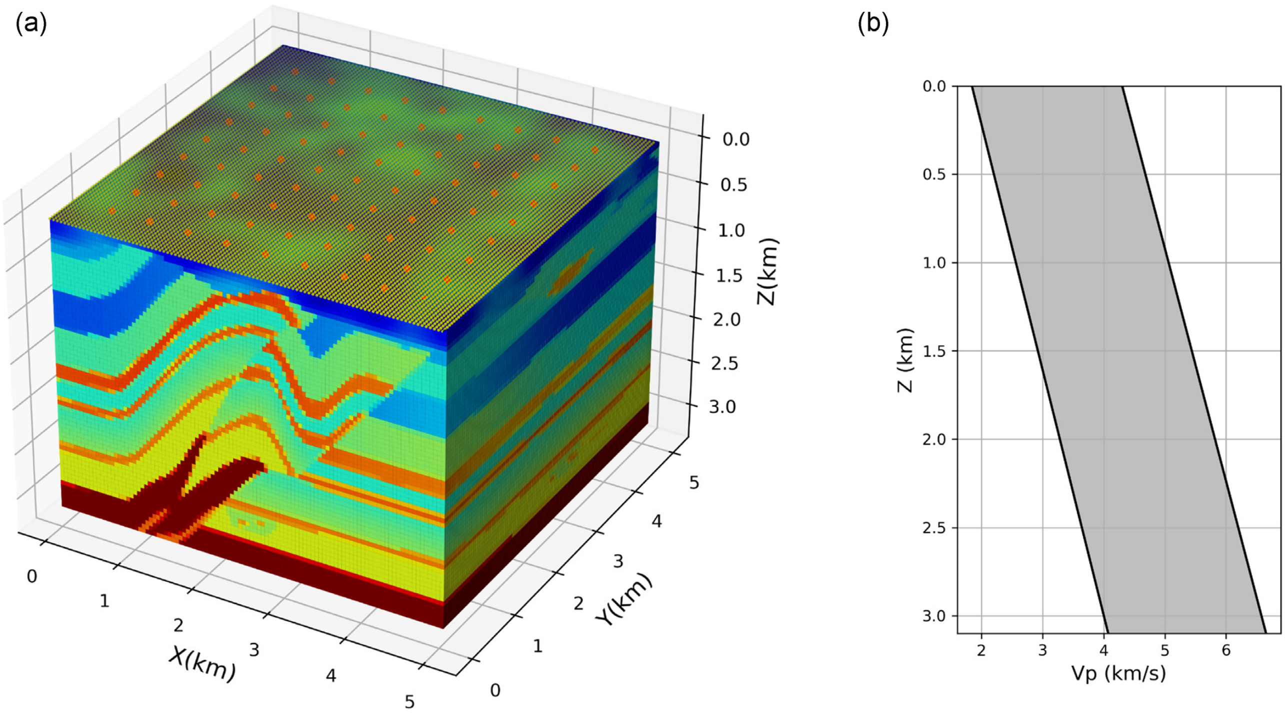

Figure 2. (a) True velocity model and acquisition geometry used in this study. Surface sources and receivers are denoted using red and yellow dots respectively. (b) Prior distribution used in the inversion: a uniform distribution with a width of 2.5 km s^{-1} at each depth.

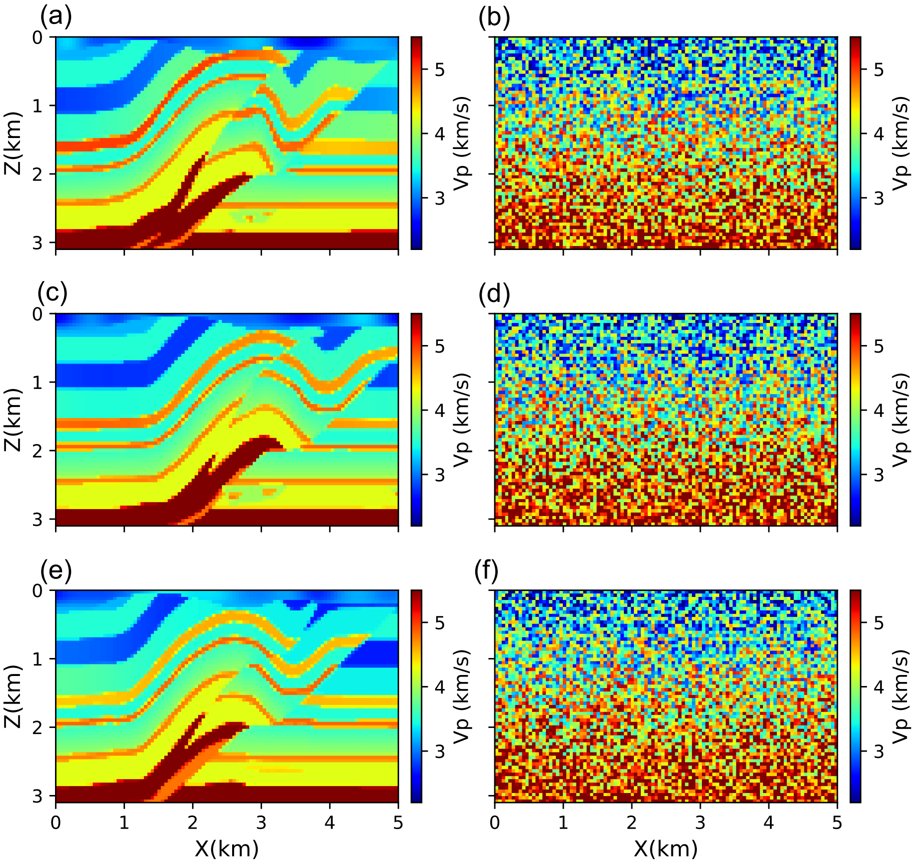

Figure 3. The true model (left-hand column) and an initial particle (right-hand column) at cross sections of $Y = 1$ km (a and b), 2.5 km (c and d) and 4 km (e and f), respectively.

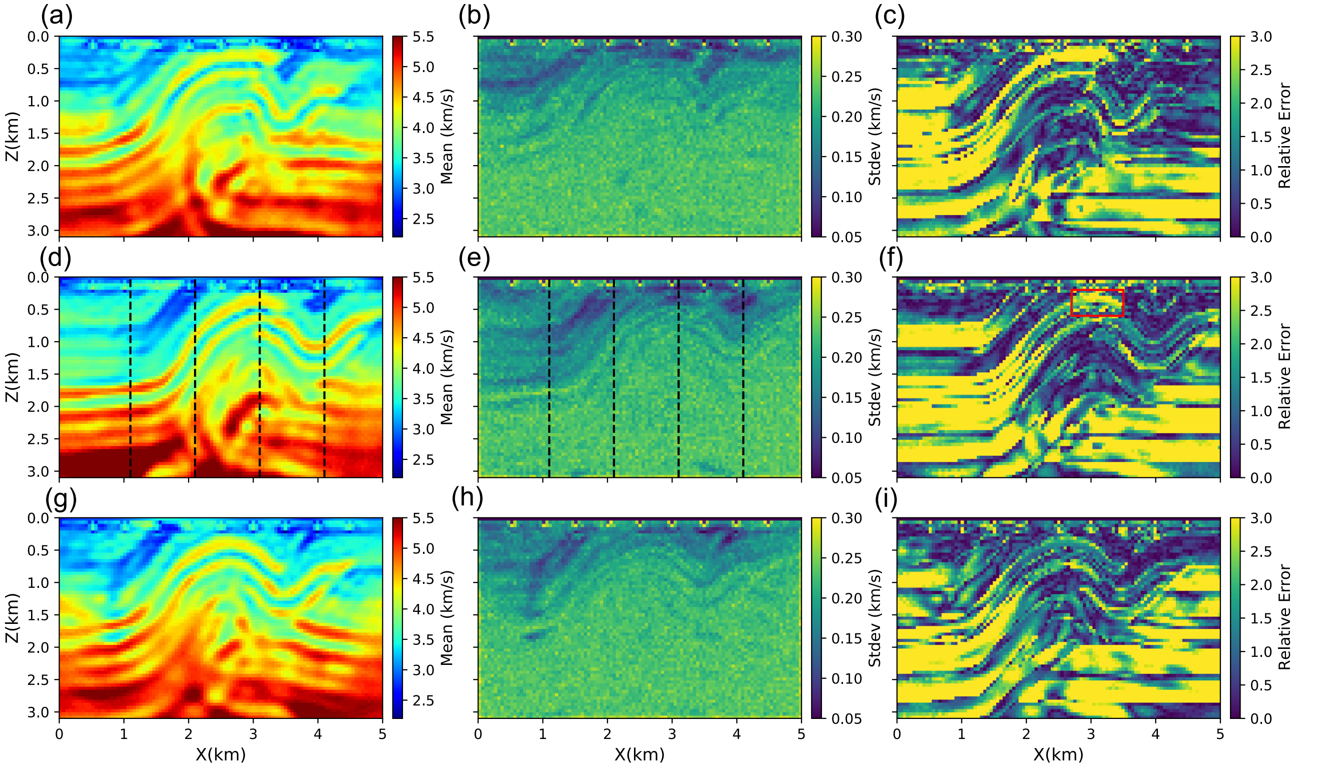

Figure 4. The mean (a, d and g), standard deviation (b, e and h) and relative error (c, f and i) obtained using ADVI over the same cross sections as in Fig. 3. The relative error is computed using $|\mathbf{m}^{mean} - \mathbf{m}^{true}|/\boldsymbol{\sigma}$ where $\boldsymbol{\sigma}$ is the standard deviation. Black dashed lines denote the well log locations referred to in the main text. Red box highlights a volume in which the true velocities are outside of the support (positive probabilities) of the prior distribution.

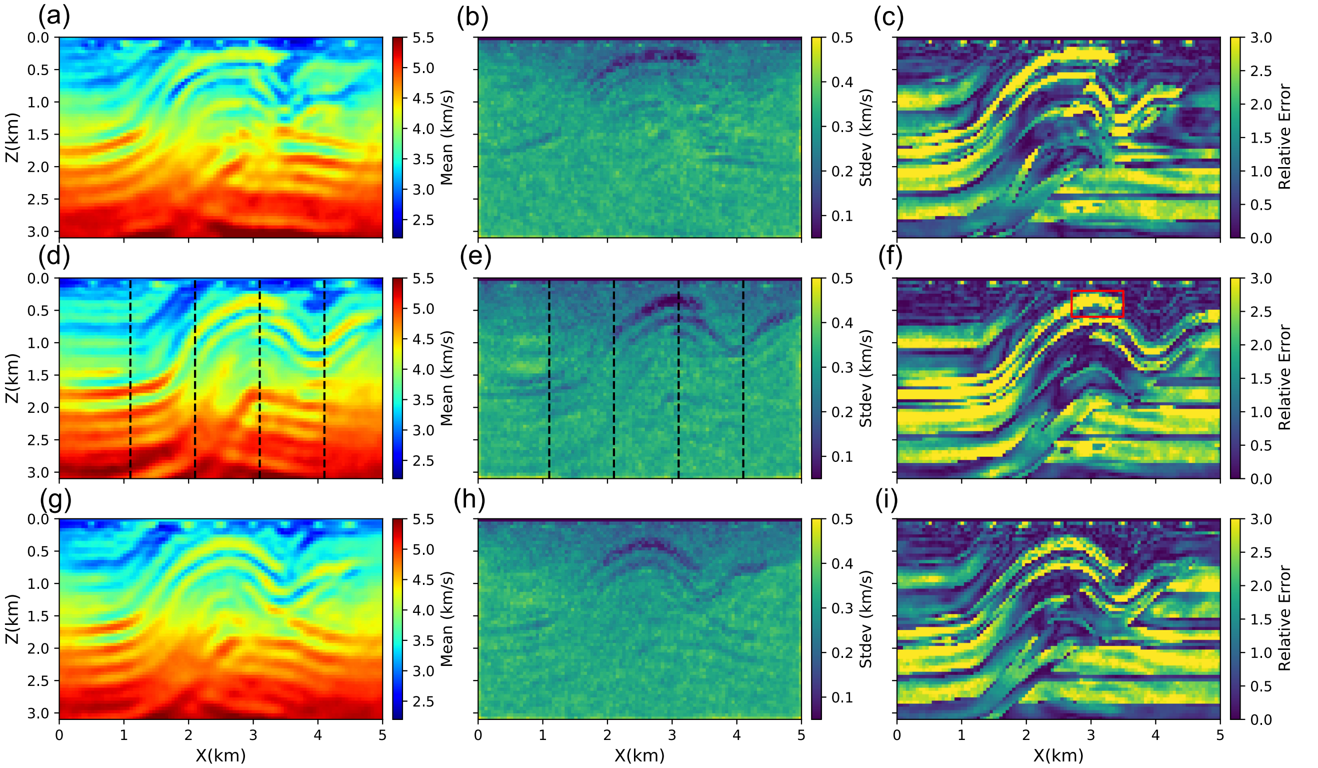

Figure 5. The mean, standard deviation and relative error obtained using SVGD. Key as in Fig. 4.

approximation. However, although the relative error is smaller than those from ADVI, there is still a large part of the model whose relative error is higher than three standard deviations which suggests that SVGD can also underestimate the uncertainty (Ba *et al.* 2021). This is probably because we use a small number of particles (400) to represent a probability distribution in an extremely high dimensional space (642 663). Consequently for those parts that are not well constrained by the data which should have a broader posterior distribution, it becomes impossible to represent the posterior distribution. Although the results can be further improved by using a larger number of particles (Zhang *et al.* 2021), this incurs a significantly higher computational cost.

Fig. 6 shows the results obtained using sSVGD. Ignoring magnitudes for the moment, the overall shapes of the mean and standard deviation models are similar to those obtained using ADVI and SVGD suggesting that these shapes may be reliable for this specific problem. Note that the mean model obtained using sSVGD is more similar to the true model, which may indicate that sSVGD produced more accurate results than ADVI or SVGD as we have discussed in Section 2. In addition, the magnitudes of the standard deviation are much higher than those obtained using ADVI or SVGD, and the relative error is also significantly smaller. For most parts the relative error obtained using sSVGD is smaller than three standard deviations, which is again indicative of the higher accuracy of sSVGD compared to ADVI or SVGD. Similarly to previous results, the deeper parts and two sides show larger errors than the rest of the model because of the lower sensitivity of our data to those parts. Note that the results obtained using ADVI and SVGD show smoother structures than those obtained using sSVGD. This is because in ADVI and SVGD the results are obtained deterministically, whereas sSVGD is a stochastic McMC method which therefore represents more randomness. A similar phenomenon was observed by Zhang & Curtis (2020b) when comparing results obtained using SVGD and HMC. We also note that the results can be further improved by running the sSVGD for longer. In all results, the shallow standard deviations show lower uncertainties at the location of higher velocity anomalies. This is probably because those high velocity anomalies in a relatively low velocity background have strong effects on the recorded waveforms, and hence are well constrained by the data. By contrast, the low velocity anomalies do not show a similar effect, which is likely because those low velocity anomalies are not strong enough to have large influences on the waveforms as the high velocity anomalies do. We note that similar effects have also been observed in 2-D FWI (Gebraad *et al.* 2020; Zhang & Curtis 2021).

In Fig. 7, we show examples of samples (particles) obtained using each method at the same cross sections as above. Overall the samples obtained using different methods show similar structures. For example, the shallow part ($Z < 1.5$ km) shows similar features to the true model, whereas the deeper part has more random structures. Similarly to the mean and standard deviation models, the sample obtained using SVGD is smoother than that obtained using sSVGD. There is no correlation between parameters in ADVI, so the sample obtained using ADVI shows random structures at pixel scale. Note that there are also small scale structures comprising clusters of a few pixels in the particles obtained using SVGD and sSVGD, which may reflect the uncertainty of the problem itself, or may appear because the methods have not fully converged. We note that such structures also appear in linearized FWI in cases when the regularisation applied is weak (Asnaashari *et al.* 2013), which suggests that the former explanation is at least partly the cause.

To analyse higher order statistics, we calculated the average correlation coefficients for parameter pairs that have the same distance in the volumes denoted using black boxes in Fig. 7. The results show that there is no significant correlation between model parameters obtained using ADVI, whereas correlation can be observed clearly in the results obtained using either SVGD or sSVGD (Fig. 8). For pairs of points up to approximately 0.15 km apart, the correlation coefficients obtained using SVGD are higher in magnitude than those obtained using sSVGD, reflecting larger spatial correlation lengths for the results of SVGD than those of sSVGD. For example, the spatial correlation length for the shallower volume is found to be 0.17 and 0.12 km for SVGD and sSVGD, respectively (Fig. 8a). This is also consistent with the observation that particles obtained using SVGD represent smoother models than those obtained using sSVGD.

To further analyse the results, we show the marginal distributions obtained using the suite of methods along four vertical profiles simulating well logs, whose locations are indicated using black dashed lines in Figs 4, 5 and 6. The results clearly show that the marginal distributions obtained using sSVGD are wider than those obtained using ADVI and SVGD as we have already observed. Across deeper parts ($Z > 1.5$ km), the true velocity values lie outside of the high probability area in the results obtained using ADVI and SVGD (Figs 9a and b), which again demonstrates that ADVI and SVGD can underestimate uncertainty. In contrast, sSVGD produces more reasonable uncertainty estimates since they at least generally include the true model in values with non-zero uncertainty. Overall the results show lower uncertainty in the shallower part ($Z < 1.5$ km) and higher uncertainty at the deeper part as we expect. Note that at the depth of 0.4 km in the third well log (denoted by a blue arrow), the marginal distributions concentrate close to the upper bound of the prior distribution. This is because the true velocity at this location is higher than the prior upper bound, which also explains the large relative error in this area (red box in Fig. 4, 5 and 6). This result provides useful insight into the performance of these methods in real applications as it is not uncommon to impose inappropriate prior information in practice.

3.2 Computational cost

In Table 1 we summarize the number of simulations, the number of CPU cores, and the wall clock time required by each method. The number of simulations provides a good metric of the overall computational cost as for each method the forward and adjoint simulations are the most time-consuming components of these calculations. Given that all of the methods can be fully parallelized, for example, the gradient calculation in each method can be performed independently for each particle (sample), the number of CPU cores together with the wall clock time provide additional insights into the computational requirement in practice.

The results show that ADVI is the cheapest method as it only requires 4000 simulations which we performed using 768 CPU cores, but we have demonstrated above that the method is likely to produce systematically biased results. However, given that the method is extremely efficient (only requiring 53.8 hr in real time), ADVI could still be used to provide a first, relatively rapid insight into the subsurface structure. In addition, as we have demonstrated in Fig. 1a, the method can be used to provide a lower bound estimate of the uncertainty. SVGD appears to be the most expensive method, which requires 400 000 simulations and takes approximately 23 d to run using 7680 CPU cores. Because of the limited number of

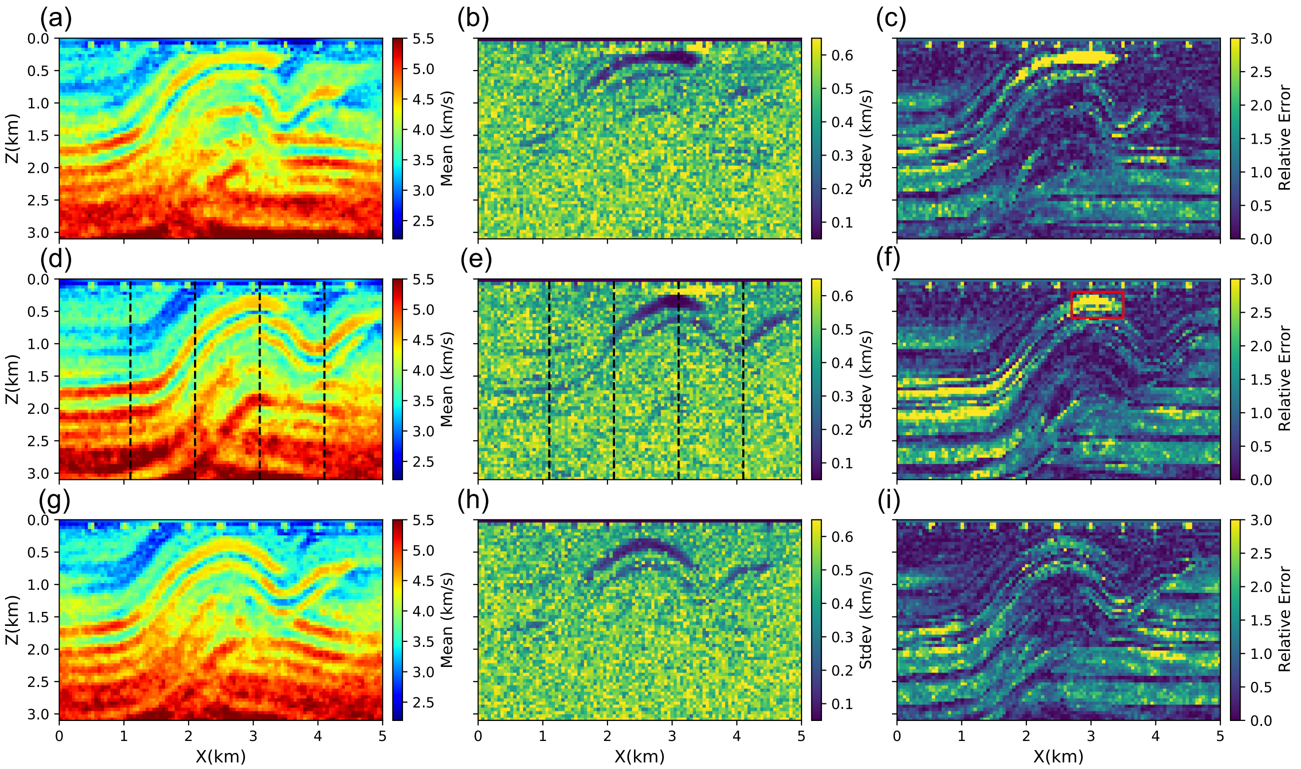

Figure 6. The mean, standard deviation and relative error obtained using sSVGD. Key as in Fig. 4.

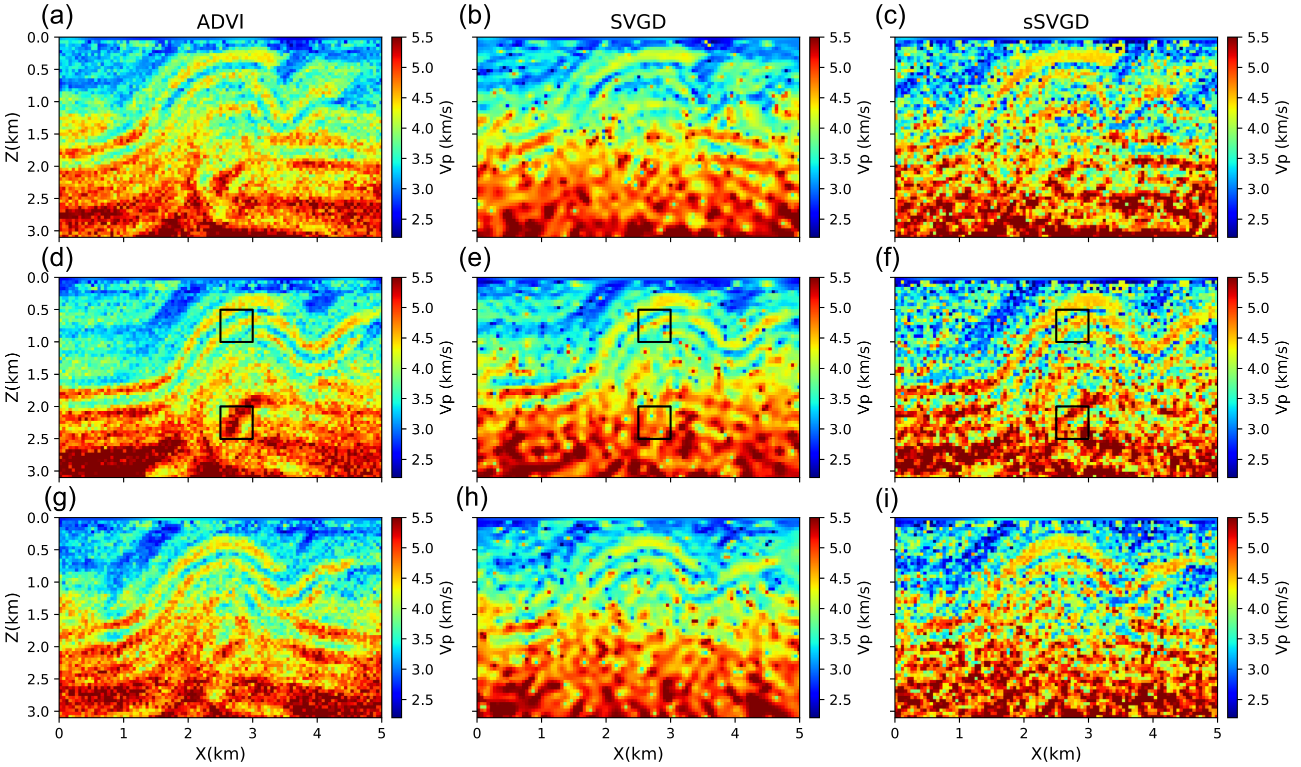

Figure 7. Example particles obtained using ADVI (a, d and g), SVGD (b,e and h) and sSVGD (c, f and i) over the same cross sections as in Fig. 3.

particles the method also provides biased results as we have shown above, which makes SVGD a less attractive method for 3-D FWI in practice. In contrast, by adding a noise term to the dynamics of SVGD, sSVGD can use a small number of particles to generate many final model samples, which makes the method relatively efficient. For example, to obtain the above results sSVGD required five times fewer simulations than SVGD. However, because of the randomness introduced by the noise term, sSVGD requires more iterations to

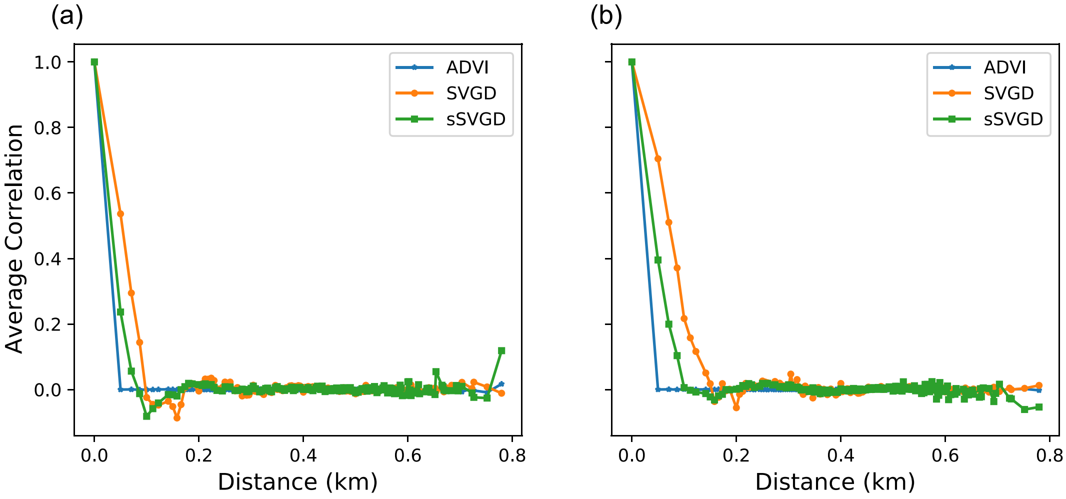

Figure 8. The average correlation coefficients plotted as a function of distance, calculated from the posterior distributions obtained using the different methods within the (a) upper and (b) lower volumes shown by black boxes in Fig. 7 (volumes are obtained by extending the area in Fig. 7 by ±250 m in the *Y* direction).

converge which makes the method only two to three times more efficient in real time. We note that an increase in efficiency was also observed in deterministic FWI by introducing stochastic elements, for example, using stochastic gradient descent (Yang *et al.* 2018; van Herwaarden *et al.* 2020). Given that sSVGD also provides the most accurate results among the three methods, the method would be a good choice for practical applications. In addition, since it is a McMC method the results of sSVGD can always be improved by performing more iterations, whereas the same method of improvement cannot be used when using ADVI or SVGD.

Note that the above comparison depends on subjective assessments of the point of convergence for each method, so the absolute computational time may not be entirely accurate. Nevertheless the comparison at least provides a reasonable insight into the efficiency of each method. We also note that all of the methods require computation of gradients, which in this study are calculated efficiently using adjoint methods. For situations in which gradients are expensive to compute, the above suite of methods may become less efficient, and in such cases other methods that do not require gradients may be preferred.

4 DISCUSSION

The primary result of this work is to show that variational methods (ADVI, SVGD and sSVGD) can be used to solve 3-D Bayesian FWI problems. For ADVI, we used a mean-field approximation to reduce the computational cost, which systematically underestimates the uncertainty. To further improve the results, a full-rank covariance matrix may be used if sufficient computational resources are available, or a sparse covariance matrix which only includes correlation information between neighbouring cells can be implemented. ADVI minimizes KL[$q||p$] to estimate the posterior distribution which can provide a lower bound estimate of the uncertainty in the mean-field case. On the other hand, methods such as the expectation propagation (Minka 2013) which minimizes KL[$p||q$] instead of KL[$q||p$], may be used to provide an upper bound estimate of the uncertainty.

We have demonstrated that for 3-D FWI SVGD can provide biased results because of the limited number of particles. Instead of increasing the number of particles which may be computationally intractable, one may try to reduce the dimensionality of the problem. For example, other parametrizations that require fewer parameters to represent the model may be used, such as Voronoi cells (Bodin & Sambridge 2009; Zhang *et al.* 2018b), wavelet parametrization (Hawkins & Sambridge 2015), Johnson-Mehl tessellation (Belhadj *et al.* 2018), Delaunay and Clough-Tocher parametrizations (Curtis & Snieder 1997) or discrete cosine transforms (Urozayev *et al.* 2022). In addition, other SVGD variants which project the high dimensional parameter space into a lower dimensional space may be used to improve the results, for example, projected SVGD (Chen & Ghattas 2020) or sliced SVGD (Gong *et al.* 2020).

By adding a noise term to the dynamics of SVGD, sSVGD becomes a McMC method with multiple interactive chains. Note that this is different from other McMC methods which run multiple interactive chains such as parallel tempering (Earl & Deem 2005; Sambridge 2013). In parallel tempering, a set of chains with different temperatures are run in parallel, and at each iteration samples in two randomly selected (or neighbouring) chains are exchanged with a Metropolis–Hastings criterion. In sSVGD, all Markov chains interact by using a kernel function and hence no sample exchange occurs between chains.

Although sSVGD provides more accurate results than ADVI and SVGD, it also requires more iterations to converge. To improve efficiency of the method, one might exploit higher order gradient information, for example, using a Hessian matrix kernel (Wang *et al.* 2019) or the stochastic Stein variational Newton method (Leviyev *et al.* 2022). Since sSVGD is a McMC method, one can further improve the accuracy of the method by implementing a Metropolis–Hastings correction step at each iteration (Metropolis & Ulam 1949; Hastings 1970), though in such cases stochastic minibatches may not be used because of the detailed balance requirement of the Metropolis–Hastings step.

Note that for both SVGD and sSVGD, the posterior distribution is likely to be under sampled given the large dimensionality (642 663) and the small number of samples (400 and 10 000, respectively). While the set of samples may not be sufficient to represent the full posterior distribution, they may at least provide reasonable mean and (in the case of sSVGD) standard deviation estimates. We also note that in practice the number of samples is always restricted by the available computational cost.

In this study we used a Uniform prior distribution. This may cause posterior pdfs to occur that are more complex than would be the case if Gaussian or other prior distributions were used that more strongly

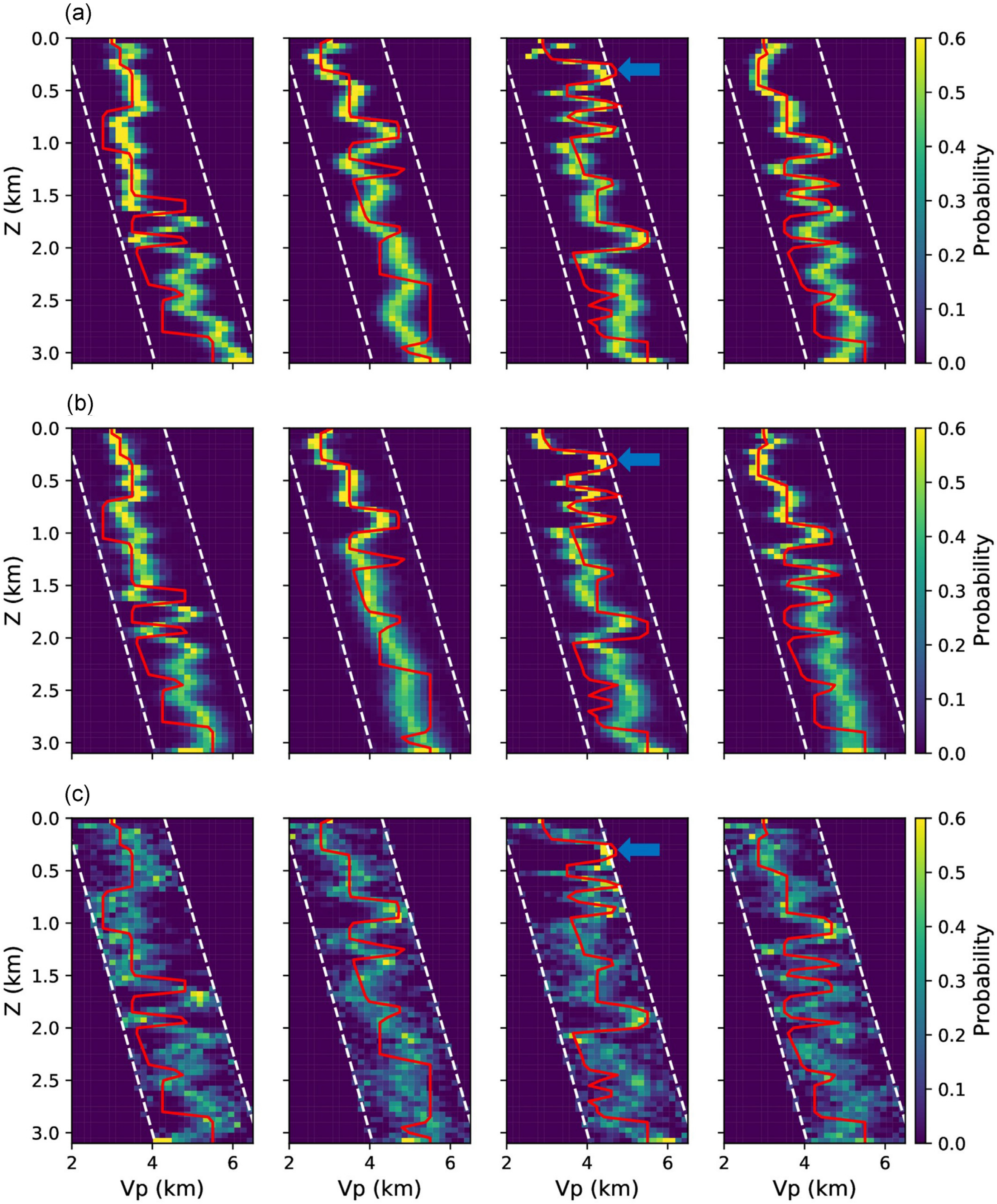

Figure 9. The marginal distributions at four well logs (black dashed line in Figs 4, 5 and 6) obtained using (a) ADVI, (b) SVGD and (c) sSVGD, respectively. Red lines show the true velocity profiles and white dashed lines show the lower and upper bound of the prior distribution. Blue arrows highlight the interval (the red box in Fig. 4) in which the true velocities are outside of the support (positive probabilities) of the prior distribution.

focus the solution towards certain regions of parameter space. This means that our posterior pdf may be harder to explore than would otherwise be the case. In practice where more knowledge about the subsurface is available, one can use a more informative prior distribution. For example, models obtained using fast traveltime tomography can be used as prior information for FWI. In addition, prior regularization or Gaussian processes may be used to produce smoother models (MacKay 2003; Ray & Myer 2019). Neural networks can also be used to encode geological information into prior distributions (Laloy *et al.* 2017; Mosser *et al.* 2020).

For the likelihood function we used Gaussian data uncertainties with a known, fixed data noise level. In practice this noise level should be determined from the data, for example, by using the maximum likelihood method (Sambridge 2013). It may also be possible

Table 1. A comparison of computational cost for the three inference methods.

Method	Number of simulations[a]	CPU cores[b]	Wall time (hr)
ADVI	4000	768	53.8
SVGD	400 000	7680	558.7
sSVGD	80 000	3840	220.8

[a] This is measured as the number of minibatch simulations.
[b] The CPU used in this study is Intel Xeon Platinum.

to estimate the noise level in the inversion process using a hierarchical Bayesian formulation (Ranganath *et al.* 2016; Malinverno & Briggs 2004). We also note that other non-Gaussian likelihood functions may be used to improve the results given that those likelihood functions are defined to represent the probability distribution of data uncertainty (Zhang *et al.* 2023).

In this study, we demonstrated variational inference methods using the overthrust model. In preparatory tests we inverted for layered 3-D models and found that the computational cost was similar to the case shown here. This suggests that increases in the true structural complexity are not necessarily reflected in increased computational cost. We note, however, that for much stronger heterogeneity the required computational cost may increase because of the increase in non-linearity caused by stronger heterogeneity. For computational efficiency we only applied the methods to a small area with a small dataset. For large subsurface volumes the number of particles and iterations required by the methods may increase significantly because of the curse of dimensionality (Curtis & Lomax 2001). However, the scaling of the computational cost is not obvious because while the cost of forward modelling increases predictably with model or particle size, these methods sample the posterior probability distribution which is expected to be far more limited in its support than might be expected from the increased dimensionality of the models. The cost will therefore depend on the specifics of the data and prior information available in each case. As a result, the methods may become computationally intractable for large subsurface volumes and large datasets. In such cases one may use experimental design methods (Curtis 2004; Maurer *et al.* 2010) to select a small part of the large dataset, and perform inversions using those selected data. Faster, approximate forward modelling methods may also be used to improve efficiency of the methods, for example neural network based modelling methods (Sirignano & Spiliopoulos 2018). We also note that apart from the mean and uncertainty models, the obtained samples can be used for real-world applications, for example, providing models for reservoir simulations or answering specific scientific questions (Arnold & Curtis 2018; Zhang & Curtis 2022; Zhao *et al.* 2022).

5 CONCLUSION

In this study we applied three different variational inference methods: ADVI, SVGD and sSVGD to 3-D FWI, and demonstrated feasibility of using these methods to solve large scale probabilistic inverse problems. The results show that ADVI with a mean-field approximation can provide rapid solutions but with systematically underestimated uncertainty. In practice, the method can therefore be used to provide a rapid initial estimate of the solution, or to provide a lower bound estimate of the uncertainty. SVGD appears to be the most expensive method, but still provides a biased solution because of the limited number of particles. By contrast, by adding a noise term in the dynamics of SVGD, sSVGD becomes a Markov chain Monte Carlo method and provides the most accurate results. We thus conclude that variational inference methods can be used to solve real-world 3-D full wave form inversion problems.

ACKNOWLEDGMENTS

The authors thank the Edinburgh Imaging Project sponsors (BP and Total) for supporting this research.

DATA AVAILABILITY STATEMENTS

The model used in this article is available in *Zenodo* at https://doi.org/10.5281/zenodo.4252588. The code underlying this paper will be shared on reasonable request to the corresponding author.

REFERENCES

Ahmed, Z., Yunyue, L. & Arthur, C., 2022. Regularized seismic amplitude inversion via variational inference, *Geophys. Prospect.,* **70**(9), 1507–1527.

Aminzadeh, F., 1997. SEG/EAGA 3-D salt and overthrust models, in *SEG/EAGE 3-D Modeling Series No. 1,* SEG.

Arnold, R. & Curtis, A., 2018. Interrogation theory, *J. geophys. Int.,* **214**(3), 1830–1846.

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.

Ba, J., Erdogdu, M.A., Ghassemi, M., Sun, S., Suzuki, T., Wu, D. & Zhang, T., 2021. Understanding the variance collapse of SVGD in high dimensions, in *International Conference on Learning Representations*. https://openreview.net/forum?id=Qycd9j5Qp9J

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

Belhadj, J., Romary, T., Gesret, A., Noble, M. & Figliuzzi, B., 2018. New parameterizations for Bayesian seismic tomography, *Inverse Problems,* **34**(6), doi:10.1088/1361-6420/aabce7.

Bishop, C.M., 2006. *Pattern Recognition and Machine Learning,* Springer.

Blatter, D., Key, K., Ray, A., Gustafson, C. & Evans, R., 2019. Bayesian joint inversion of controlled source electromagnetic and magnetotelluric data to image freshwater aquifer offshore New Jersey, *J. geophys. Int.,* **218**(3), 1822–1837.

Blei, D.M., Kucukelbir, A. & McAuliffe, J.D., 2017. Variational inference: a review for statisticians, *J. Am. Stat. Assoc.,* **112**(518), 859–877.

Bodin, T. & Sambridge, M., 2009. Seismic tomography with the reversible jump algorithm, *J. geophys. Int.,* **178**(3), 1411–1436.

Bodin, T., Sambridge, M., Tkalčić, H., Arroucau, P., Gallagher, K. & Rawlinson, N., 2012. Transdimensional inversion of receiver functions and surface wave dispersion, *J. geophys. Res.,* **117**(B2), doi:10.1029/2011JB008560.

Bosch, M., Meza, R., Jiménez, R. & Hönig, A., 2006. Joint gravity and magnetic inversion in 3D using Monte Carlo methods, *Geophysics,* **71**(4), G153–G156.

Bozdağ, E., Trampert, J. & Tromp, J., 2011. Misfit functions for full waveform inversion based on instantaneous phase and envelope measurements, *J. geophys. Int.,* **185**(2), 845–870.

Bozdağ, E., Peter, D., Lefebvre, M., Komatitsch, D., Tromp, J., Hill, J., Podhorszki, N. & Pugmire, D., 2016. Global adjoint tomography: first-generation model, *J. geophys. Int.,* **207**(3), 1739–1766.

Brooks, S., Gelman, A., Jones, G. & Meng, X.-L., 2011. *Handbook of Markov chain Monte Carlo,* CRC Press.

Brossier, R., Operto, S. & Virieux, J., 2010. Which data residual norm for robust elastic frequency-domain full waveform inversion?, *Geophysics,* **75**(3), R37–R46.

Burdick, S. & Lekić, V., 2017. Velocity variations and uncertainty from transdimensional P-wave tomography of North America, *J. geophys. Int.,* **209**(2), 1337–1351.

Chen, M., Niu, F., Liu, Q., Tromp, J. & Zheng, X., 2015. Multiparameter adjoint tomography of the crust and upper mantle beneath east Asia: 1. Model construction and comparisons, *J. geophys. Res.,* **120**(3), 1762–1786.

Chen, P. & Ghattas, O., 2020. Projected stein variational gradient descent, *Adv. Neural Inform. Process. Syst.,* **33,** 1947–1958.

Chen, P., Zhao, L. & Jordan, T.H., 2007. Full 3D tomography for the crustal structure of the Los Angeles region, *Bull. seism. Soc. Am.,* **97**(4), 1094–1120.

Chen, T., Fox, E. & Guestrin, C., 2014. Stochastic gradient Hamiltonian Monte Carlo, *International Conference on Machine Learning,* pp. 1683–1691, PMLR.

Curtis, A., 2004. Seismic survey design-theory of model-based geophysical survey and experimental design, part 1, *Leading Edge,* **23**(10), 997–1006.

Curtis, A. & Lomax, A., 2001. Prior information, sampling distributions, and the curse of dimensionality, *Geophysics,* **66**(2), 372–378.

Curtis, A. & Snieder, R., 1997. Reconditioning inverse problems using the genetic algorithm and revised parameterization, *Geophysics,* **62**(5), 1524–1532.

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

Duane, S., Kennedy, A.D., Pendleton, B.J. & Roweth, D., 1987. Hybrid Monte Carlo, *Phys. Lett. B,* **195**(2), 216–222.

Earl, D.J. & Deem, M.W., 2005. Parallel tempering: theory, applications, and new perspectives, *Phys. Chem. Chem. Phys.,* **7**(23), 3910–3916.

Fichtner, A., Bunge, H.-P. & Igel, H., 2006. The adjoint method in seismology: I. Theory, *Phys. Earth planet. Inter.,* **157**(1-2), 86–104.

Fichtner, A., Kennett, B.L., Igel, H. & Bunge, H.-P., 2008. Theoretical background for continental-and global-scale full-waveform inversion in the time–frequency domain, *J. geophys. Int.,* **175**(2), 665–685.

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, *J. geophys. Int.,* **179**(3), 1703–1725.

Fichtner, A. *et al.*, 2018a. The collaborative seismic earth model: generation 1, *Geophys. Res. Lett.,* **45**(9), 4007–4016.

Fichtner, A., Zunino, A. & Gebraad, L., 2018b. Hamiltonian Monte Carlo solution of tomographic inverse problems, *J. geophys. Int.,* **216**(2), 1344–1363.

French, S. & Romanowicz, B., 2014. Whole-mantle radially anisotropic shear velocity structure from spectral-element waveform tomography, *J. geophys. Int.,* **199**(3), 1303–1327.

Galetti, E. & Curtis, A., 2018. Transdimensional electrical resistivity tomography, *J. geophys. Res.,* **123**(8), 6347–6377.

Galetti, E., Curtis, A., Meles, G.A. & Baptie, B., 2015. Uncertainty loops in travel-time tomography from nonlinear wave physics, *Phys. Rev. Lett.,* **114**(14), doi:10.1103/PhysRevLett.114.148501.

Galetti, E., Curtis, A., Baptie, B., Jenkins, D. & Nicolson, H., 2017. Transdimensional love-wave tomography of the British Isles and shear-velocity structure of the east Irish Sea Basin from ambient-noise interferometry, *J. geophys. Int.,* **208**(1), 36–58.

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

Gauthier, O., Virieux, J. & Tarantola, A., 1986. Two-dimensional nonlinear inversion of seismic waveforms: numerical results, *Geophysics,* **51**(7), 1387–1403.

Gebraad, L., Boehm, C. & Fichtner, A., 2020. Bayesian elastic full-waveform inversion using Hamiltonian Monte Carlo, *J. geophys. Res.,* **125**(3), e2019JB018428, doi:10.31223/osf.io/qftn5.

Gee, L.S. & Jordan, T.H., 1992. Generalized seismological data functionals, *J. geophys. Int.,* **111**(2), 363–390.

Gong, C., Peng, J. & Liu, Q., 2019. Quantile stein variational gradient descent for batch Bayesian optimization, *International Conference on Machine Learning,* pp. 2347–2356, PMLR.

Gong, W., Li, Y. & Hernández-Lobato, J.M., 2020. Sliced kernelized stein discrepancy, preprint (arXiv:2006.16531).

Guo, P., Visser, G. & Saygin, E., 2020. Bayesian trans-dimensional full waveform inversion: synthetic and field data application, *J. geophys. Int.,* **222**(1), 610–627.

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

Hawkins, R. & Sambridge, M., 2015. Geophysical imaging using transdimensional trees, *J. geophys. Int.,* **203**(2), 972–1000.

Hukushima, K. & Nemoto, K., 1996. Exchange Monte Carlo method and application to spin glass simulations, *J. Phys. Soc. Japan,* **65**(6), 1604–1608.

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

Kotsi, M., Malcolm, A. & Ely, G., 2020. 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.

Kubrusly, C. & Gravier, J., 1973. Stochastic approximation algorithms and applications, in *1973 IEEE Conference on Decision and Control Including the 12th Symposium on Adaptive Processes,* pp. 763–766, IEEE.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A. & Blei, D.M., 2017. Automatic differentiation variational inference, *J. Mach. Learn. Res.,* **18**(1), 430–474.

Kullback, S. & Leibler, R.A., 1951. On information and sufficiency, *Ann. Math. Stat.,* **22**(1), 79–86.

Laloy, E., Hérault, R., Lee, J., Jacques, D. & Linde, N., 2017. Inversion using a new low-dimensional representation of complex binary geological media based on a deep neural network, *Adv. Water Resour.,* **110,** 387–405.

Lei, W. *et al.*, 2020. Global adjoint tomography-model glad-m25, *J. geophys. Int.,* **223**(1), 1–21.

Leviyev, A., Chen, J., Wang, Y., Ghattas, O. & Zimmerman, A., 2022. A stochastic Stein Variational Newton method, preprint (arXiv:2204.09039).

Liu, Q. & Gu, Y., 2012. Seismic imaging: From classical to adjoint tomography, *Tectonophysics,* **566,** 31–66.

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.

Luo, Y. & Schuster, G.T., 1991. Wave-equation traveltime inversion, *Geophysics,* **56**(5), 645–653.

Ma, Y.-A., Chen, T. & Fox, E., 2015. A complete recipe for stochastic gradient MCMC, in *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems,* Vol. 2, pp. 2917–2925, MIT Press.

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

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

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

Malinverno, A., Leaney, S. *et al.*, 2000. A Monte Carlo method to quantify uncertainty in the inversion of zero-offset VSP data, in *2000 SEG Annual Meeting,* Society of Exploration Geophysicists.

Martin, J., Wilcox, L.C., Burstedde, C. & Ghattas, O., 2012. A stochastic newton MCMC method for large-scale statistical inverse problems with application to seismic inversion, *SIAM J. Scient. Comput.,* **34**(3), A1460–A1487.

Maurer, H., Curtis, A. & Boerner, D.E., 2010. Recent advances in optimized geophysical survey design, *Geophysics,* **75**(5), 75A177–75A194.

Métivier, L., Brossier, R., Mérigot, Q., Oudet, E. & Virieux, J., 2016. Measuring the misfit between seismograms using an optimal transport distance: application to full waveform inversion, *J. geophys. Int.,* **205**(1), 345–377.

Metropolis, N. & Ulam, S., 1949. The Monte Carlo method, *J. Am. Stat. Assoc.,* **44**(247), 335–341.

Minka, T.P., 2013. Expectation propagation for approximate Bayesian inference, preprint (arXiv:1301.2294).

Minsley, B.J., 2011. A trans-dimensional Bayesian Markov Chain Monte Carlo algorithm for model assessment using frequency-domain electromagnetic data, *J. geophys. Int.,* **187**(1), 252–272.

Mosegaard, K. & Tarantola, A., 1995. Monte Carlo sampling of solutions to inverse problems, *J. Geophys. Res.,* **100**(B7), 12 431–12 447.

Mosser, L., Dubrule, O. & Blunt, M.J., 2020. Stochastic seismic waveform inversion using generative adversarial networks as a geological prior, *Math. Geosci.,* **52**(1), 53–79.

Nawaz, M.A. & Curtis, A., 2018. Variational Bayesian inversion (VBI) of quasi-localized seismic attributes for the spatial distribution of geological facies, *J. geophys. Int.,* **214**(2), 845–875.

Nawaz, M.A. & Curtis, A., 2019. Rapid discriminative variational Bayesian inversion of geophysical data for the spatial distribution of geological properties, *J. geophys. Res.,* **124**(6), 5867–5887.

Nawaz, M.A., Curtis, A., Shahraeeni, M.S. & Gerea, C., 2020. Variational Bayesian inversion of seismic attributes jointly for geological facies and petrophysical rock properties, *Geophysics,* **85**(4), 1–78.

O'Hagan, A. & Forster, J.J., 2004. *Kendall's Advanced Theory of Statistics,* Vol. 2B: Bayesian Inference, Arnold.

Piana Agostinetti, N., Giacomuzzi, G. & Malinverno, A., 2015. Local three-dimensional earthquake tomography by trans-dimensional Monte Carlo sampling, *J. geophys. Int.,* **201**(3), 1598–1617.

Pinder, T., Nemeth, C. & Leslie, D., 2020. Stein variational Gaussian processes, preprint (arXiv:2009.12141).

Plessix, R.-E., 2006. A review of the adjoint-state method for computing the gradient of a functional with geophysical applications, *J. geophys. Int.,* **167**(2), 495–503.

Pratt, R.G., 1999. Seismic waveform inversion in the frequency domain, part 1: theory and verification in a physical scale model, *Geophysics,* **64**(3), 888–901.

Prieux, V., Brossier, R., Operto, S. & Virieux, J., 2013. Multiparameter full waveform inversion of multicomponent ocean-bottom-cable data from the Valhall field. Part 1: imaging compressional wave speed, density and attenuation, *J. geophys. Int.,* **194**(3), 1640–1664.

Ramgraber, M., Weatherl, R., Blumensaat, F. & Schirmer, M., 2021. Non-Gaussian parameter inference for hydrogeological models using stein variational gradient descent, *Water Resour. Res.,* **57**(4), e2020WR029339, doi:10.1029/2020WR029339.

Ranganath, R., Tran, D. & Blei, D., 2016. Hierarchical variational models, in *International Conference on Machine Learning,* pp. 324–333.

Ray, A. & Myer, D., 2019. Bayesian geophysical inversion with trans-dimensional Gaussian process machine learning, *J. geophys. Int.,* **217**(3), 1706–1726.

Ray, A., Alumbaugh, D.L., Hoversten, G.M. & Key, K., 2013. Robust and accelerated Bayesian inversion of marine controlled-source electromagnetic data using parallel tempering, *Geophysics,* **78**(6), E271–E280.

Ray, A., Kaplan, S., Washbourne, J. & Albertin, U., 2017. Low frequency full waveform seismic inversion within a tree based Bayesian framework, *J. geophys. Int.,* **212**(1), 522–542.

Rezende, D.J. & Mohamed, S., 2015. Variational inference with normalizing flows, *Proceedings of the 32nd International Conference on Machine Learning* **37** 1530–-1538 PMLR.

Robbins, H. & Monro, S., 1951. A stochastic approximation method, *Ann. Math. Stat.,* **22**(3), 400–407.

Roberts, G.O., Tweedie, R.L. *et al.*, 1996. Exponential convergence of Langevin distributions and their discrete approximations, *Bernoulli,* **2**(4), 341–363.

Rossi, L., 2017. Bayesian gravity inversion by Monte Carlo methods, *Doctoral dissertation*, Department of Environmental and Civil Engineering, Politecnico di Milano.

Ryan, H., 1994. Ricker, Ormsby; Klander, Bntterwo-a choice of wavelets, *CSEG Recorder,* **19**(07). http://www.dim.uchile.cl/~fmaldonado/Documentos/trabajos/lab/sep94-choice-of-wavelets.pdf

Ryberg, T., Kirsch, M., Haberland, C., Tolosana-Delgado, R., Viezzoli, A. & Gloaguen, R., 2022. Ambient seismic noise analysis of large-n data for mineral exploration in the central Erzgebirge, Germany, *Solid Earth,* **13**(3), 519–533.

Sambridge, M., 2013. A parallel tempering algorithm for probabilistic sampling and multimodal optimization, *J. geophys. Int.,* **196**(1), 357–374.

Sambridge, M., Jackson, A. & Valentine, A.P., 2022. Geophysical inversion and optimal transport, *J. geophys. Int.,* **231**(1), 172–198.

Sen, M.K. & Biswas, R., 2017. Transdimensional seismic inversion using the reversible jump Hamiltonian Monte Carlo algorithm, *Geophysics,* **82**(3), R119–R134.

Shen, W., Ritzwoller, M.H., Schulte-Pelkum, V. & Lin, F.-C., 2012. Joint inversion of surface wave dispersion and receiver functions: a Bayesian Monte-Carlo approach, *J. geophys. Int.,* **192**(2), 807–836.

Siahkoohi, A., Rizzuti, G. & Herrmann, F.J., 2020a. Uncertainty quantification in imaging and automatic horizon tracking—a Bayesian deep-prior based approach, in *SEG Technical Program Expanded Abstracts 2020,* pp. 1636–1640, Society of Exploration Geophysicists.

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

Sirignano, J. & Spiliopoulos, K., 2018. DGM: a deep learning algorithm for solving partial differential equations, *J. Comput. Phys.,* **375,** 1339–1364.

Smith, J.D., Ross, Z.E., Azizzadenesheli, K. & Muir, J.B., 2022. Hyposvi: hypocentre inversion with stein variational inference and physics informed neural networks, *J. geophys. Int.,* **228**(1), 698–710.

Tape, C., Liu, Q., Maggi, A. & Tromp, J., 2009. Adjoint tomography of the southern California crust, *Science,* **325**(5943), 988–992.

Tarantola, A., 1984. Inversion of seismic reflection data in the acoustic approximation, *Geophysics,* **49**(8), 1259–1266.

Tarantola, A., 1988. Theoretical background for the inversion of seismic waveforms, including elasticity and attenuation, in *Scattering and Attenuations of Seismic Waves, Part I,* pp. 365–399, Springer.

Tarantola, A., 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation,* Vol. **89,** SIAM.

Stan Development Team *et al*, 2016. Stan modeling language users guide and reference manual, Technical report.

Tran, D., Ranganath, R. & Blei, D.M., 2015. The variational Gaussian process, preprint (arXiv:1511.06499).

Tromp, J., Tape, C. & Liu, Q., 2005. Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels, *J. geophys. Int.,* **160**(1), 195–216.

Urozayev, D., Ait-El-Fquih, B., Hoteit, I. & Peter, D., 2022. A reduced-order variational Bayesian approach for efficient subsurface imaging, *J. geophys. Int.,* **229**(2), 838–852.

van Herwaarden, D.P., Boehm, C., Afanasiev, M., Thrastarson, S., Krischer, L., Trampert, J. & Fichtner, A., 2020. Accelerated full-waveform inversion using dynamic mini-batches, *J. geophys. Int.,* **221**(2), 1427–1438.

Van Leeuwen, T. & Mulder, W., 2010. A correlation-based misfit criterion for wave-equation traveltime tomography, *J. geophys. Int.,* **182**(3), 1383–1394.

Virieux, J. & Operto, S., 2009. An overview of full-waveform inversion in exploration geophysics, *Geophysics,* **74**(6), WCC1–WCC26.

Wang, D., Tang, Z., Bajaj, C. & Liu, Q., 2019. Stein variational gradient descent with matrix-valued kernels, in *Advances in Neural Information Processing Systems,* pp. 7836–7846.

Warner, M. & Guasch, L., 2016. Adaptive waveform inversion: theory, *Geophysics,* **81**(6), R429–R445.

Warner, M. *et al.*, 2013. Anisotropic 3D full-waveform inversion, *Geophysics,* **78**(2), R59–R80.

Welling, M. & Teh, Y.W., 2011. Bayesian learning via stochastic gradient langevin dynamics, in *Proceedings of the 28th International Conference On Machine Learning (ICML-11),* pp. 681–688, Citeseer.

Yang, H., Jia, J., Wu, B. & Gao, J., 2018. Mini-batch optimized full waveform inversion with geological constrained gradient filtering, *J. appl. Geophys.,* **152,** 9–16.

Young, M.K., Rawlinson, N. & Bodin, T., 2013. Transdimensional inversion of ambient seismic noise for 3D shear velocity structure of the Tasmanian crust, *Geophysics,* **78**(3), WB49–WB62.

Yuan, Y.O., Bozdağ, E., Ciardelli, C., Gao, F. & Simons, F.J., 2020. The exponentiated phase measurement, and objective-function hybridization for adjoint waveform tomography, *J. geophys. Int.,* **221**(2), 1145–1164.

Zhang, C. & Chen, T., 2022. Bayesian slip inversion with automatic differentiation variational inference, *J. geophys. Int.,* **229**(1), 546–565.

Zhang, C., Bütepage, J., Kjellström, H. & Mandt, S., 2018a. Advances in variational inference, *IEEE Trans. Pattern Anal. Mach. Intell.,* **41**(8), 2008–2026.

Zhang, X. & Curtis, A., 2020a. Seismic tomography using variational inference methods, *J. geophys. Res.,* **125**(4), e2019JB018589, doi:10.1029/2019JB018589.

Zhang, X. & Curtis, A., 2020b. Variational full-waveform inversion, *J. geophys. Int.,* **222**(1), 406–411.

Zhang, X. & Curtis, A., 2021. Bayesian full-waveform inversion with realistic priors, *Geophysics,* **86**(5), 1–20.

Zhang, X. & Curtis, A., 2022. Interrogating probabilistic inversion results for subsurface structural information, *J. geophys. Int.,* **229**(2), 750–757.

Zhang, X., Curtis, A., Galetti, E. & de Ridder, S., 2018b. 3-D Monte Carlo surface wave tomography, *J. geophys. Int.,* **215**(3), 1644–1658.

Zhang, X., Hansteen, F., Curtis, A. & de Ridder, S., 2020a. 1D, 2D and 3D Monte Carlo ambient noise tomography using a dense passive seismic array installed on the North Sea seabed, *J. geophys. Res.,* **125**(2), e2019JB018552, doi:10.1029/2019JB018552.

Zhang, X., Roy, C., Curtis, A., Nowacki, A. & Baptie, B., 2020b. Imaging the subsurface using induced seismicity and ambient noise: 3-D tomographic Monte Carlo joint inversion of earthquake body wave traveltimes and surface wave dispersion, *J. geophys. Int.,* **222**(3), 1639–1655.

Zhang, X., Nawaz, M.A., Zhao, X. & Curtis, A., 2021. An introduction to variational inference in geophysical inverse problems, *Adv. Geophys.,* **62,** 73–140.

Zhang, X., Zheng, Y. & Curtis, A., 2023. Surface wave dispersion inversion using an energy likelihood function, *J. geophys. Int.,* **232**(1), 523–536.

Zhao, X., Curtis, A. & Zhang, X., 2021. Bayesian seismic tomography using normalizing flows, *J. geophys. Int.,* **228**(1), 213–239.

Zhao, X., Curtis, A. & Zhang, X., 2022. Interrogating subsurface structures using probabilistic tomography: an example assessing the volume of irish sea basins, *J. geophys. Res.,* **127**(4), e2022JB024098, doi:10.1029/2022JB024098.

Zhao, Z. & Sen, M.K., 2019. A gradient based MCMC method for FWI and uncertainty analysis, in *SEG Technical Program Expanded Abstracts 2019,* pp. 1465–1469, Society of Exploration Geophysicists.