

3D Bayesian Variational Surface Wave Tomography and Application to the Southwest China

Wenda Yang¹, Xin Zhang^{1*}

¹ State Key Laboratory of Deep Earth Exploration and Imaging, School of Engineering and Technology, China

University of Geosciences, Beijing 100083, China

* Corresponding author: Xin Zhang. E-mail: xzhang@cugb.edu.cn

Abstract

Seismic surface wave tomography uses surface wave information to obtain velocity structures in the subsurface. Due to data noise and nonlinearity of the problem, surface wave tomography often has non-unique solutions. It is therefore required to quantify uncertainty of the results in order to better interpret the resulting images. Bayesian inference is the most widely-used method for this purpose. However, the commonly-used Monte Carlo methods require huge computational cost and remains intractable in high-dimensional problems. Variational inference uses optimization to solve Bayesian inverse problems, and therefore can be more efficient in the case of large datasets and high-dimensional parameter spaces. Variational inference has been applied to 2-D surface wave tomographic problems. In this study, we extend the method to 3-D surface wave tomography by directly inverting for 3-D spatial structures from frequency-dependent travel time measurements. Specifically, we apply three variational methods, mean-field automatic differential variational inference (mean-field ADVI), physically structured variational inference (PSVI) and stochastic Stein variational gradient descent (sSVGD) to surface wave tomographic problems using both synthetic data and real data in the Southwest China. The results show that all methods can provide accurate velocity estimates, while sSVGD produces more reasonable uncertainty estimates than mean-field ADVI and PSVI because of

Gaussian assumption used in the later methods. In the real data case, the variational methods provide more detailed velocity structures than those obtained using traditional methods, along with reliable uncertainty estimates. We therefore conclude that variational surface wave tomography can be applied fruitfully to many realistic problems.

1. Introduction

Understanding property variations of the Earth's crust is important to geoscientists, because these variations can provide critical insights into tectonic processes, seismic hazard assessment, natural resource exploration, and environmental monitoring. Seismic surface wave tomography is one popular method used to study these variations. Surface waves are low-frequency seismic waves that travel along the Earth's surface, and are dispersive, meaning that their velocity depends on the frequency of the wave (Aki & Richard 2002). By analyzing this dispersion characteristics of surface waves, one can infer velocity structure beneath the surface. Seismic surface wave tomography has been applied in various geological settings, from global-scale studies of the Earth's mantle to local investigations of the crust (Trampert & Woodhouse 1995; Shapiro & Ritzwoller 2002; Meier *et al.* 2007a,b; Zielhuis & Nolet 1994; Curtis *et al.* 1998; Simons *et al.* 2002).

Seismic surface wave tomography can be conducted in either a two-step strategy or a one-step strategy. In the two-step strategy, one first constructs a series of 2-D phase or group velocity maps for different periods tomographically at each geographic point of interest using arrival times measured at each period; the 1-D dispersion curve at each geographic point is then inverted to obtain a 1-D shear velocity model beneath this point. Those 1-D shear velocity structures are finally interpolated to construct a 3-D model (Nakanishi & Anderson 1983; Trampert & Woodhouse 1995). Since the two steps are conducted separately and sequentially, the solution to the 1-D depth inversion cannot interact directly with the 2-D phase or group velocity tomography step. In addition, in the second step the 1-D dispersion inversions are conducted separately, and consequently horizontal spatial correlation information can be lost in the obtained 3-D model. As a result, the results obtained using this strategy may be biased (Zhang *et al.* 2018). By contrast, in the one-step strategy one constructs a 3-D shear velocity model directly from frequency-dependent travel time measurements, which can naturally

impose spatial correlations in the model and produce more accurate results (Fang *et al.* 2015; Zhang *et al.* 2018). In either strategy, surface wave tomography involves solving one or a set of inverse problems.

Surface wave inverse problems are traditionally solved using optimization, in which one seeks an optimal solution by iteratively perturbing a set of starting model parameters to minimize the misfit between observed data and data predicted by the known physics from the model (Trampert & Woodhouse 1995; Ritzwoller *et al.* 2002). Since most tomographic problems are non-linear and under-determined, regularization is often imposed to incorporate additional constraints on model parameters. (Aki & Lee 1976; Dziewonski & Woodhouse 1987; Iyer & Hirahara 1993; Tarantola 2005). However, the regularization is often chosen by ad hoc means (often trial and error), and valuable information in the data can be concealed by the regularization (Zhdanov 2002). In addition, since optimization methods typically provide a single optimal solution, it is difficult to quantify uncertainty in these methods. This limitation is particularly problematic in fields like geophysics and engineering, where understanding the range of possible solutions and their likelihood is crucial for robust decision-making.

To address these challenges, Bayesian inference methods have been introduced to geophysics. Unlike traditional optimization methods, Bayesian inference treats the unknown parameters as random variables and uses Bayes' theorem to update a prior distribution with new information contained in the data to obtain a so-called posterior probability density functions (pdfs), which describes uncertainty in the solution. A common way to solve Bayesian inference problems is Markov chain Monte Carlo (MCMC) sampling. The method generates a set (or chain) of samples from the posterior pdf by taking a structured random walk through a parameter space (Geyer 2011; Bishop & Nasrabadi 2006; Andrieu & Thoms 2008); thereafter these samples can be used to estimate useful information about that pdf (mean, standard deviation, *etc.*). The Metropolis-Hastings (MH) algorithm (Chib & Greenberg 1995) is one such method, which has been widely applied in geophysics to solve non-linear inverse problems (Mosegaard & Tarantola 1995; Sambrige & Mosegaard 2002), such as seismic travel time tomography (Debski 2004; Zhang *et al.* 2020 a, b), surface wave dispersion inversion (Molnar *et al.* 2010; Gosselin *et al.* 2017), and full waveform inversion (Kotsi *et al.* 2020; Fu & Innanen 2021; Zhu 2018). The method has further been extended to trans-dimensional inversion, called the reversible jump MCMC (rj-MCMC) method (Green 1995), which allows for changes in the dimensionality of the model space,

and consequently the parameterization itself can be adapted to the data. The *rw*-MCMC algorithm has been widely used to solve 1-D dispersion inversion (Bodin *et al.* 2012; Young *et al.* 2013; Galetti *et al.* 2017; Zhao *et al.* 2022; Gao & Lekić 2018) and 2-D phase or group velocity tomographic problems (Bodin & Sambridge 2009; Galetti *et al.* 2015, 2017; Burdick & Lekić 2017; Zhao *et al.* 2022; Kästle *et al.* 2024). However, due to its random-walk behavior, the method becomes inefficient in high dimensional space. As a result, its application in one-step 3-D surface wave tomography has been limited (Zhang *et al.* 2018, 2019; Rahimi *et al.* 2024). To improve efficiency of MCMC sampling, more advanced methods have been introduced to geophysics, such as Hamiltonian Monte Carlo (Fichtner *et al.* 2019; Gebraad *et al.* 2020; Aleardi *et al.* 2020), stochastic Newton MCMC (Martin *et al.* 2012; Zhao & Sen 2021; Huang *et al.* 2020), Langevin Monte Carlo (Durmus *et al.* 2019; Izzatullah *et al.* 2021), and Gaussian mixture MCMC (Tilman *et al.* 2020; Zhang *et al.* 2021). Nevertheless, these methods still demand enormous computational resources, and therefore cannot be applied to one-step 3-D surface wave tomography.

Variational inference addresses Bayesian inference problems by approximating the posterior distribution within a specified family of pdfs. This is achieved by minimizing a measure of the difference between the posterior pdf and the approximating pdf, for example, the Kullback-Leibler (KL) divergence (Kullback & Leibler 1951). The method therefore employs an optimization-based approach rather than random sampling, and can potentially offer greater computational efficiency and improved scalability for high-dimensional problems compared to MCMC methods (Blei & Jordan 2006; Blei *et al.* 2017).

In variational inference, the choice of the approximating family of pdfs is important as it determines the accuracy of the results and the difficulty of the optimization problem. Different families can lead to different inversion results (Zhang *et al.* 2021). A common choice is to use a mean-field approximation family (Bishop & Nasrabadi 2006; Blei *et al.* 2017), which decomposes the joint probability distribution into a product of mutually independent marginal distributions (Jordan *et al.* 1999). In Geophysics the method has been used to invert for the spatial distribution of geological facies given seismic data (Nawaz & Curtis 2018, 2019; Nawaz *et al.* 2020). Nevertheless, the method neglects correlation information between parameters, and usually involves tedious derivations and bespoke implementations for each type of problem which restricts their applicability (Bishop & Nasrabadi 2006;

Blei *et al.* 2017; Nawaz & Curtis 2018, 2019). To make variational methods easier to use, Kucukelbir *et al.* (2017) proposed an automatic differential variational inference (ADVI) method based on a Gaussian variational family, which can easily be applied to many Bayesian inference 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). However, ADVI becomes expensive to compute in high dimensional space when using a full-rank covariance matrix. By using a mean-field approximation, the method can be more efficient, but ignores correlation between parameters and produces biased results (Zhang *et al.* 2023). To reduce this issue, Zhao & Curtis (2024b) introduced a method called Physically Structured Variational Inference (PSVI), by incorporating spatially localized correlation structures into the variational distribution, which takes into account only correlation of spatial proximity parameters. Nevertheless, all of these ADVI methods can lead to biased posterior estimates due to their implicit Gaussian approximation, for example, for multi-modal posterior distributions (Zhang *et al.* 2021).

Stein variational gradient descent (SVGD) provides a more accurate alternative for probabilistic inference by using a set of tunable particles (representative models) to represent the posterior pdf. By minimizing the KL divergence between the particle distribution and the target distribution, SVGD iteratively reshapes the particles ensemble to align with the posterior probability distribution (Liu & Wang 2016; Leviyev *et al.* 2022). Zhang & Curtis (2020a) applied SVGD to 2-D travel time tomography, and demonstrated that the method can provide accurate and efficient estimates of posterior distributions compared to MCMC methods. The method has later been applied to a variety of geophysical problems, including seismic travel time tomography (Zhang & Curtis 2023; Agata *et al.* 2023), earthquake location (Smith *et al.* 2022), hydrogeological inversion (Ramgraber *et al.* 2021), and full waveform inversion (Zhang & Curtis 2020b, 2021; Zhang *et al.* 2023).

Despite the growing interest in variational inference, the method is mainly used to solve 1-D or 2-D problems, and its application to 3-D tomographic problems remains limited (Zhang *et al.* 2021; Zhang *et al.* 2023; Yang *et al.* 2025; Zhao & Curtis 2025). In this study, we use variational inference methods to solve the one-step 3-D surface wave tomographic problems. In the following, we first describe the basic concept of 3-D surface wave tomography and variational inference methods, including ADVI, SVGD and stochastic SVGD (sSVGD, Gallego & Insua 2018). In section 3, we apply

the suite of methods to a synthetic 3-D surface wave tomography problem and compare their results to assess performance of those methods. In section 4, we use these methods to study the shear wave velocity structure of the crust and upper mantle in southwest China using Rayleigh wave dispersion data extracted from ambient noise (Liu *et al.* 2021). The results show that variational inference can obtain higher-resolution shear velocity models than those obtained using traditional methods, along with reliable uncertainty estimates. We therefore conclude that variational inference is practical and feasible in solving 3-D high dimensional surface wave tomographic problems, and can be applied fruitfully to image the Earth's subsurface.

2. Methods

2.1. Variational inference

Bayesian inference calculates the posterior probability density $p(\mathbf{m}|\mathbf{d}_{obs})$ for model parameters $\mathbf{m}$ by revising prior beliefs $p(\mathbf{m})$ in light of new data. This revision is governed by Bayes' theorem:

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

where $p(\mathbf{d}_{obs}|\mathbf{m})$ is the likelihood which describes the probability of observing data $\mathbf{d}_{obs}$ given $\mathbf{m}$, $p(\mathbf{m})$ represents the prior pdf which describes information that is known independently of data, and $p(\mathbf{d}_{obs})$ is a normalization factor called the evidence. The likelihood function is often assumed to be a Gaussian distribution around the data predicted from model $\mathbf{m}$ using the known physical relationship, as this is often a reasonable approximation to the pdf of noise or errors in the measured data.

Variational inference tackles Bayesian inference using optimization. The method approximates the posterior pdf $p(\mathbf{m}|\mathbf{d}_{obs})$ by finding a distribution $q^*(\mathbf{m})$ within a predefined family of known probability distributions $\mathcal{Q} = \{q(\mathbf{m})\}$ that minimizes the KL divergence between $q(\mathbf{m})$ and the true posterior pdf:

$$q^*(\mathbf{m}) = \arg \min_{q \in \mathcal{Q}} KL[q(\mathbf{m})||p(\mathbf{m}|\mathbf{d}_{obs})] \quad (2)$$

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

$$KL[q(\mathbf{m})||p(\mathbf{m}|\mathbf{d}_{obs})] = \mathbb{E}_q[\log q(\mathbf{m})] - \mathbb{E}_q[\log p(\mathbf{m}|\mathbf{d}_{obs})] \quad (3)$$

where the expectation is taken with respect to the distribution $q(\mathbf{m})$. The KL divergence is

nonnegative and only equals zero when $q(\mathbf{m}) = p(\mathbf{m}|\mathbf{d}_{obs})$ (Kullback & Leibler 1951). Expanding the posterior pdf using equation (1), the KL divergence becomes:

$$KL[q(\mathbf{m})||p(\mathbf{m}|\mathbf{d}_{obs})] = \mathbb{E}_q[\log q(\mathbf{m})] - \mathbb{E}_q[\log p(\mathbf{m}, \mathbf{d}_{obs})] + \log p(\mathbf{d}_{obs}) \quad (4)$$

However, the evidence term $\log p(\mathbf{d}_{obs})$ is intractable, as it involves a high-dimensional integral whose evaluation scales exponentially with the number of model parameters. We thus rearrange equation (4) to obtain the evidence lower bound (ELBO):

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

Since the $KL[q||p] \geq 0$, equation (5) defines a lower bound for the evidence $\log p(\mathbf{d}_{obs})$. Because the evidence term is a constant for a given problem, minimizing the KL-divergence is equivalent to maximizing the ELBO. Consequently, variational inference in equation (2) can be expressed as:

$$q^*(\mathbf{m}) = \arg \max_{q \in \mathcal{Q}} \text{ELBO}[q(\mathbf{m})] \quad (6)$$

The selection of the variational family $\mathcal{Q}$ is critical as it determines both the approximation accuracy and the computational tractability of the optimization problem. Different choices of $\mathcal{Q}$ lead to different algorithms. In this study we describe two types of approaches: automatic differentiation variational inference (ADVI), and Stein variational gradient descent (SVGD), including its stochastic variant (sSVGD), and evaluate their performance in 3D surface wave inversion problems.

2.2 Automatic differential variational inference (ADVI)

ADVI utilizes Gaussian distributions as variational family (Kucukelbir *et al.* 2017). Since geophysical parameters often exhibit hard constraints (e.g., seismic velocity is always positive), we first use an invertible transform $T: \theta = T(\mathbf{m})$ to map constrained parameters $\mathbf{m}$ into an unconstrained space θ . A commonly-used transform is logit function (Team *et al.* 2016; Zhang & Curtis 2020a):

$$m_i = f(\theta_i) = a_i + \frac{b_i - a_i}{1 + \exp(-\theta_i)}$$

$$\theta_i = f^{-1}(m_i) = \log(m_i - a_i) - \log(b_i - m_i) \quad (7)$$

where m_i is the i^{th} model parameter bounded by the lower and upper bounds a_i and b_i respectively, θ_i is the transformed variable in an unconstrained space. In this unconstrained space the joint probability $p(\mathbf{m}, \mathbf{d}_{obs})$ becomes:

$$p(\theta, \mathbf{d}_{\text{obs}}) = p(\mathbf{m}, \mathbf{d}_{\text{obs}}) \left| \det J_{T^{-1}}(\theta) \right| \quad (8)$$

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

$$q(\theta; \xi) = \mathcal{N}(\theta | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\theta | \boldsymbol{\mu}, \mathbf{L}\mathbf{L}^\top) \quad (9)$$

where ξ represents variational parameters, that is the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$.

The covariance matrix is reparameterized using a Cholesky factorization $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$ with a lower-triangular matrix $\mathbf{L}$ to ensure positive semidefiniteness. The off-diagonal values of $\mathbf{L}$ describe correlation between model parameters. However, a full-rank covariance matrix becomes computationally intractable for high dimensional problems (as in 3D FWI; Zhang *et al.* 2023) as it requires $n(n+1)/2$ parameters to construct $\mathbf{L}$ where n is the number of parameters. To reduce this issue, a mean-field approximation may be employed by using a diagonal $\mathbf{L}$ (Kucukelbir *et al.* 2017), but this choice tends to strongly underestimate posterior parameter uncertainties because correlation between parameters is neglected (Kucukelbir *et al.* 2017; Zhang *et al.* 2023).

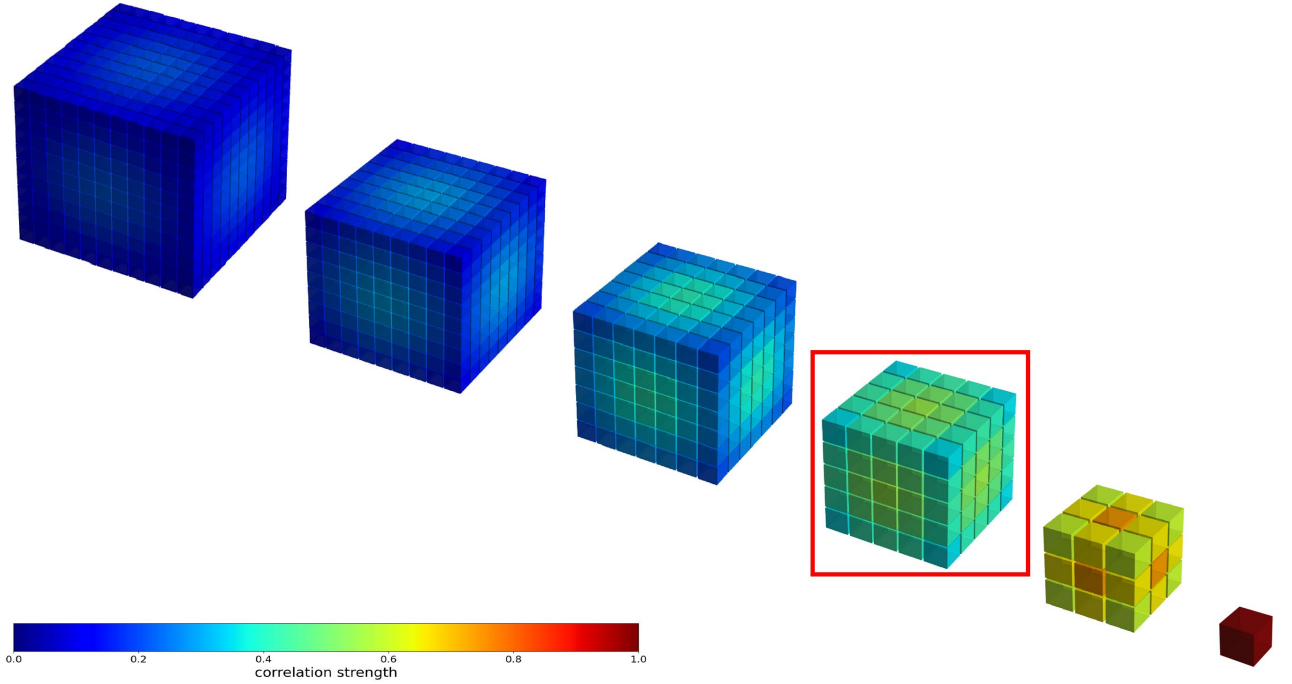


Figure 1. The parameter correlation strength in a 3D model discretized into square grid cells ($n_x \times n_y \times n_z$), with the central cell (highlighted in red) serving as a reference. The correlation magnitude between the central cell and surrounding cells exhibits an inverse relationship with spatial distance – cells proximal to the central reference show stronger correlation, while those at greater distances display progressively weaker association. This indicates that distal cells maintain only negligible correlation with the central cell.

Alternatively, one can construct an $\mathbf{L}$ which only captures correlation between spatially proximate volumes (Zhao & Curtis 2024b), as in reality the magnitude of correlation between two cells often decreases with their distance apart (Fig. 1). Thus, for an n -dimensional problem displayed in Fig. 1 ($n = nx \times ny \times nz$), $\mathbf{L}$ can be set as a sparse matrix to capture only correlation between proximate cells (see Appendix A for details on constructing $\mathbf{L}$). By doing this, the covariance matrix $\Sigma = \mathbf{L}\mathbf{L}^T$ only contains correlation between pairs of adjacent cells in each direction (cells surrounding the central red cell in Fig. 1). The method therefore captures the structural correlation of actual physics, which is referred as physically structured variational inference (PSVI), and has been shown to produce more accurate results than mean-field ADVI, and are more computational efficient than full-rank ADVI (Zhao & Curtis 2024b).

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

$$\xi^* = \arg \max_{\xi} \text{ELBO}[q(\theta; \xi)] = \arg \max_{\xi} \mathbb{E}_q[\log p(T^{-1}(\theta), \mathbf{d}_{\text{obs}}) + \log |\det \mathbf{J}_{T^{-1}}(\theta)|] - \mathbb{E}_q[\log q(\theta; \xi)] \quad (10)$$

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 μ and $\mathbf{L}$ are:

$$\nabla_{\mu} \text{ELBO} = \mathbb{E}_{N(\eta|0, \mathbf{I})} \left[\nabla_{\mathbf{m}} \log p(\mathbf{m}, \mathbf{d}_{\text{obs}}) \nabla_{\theta} T^{-1}(\theta) + \nabla_{\theta} \log |\det \mathbf{J}_{T^{-1}}(\theta)| \right] \quad (11)$$

$$\nabla_{\mathbf{L}} \text{ELBO} = \mathbb{E}_{N(\eta|0, \mathbf{I})} \left[\left(\nabla_{\mathbf{m}} \log p(\mathbf{m}, \mathbf{d}_{\text{obs}}) \nabla_{\theta} T^{-1}(\theta) + \nabla_{\theta} \log |\det \mathbf{J}_{T^{-1}}(\theta)| \right) \eta^T \right] + (\mathbf{L}^{-1})^T \quad (12)$$

where η is a random variable generated from a standard Normal distribution $N(\mathbf{0}, \mathbf{I})$. The expectations in the above equations can be estimated by Monte Carlo (MC) integration, which in practice only requires a low number of samples because the optimization is performed over many iterations so that statistically the gradients will lead to convergence toward the correct solution (Kucukelbir *et al.* 2017). The final approximation $q^*(\mathbf{m})$ is obtained by transforming $q^*(\theta)$ back to the original space.

2.3 Stein variational gradient descent (SVGD)

SVGD approximates the target posterior pdf using a set of models (called particles) $\{\mathbf{m}_i\}_{i=1}^n$. In SVGD, these particles are transported toward the posterior distribution by iteratively minimizing the KL divergence between the particle distribution and the posterior pdf (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). SVGD updates particles using a smooth perturbation:

$$T(\mathbf{m}_i) = \mathbf{m}_i + \epsilon \boldsymbol{\phi}(\mathbf{m}_i) \quad (13)$$

where $\boldsymbol{\phi} = [\phi_1, \dots, \phi_n]$ 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):

$$\nabla_{\epsilon} \text{KL}[q_T \parallel p] \Big|_{\epsilon=0} = -\mathbb{E}_q \left[\text{trace} \left(A_p \boldsymbol{\phi}(\mathbf{m}) \right) \right] \quad (14)$$

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

Liu & Wang (2016) showed that the optimal $\boldsymbol{\phi}^*$ which maximizes the expectation in equation (14) can be derived using kernel functions. Say $x, y \in X$ and define a mapping ψ from X to a space where an inner product $\langle \cdot, \cdot \rangle$ 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):

$$\boldsymbol{\Phi}^* \propto E_{\{\mathbf{m}' \sim q\}} \left[A_p k(\mathbf{m}', \mathbf{m}) \right] \quad (15)$$

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

$$\boldsymbol{\phi}_l^*(\mathbf{m}) = \frac{1}{n} \sum_{j=1}^n \left[k(\mathbf{m}_j^l, \mathbf{m}) \nabla_{\mathbf{m}_j^l} \log p(\mathbf{m}_j^l | \mathbf{d}_{\text{obs}}) + \nabla_{\mathbf{m}_j^l} k(\mathbf{m}_j^l, \mathbf{m}) \right] \quad (16)$$

$$\mathbf{m}_i^{l+1} = \mathbf{m}_i^l + \epsilon^l \boldsymbol{\phi}_l^*(\mathbf{m}_i^l) \quad (17)$$

where superscript l denotes the l^{th} iteration and n is the number of particles. 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):

$$k(\mathbf{m}, \mathbf{m}') = \exp \left[-\frac{\|\mathbf{m} - \mathbf{m}'\|^2}{2h^2} \right] \quad (18)$$

where h is a bandwidth parameter that controls the intensity of particle interactions depending on

distance. As suggested by several studies (Liu & Wang 2016; Zhang & Curtis 2020a,b), we choose $h = \tilde{d} / \sqrt{2 \log n}$ where $\tilde{d}$ is the median of pairwise distances between all particles. This choice ensures balanced contributions between each particle's local gradient and gradients from all other particles.

Note that for the RBF kernel, the second term of ϕ^* in equation (16) becomes $\sum_j \frac{\mathbf{m} - \mathbf{m}_j}{h^2} k(\mathbf{m}_j, \mathbf{m})$ 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 toward high probability areas.

2.4 Stochastic SVGD

Despite its wide applications (Gong *et al.* 2019; Zhang & Curtis 2020a,b; Pinder *et al.* 2020), SVGD can produce biased estimations in high-dimensional space due to finite-particle implementation and computational constraints (Ba *et al.* 2021). To improve accuracy, Gallego & Insua (2018) proposed a variant of SVGD, called stochastic SVGD (sSVG), which transforms SVGD into an MCMC method with multiple interacting chains by introducing a Gaussian noise term to the dynamics of SVGD. 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 particles from the beginning. In addition, this method guarantees asymptotic convergence as the number of iterations tends to infinity, which cannot be achieved by SVGD with finite number of particles (Leviyev *et al.* 2022).

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

$$d\mathbf{z} = \mathbf{f}(\mathbf{z})dt + \sqrt{2\mathbf{D}(\mathbf{z})}d\mathbf{W}(t) \quad (19)$$

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 (Oksendal 2013). If we denote the posterior distribution as $p(\mathbf{z})$, Ma *et al.* (2015) proposed a SDE that converges to the distribution $p(\mathbf{z})$:

$$f(\mathbf{z}) = [\mathbf{D}(\mathbf{z}) + \mathbf{Q}(\mathbf{z})] \nabla \log p(\mathbf{z}) + \Gamma(\mathbf{z}) \quad (20)$$

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.

By discretizing the SDE using Euler–Maruyama, we obtain a practical algorithm:

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

where $\mathcal{N}(\mathbf{0}, 2\epsilon_t \mathbf{D}(\mathbf{z}_t))$ represents a Gaussian distribution. The gradient $\nabla \log p(\mathbf{z}_t)$ can be computed using the entire data set, or uniformly randomly selected mini-batches of the data which produces a stochastic gradient approximation. In either case the above process converges to the posterior distribution asymptotically as $\epsilon_t \rightarrow 0$ and $t \rightarrow \infty$ (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 $\mathbf{Q} = \begin{bmatrix} \mathbf{0} & -\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}$, the stochastic Hamiltonian Monte Carlo (HMC) method can be derived (Chen *et al.* 2014).

In sSVGD, for the set of particles $\{\mathbf{m}_i\}$ defined in the above section, we construct an augmented space $\mathbf{z} = (\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_n) \in \mathbb{R}^{nd}$ that concatenates the n particles, and use equation (21) to obtain a valid sampler that runs multiple (n) interacted chains in parallel to generate samples from the posterior distribution $p(\mathbf{z}) = \prod_{i=1}^n p(\mathbf{m}_i | \mathbf{d}_{obs})$. Define a matrix $\mathbf{K}$:

$$\mathbf{K} = \frac{1}{n} \begin{bmatrix} k(\mathbf{m}_1, \mathbf{m}_1) \mathbf{I}_{d \times d} & \cdots & k(\mathbf{m}_1, \mathbf{m}_n) \mathbf{I}_{d \times d} \\ \vdots & \ddots & \vdots \\ k(\mathbf{m}_n, \mathbf{m}_1) \mathbf{I}_{d \times d} & \cdots & k(\mathbf{m}_n, \mathbf{m}_n) \mathbf{I}_{d \times d} \end{bmatrix} \quad (23)$$

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). Let $\mathbf{D}_k = \mathbf{K}$, $\mathbf{Q}_k = \mathbf{0}$, equation (21) becomes:

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

Comparing this equation with equations (16) and (17), SVGD can be regarded as a special case of equation (24) without Gaussian noise term. This equation therefore defines a stochastic gradient MCMC method with SVGD gradients, which is called stochastic SVGD (sSVGD). According to the discussion above, this process converges to the posterior distribution $p(\mathbf{z}) = \prod_{i=1}^n p(\mathbf{m}_i | \mathbf{d}_{obs})$ asymptotically. Note that when the number of particles is large enough, the noise term would be tiny according to equation (23). Consequently in such case the method produces the same results as standard SVGD.

In order to use equation (24) to sample the posterior distribution, we need to draw samples from the Gaussian distribution $\mathcal{N}(0, 2\epsilon_t \mathbf{K})$. 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 $\mathbf{D}_k$:

$$\mathbf{D}_k = \frac{1}{n} \begin{bmatrix} \overline{\mathbf{K}} & & \\ & \ddots & \\ & & \overline{\mathbf{K}} \end{bmatrix} \quad (25)$$

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

$$\mathbf{P} = \begin{bmatrix} 1 & & & & & & & & & \\ & & 1 & & & & & & & \\ & & & & \ddots & & & & & \\ - & - & - & - & - & - & 1 & - & - & \\ & 1 & & & & & & & & \\ & & & 1 & & & & & & \\ & & & & \ddots & & & & & \\ - & - & - & - & - & - & 1 & - & - & \\ & \ddots & & \ddots & \ddots & \ddots & \ddots & & & \\ - & - & 1 & - & - & - & - & - & - & \\ & & & & 1 & & & & & \\ & & & & & \ddots & & & & \\ & & & & & & & & & 1 \end{bmatrix} \quad (26)$$

This permutation matrix transforms vector $\mathbf{z}$ from particle-major ordering (sequential particle listing) to dimension-major ordering (first coordinates of all particles, followed by second coordinates, etc.). The noise term η can therefore be generated using:

$$\begin{aligned} \eta &\sim \mathcal{N}(0, 2\epsilon_t \mathbf{K}) \\ &\sim \sqrt{2\epsilon_t} \mathbf{P}^T \mathbf{P} \mathcal{N}(0, \mathbf{K}) \\ &\sim \sqrt{2\epsilon_t} \mathbf{P}^T \mathcal{N}(0, \mathbf{D}_k) \\ &\sim \sqrt{2\epsilon_t} \mathbf{P}^T \mathbf{L}_{\mathbf{D}_k} \mathcal{N}(0, \mathbf{I}) \end{aligned} \quad (27)$$

Given the block-diagonal structure of $\mathbf{D}_k$, its Cholesky factor $\mathbf{L}_{\mathbf{D}_k}$ can be calculated easily as we only need to calculate the lower triangular Cholesky decomposition of matrix $\overline{\mathbf{K}}$. Since the number of particles n is typically modest in practice, evaluation of the noise term is computationally negligible. Equation (24) can therefore be directly employed to sample the posterior distribution.

2.5 3D variational surface wave tomography

In order to use variational inference methods in 3D surface wave tomography, it is required to efficiently compute travel times and their gradients with respect to model parameters at varying periods along arbitrary source-receiver paths. Ideally a fully 3-D wavefield simulation (e.g., spectral-element method) could be used but these are generally computationally too expensive (Chen *et al.* 2014; Gao & Shen 2014). To balance accuracy and computational efficiency, we adopt a ray tracing-based approximation method (Fang *et al.* 2015), which first calculates phase or group velocity dispersion curves at each geographic point using the 1-D velocity profile beneath this point, and then computes source-receiver travel times at each period using the calculated phase or group velocity map at this period. Under this approximation, for a logarithm posterior pdf function $\mathbf{F}$ (e.g., $\mathbf{F} = \log p(\mathbf{m}, \mathbf{d}_{obs})$) as in equation 12), the gradient $\nabla_{\mathbf{m}} \mathbf{F}$ can be calculated using chain rule:

$$\nabla_{\mathbf{m}} \mathbf{F} = \sum_i^N \sum_{\tau}^T \frac{\partial F_{i,\tau}}{\partial t_i(\tau)} \frac{\partial t_i(\tau)}{\partial C_i(\tau)} \frac{\partial C_i(\tau)}{\partial \mathbf{m}} \quad (28)$$

where N and T represent the number of source-receiver paths and discrete periods in surface wave dispersion measurements, respectively. $\frac{\partial F_{i,\tau}}{\partial t_i(\tau)}$ quantifies the sensitivity of the function $\mathbf{F}$ to the i th phase travel-time perturbations at each measurement period. $\frac{\partial t_i(\tau)}{\partial C_i(\tau)}$ is calculated using the fast marching method (Rawlinson & Sambridge 2004), and $\frac{\partial C_i(\tau)}{\partial \mathbf{m}}$ represents the Frechét kernel of phase velocity with respect to model parameters (e.g., shear wave velocity), computed using matrix propagator methods (Herrmann 2013).

3. Synthetic tests

We first apply the suite of methods to a 3D synthetic example. The true model is composed of three homogeneous layers with shear wave velocities of 3.0, 4.0 and 5.0 km/s, respectively. In the second

layer we constructed a square high-speed anomaly (4.5 km/s) and a low-speed anomaly (3.5 km/s) in two specified locations (Fig. 2a and b). The P-wave velocity (v_p) is determined by a fixed $\frac{v_p}{v_s}$ ratio with a typical crustal value of 1.73. Density ρ is expressed as a function of v_p through an empirical relationship: $\rho = 2.35 + 0.036 \times (v_p - 3.0)^2$ (Kurita 1973). We deployed 173 receivers whose locations are set as the actual station distribution in Southwest China (Liu *et al.* 2021). Each receiver is also treated as a virtual source to simulate a typical ambient noise tomography (Shapiro *et al.* 2005). We discretized the velocity model using a $27 \times 23 \times 71$ grid, and calculated the phase velocity travel-times of Rayleigh waves at periods from 5 to 50 s between each station pair using the forward method described above. In addition, two percent Gaussian noise is added to the data. This produces a total of 14,878 inter-receiver travel-times for each period.

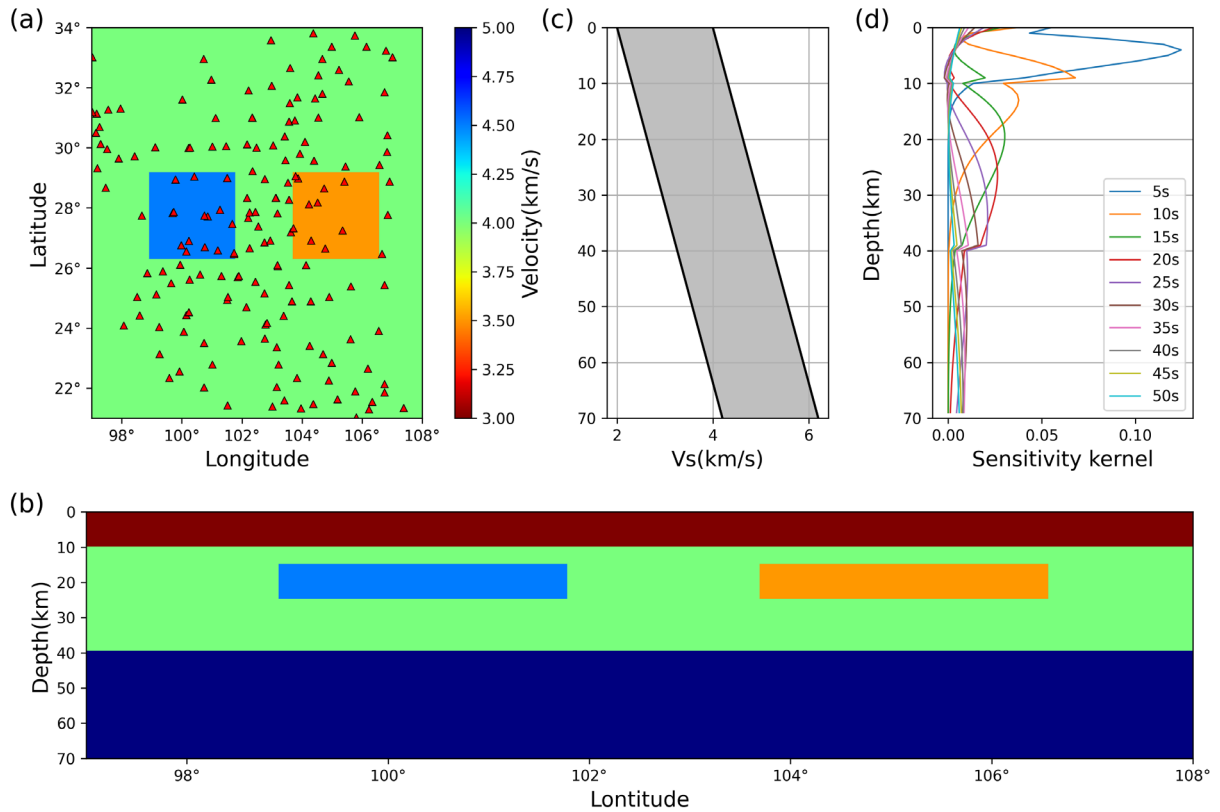


Figure 2. True model used for the synthetic test. Red triangles show the locations of receivers. (a) Horizontal section of the true model at depth of 20 km; (b) Vertical section at latitude=27.5°N; (c) Prior distribution used in the inversion: a Uniform distribution with a width of 2.2 km/s at each depth; (d) Phase velocity sensitivity kernels calculated for a 1D profile with the three homogeneous layers (without those anomalies in the second layer).

For variational inversion, the velocity model $\mathbf{m}$ is parameterized on a regular $0.5^\circ \times 0.5^\circ$ horizontal grid, with vertical intervals of 5 km between 0 and 40 km depth and 10 km between 40 and 70 km depth. The noise level is fixed to be 2% of travel-times for all inversions. The prior pdf of the velocity in each cell is set to be a Uniform distribution over an interval of 2.2 km/s at each depth to encompass the true model (Fig. 2c). In addition, we impose spatial smoothness constraints on model parameters. To achieve this, we define a second-order finite-difference operator $\mathbf{S}$ that calculates the curvature of model parameters between adjacent grid cells in the 3D velocity model. By assuming a Gaussian distribution for the product $\mathbf{Sm}$, we enforce smoothness through:

$$p(\mathbf{Sm}) \propto \exp\left(-\frac{1}{2\sigma_s^2}(\mathbf{Sm})^T(\mathbf{Sm})\right) \quad (29)$$

where σ_s is a hyperparameter controlling the strength of spatial smoothing. Equation (29) can be interpreted as applying Tikhonov regularization with matrix $\mathbf{S}$ to $\mathbf{m}$ (Golub *et al.* 1999). By combining with the Uniform prior distribution $p(\mathbf{m})$, the resulting smoothed prior pdf can be expressed as:

$$P_{\text{smooth}}(\mathbf{m}) \propto p(\mathbf{Sm})p(\mathbf{m}) \quad (30)$$

Here we choose different σ_s for different directions, that is, $\sigma_x=0.2$, $\sigma_y=0.2$, $\sigma_z=1.0$ in the latitude, longitude, and depth direction, respectively.

For mean field ADVI and PSVI, we set the initial Gaussian distribution in the unconstrained space to be a standard Normal distribution $N(\theta|0, I)$. In PSVI we employ a $5 \times 5 \times 5$ correlation kernel to model the main correlation between model parameters (see Fig. 1 and Appendix A). The initial distribution is then updated using the ADAM algorithm (Kingma & Ba 2014) for 2,000 iterations for each method. At each iteration the gradients are calculated using eight Monte Carlo samples. The obtained probability distribution is transformed back to the original space to obtain the final approximation, from which we generate 1,000 samples to compute the mean and standard deviation to visualize the results.

For sSVGD we start from 40 particles that are generated from the prior distribution, and transform them to the unconstrained space using equation (7). Those particles are then updated (sampled) using equation (24) for 2,000 iterations with a burn-in period of 1,000. To reduce the memory and storage cost, we only retain every second sample after the burn-in period. This results in a total of 20,000

samples, which are transformed back to the original space to calculate statistics of the estimated posterior pdf.

3.1 Model comparison

Fig. 3 shows horizontal profiles of the results at the depth of 20 km obtained using the suite of methods. Overall, the mean models capture the true structure, including the high- and low-velocity anomalies in the second layer. However, the velocity magnitudes are slightly different — the velocity of the high-velocity anomaly is lower (≈ 4.4 km/s) than the true value (4.5 km/s), and the velocity of the low-velocity anomaly is higher (≈ 3.6 km/s) than the true value (3.5 km/s). This may reflect uncertainty of the problem itself or may be caused by the smooth prior constraint (see discussion below and [Appendix B](#)). The standard deviation models obtained using the three methods generally exhibit lower uncertainty in the central region and higher uncertainty along the periphery as expected, which reflects the insufficient data coverage outside the station array. In addition, the results obtained using mean-field ADVI show smaller uncertainty than those from PSVI and sSVGD because of absence of correlation in mean-field ADVI (Blei *et al.* 2017; Zhang *et al.* 2021). Moreover, while mean-field ADVI yields generally low standard deviation within the array, PSVI and sSVGD show more complex features. For example, both results obtained using PSVI and sSVGD exhibits relatively lower standard deviation at the location of the high velocity anomaly because of relatively denser station distribution, or may be caused by the fact that rays concentrate at high velocity anomalies. Note that there are small scale structures in all three standard deviation models which likely reflects the uncertainty of the problem itself.

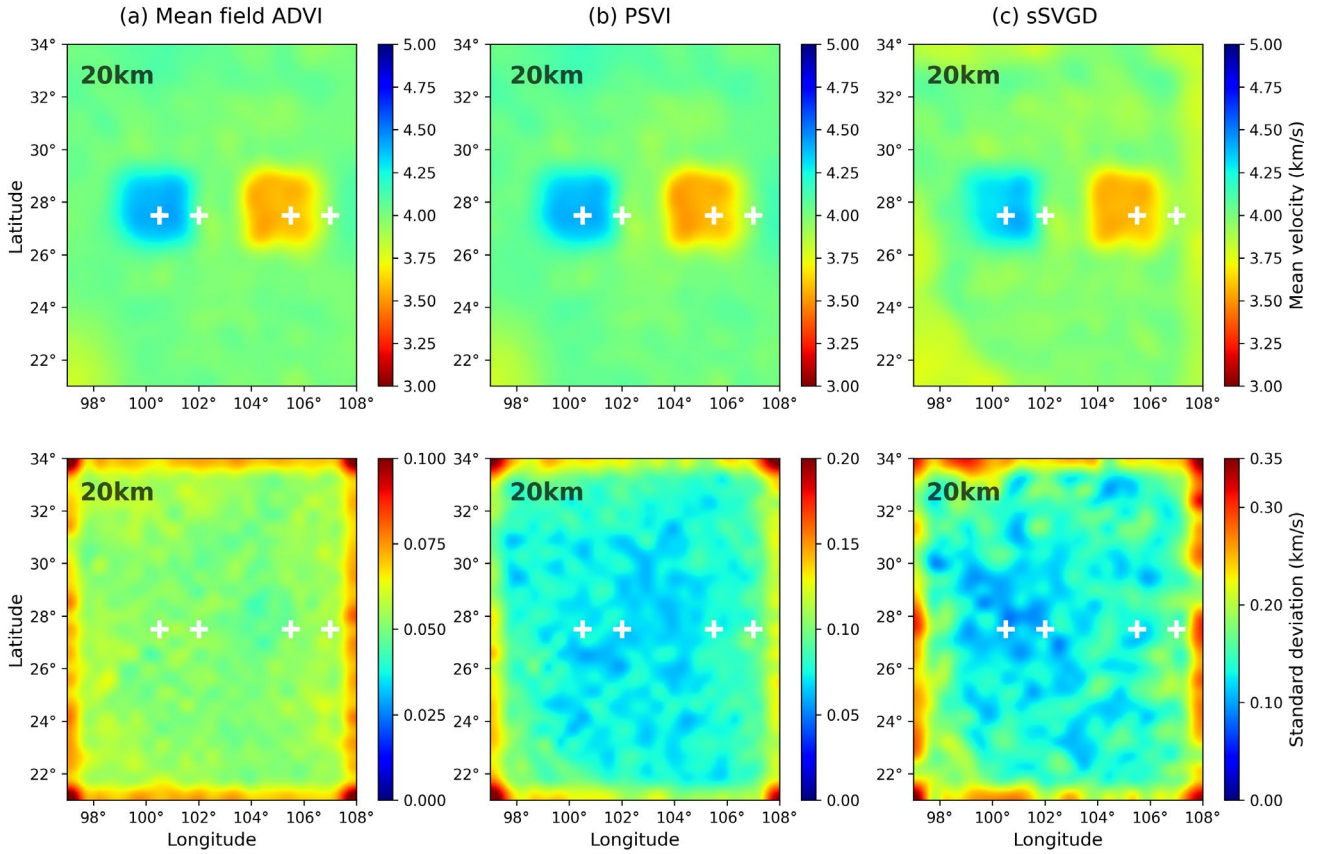


Figure 3. The mean (top row) and standard deviation (bottom row) at the horizontal slice of depth = 20 km obtained using (a) mean field ADVI, (b) PSVI, and (c) sSVGD, respectively. The white pluses show the point location which are referred to in the text.

Fig. 4 shows the results at the vertical profile of latitude 27.5°N. Overall, the mean models show similar structure to that of the true model, including the three layers and the two rectangle anomalies within the second layer. The top layer is clearly recovered due to the high sensitivity of surface waves at shallow depth (Fig. 2d), except that the layer boundary is deeper at two sides because of low data coverage in those areas. Within the second layer, the location of the low velocity anomaly is slightly deeper than that of the high velocity anomaly in all three results. Given that the methods are quite different in their theoretical foundation, the results likely reflect the uncertainty of the problem itself. The standard deviation models generally show increasing uncertainty with depth, which reflects the decreasing sensitivity of surface waves when depth increases. Note that the uncertainty is lower in the bottom layer, which is likely due to the cumulative sensitivity at longer periods for depths beyond 70 km because of the imposed half-space boundary at this depth. Similar phenomenon has also been observed in several previous studies (Bodin *et al.* 2012, Zhang *et al.* 2018). There are also high uncertainties around 10km and 40 km depth across all standard deviation models, which is consistent

with the rapid decrease in the sensitivity kernels of surface waves across different periods (Fig. 2d).

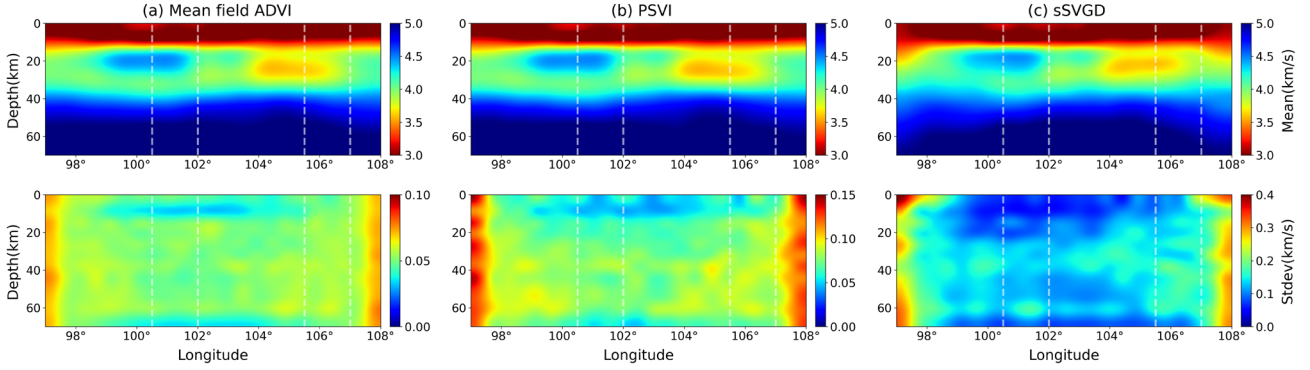


Figure 4. The mean (top row) and standard deviation (bottom row) at the vertical slice of Latitude=27.5°N obtained using (a) mean field ADVI, (b) PSVI, and (c) sSVGD, respectively. White dashed lines denote the 1D profile location referred to in the main text.

To further evaluate the results, in Fig. 5 we compare the posterior marginal distributions derived from the three methods along four vertical profiles (white pluses in Fig. 3 and white dashed lines in Fig. 4). Overall, the marginal distributions obtained using the three methods show similar features. For example, all results show generally high probability around the true value. However, there exists details that differ in the results. For example, the distribution obtained using sSVGD are generally wider than those from ADVI and PSVI, which suggests that ADVI and PSVI may underestimate the uncertainty because of their Gaussian assumption (Zhang & Curtis 2020b). In addition, along the third profile between depths of 15–20 km and 25–30 km, the true velocity values fall outside the high-probability regions in the results from ADVI and PSVI (Fig. 5a and 5b), which further indicates that these methods can underestimate uncertainty. By contrast, sSVGD produces more reasonable uncertainty estimates as the distribution generally encompasses the true value within regions of non-zero probability. At around 40 km depth, the true velocity lies outside the high-probability region in all results, which is consistent with the high uncertainties shown in Fig. 4. This is likely caused by the lower sensitivity at this depth and the smoothness prior constraints (see Appendix B).

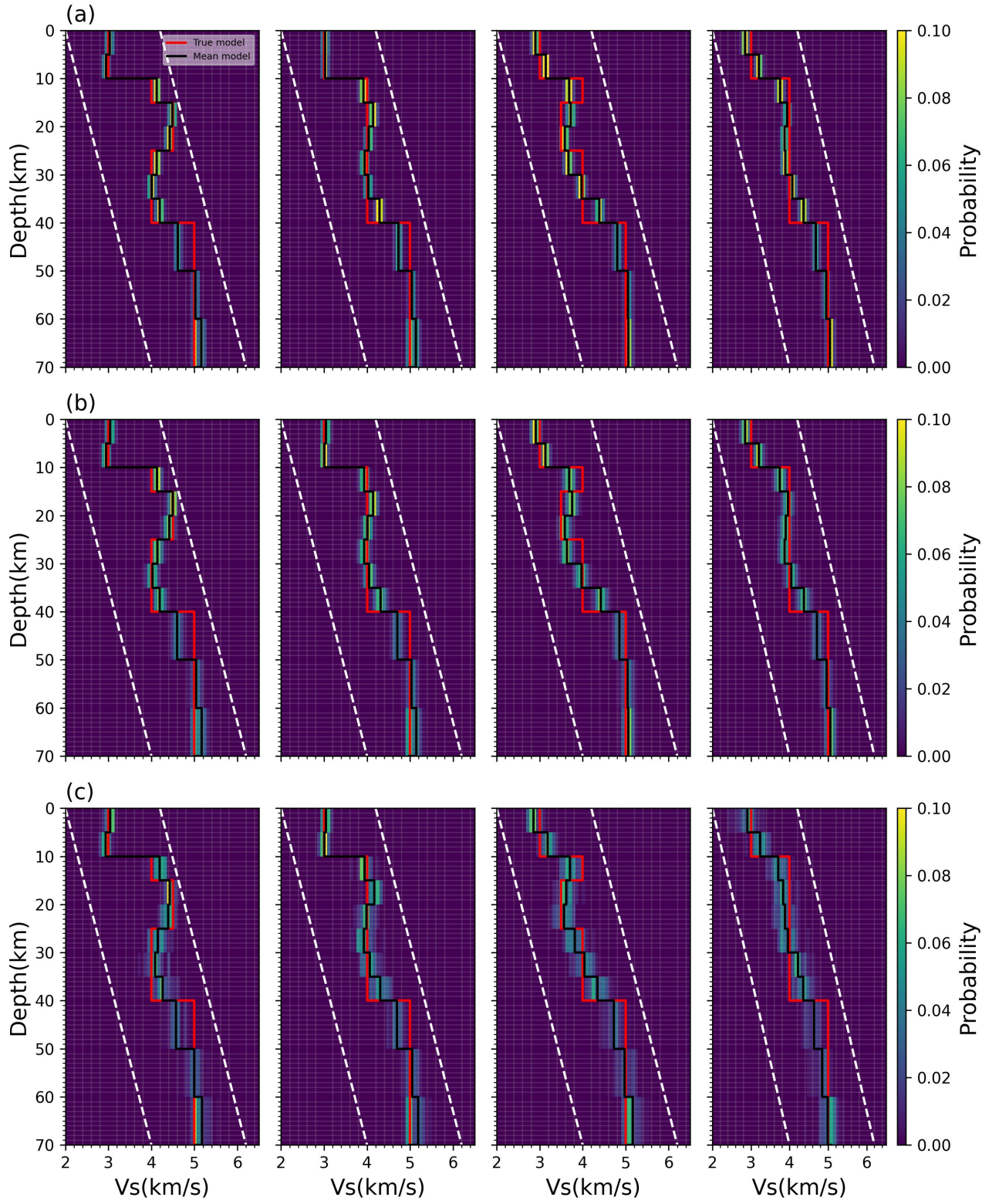


Figure 5. The marginal distributions at four vertical profiles obtained using (a) mean field ADVI, (b) PSVI and (c) sSVG respectively. The locations of these four profiles are represented by white pluses and dashed lines in Figure 3 and 4. White dashed lines show the lower and upper bound of the prior distribution, and black and red lines show the mean and true velocity values.

Strong prior information tends to favor overly smoothed velocity models, as strong smoothness constraints limit the velocity differences between adjacent grid cells. To study effects of the smoothness prior constraints, we compare results using both smoothed and non-smoothed prior pdfs for each method, and analyze their respective consequences.

The results show that the mean velocity models obtained using smoothed prior pdfs exhibit greater spatial continuity with reduced small-scale variations compared to those obtained using non-smoothed prior pdfs (Fig. B1 in appendix B), and the standard deviations obtained using smoothed prior pdfs show lower uncertainty as large velocity contrasts between neighbouring cells are precluded by the prior information. In addition, the results obtained using non-smoothed prior pdfs show a clear low-uncertainty anomaly at the location of the high-velocity anomaly, which becomes less pronounced after applying smoothing constraints. This is likely because the Gaussian smoothing enhances spatial correlations between adjacent cells, and thereby reduces contrast in standard deviation values across velocity boundaries. Similarly, the posterior marginal pdfs obtained using smoothed prior pdfs display narrower probability ranges (Fig. B2).

However, from a Bayesian perspective, without prior knowledge of the true Earth's smoothness, there is no objective reason to prefer any one of these posterior distributions – each reflects a different aspect of posterior uncertainty. While prior intuition may suggest that the models in the top panels of Fig. B1 are under-smoothed, establishing an appropriate prior probability distribution over smoothness levels remains challenging. In such cases, one may try to perform a set of inversions with different smoothness levels, each of which represents a prior belief of smoothness of the subsurface.

3.2 Computational cost

Table 1 summarizes the number of simulations and the CPU hours required for each method. The number of simulations serves as a reliable metric for comparing overall computational cost, since forward simulations constitute the most time-consuming component in all methods. Given that all approaches can be fully parallelized, we also report wall time to provide further insight into their practical computational requirements.

Table 1. A comparison of computational cost for the 3 inference methods.

Method	Number of simulations	CPU hours ^a	Wall time (hours)
--------	-----------------------	------------------------	-------------------

Mean field ADVI	16000	477.74	23.89
PSVI	16000	586.09	29.30
sSVGD	80000	1633.79	65.35

^a The CPU used in this study is Intel Xeon Gold 6258R.

The results indicate that mean-field ADVI is the least computationally expensive method, which required 16,000 simulations with 477.74 CPU hours; however, as demonstrated above, it tends to produce systematically biased results. PSVI is nearly as efficient as mean-field ADVI, which took the same number of simulations with slightly more CPU hours (586.09) because of larger covariance matrix computation, yet it improves accuracy and delivers uncertainty estimates comparable to those of sSVGD. This suggests that PSVI method can be useful for obtaining initial, relatively rapid estimates of subsurface structure. sSVGD took the most computational cost, but likely produces the most accurate results as it does not make Gaussian assumption as in mean-field ADVI and PSVI. 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. Nonetheless, the results offer a reasonable indication of the relative efficiency of each method.

4. Application to Southwest China

In this section, we apply the above methods to data collected in Southwest China at the southeastern margin of the Tibetan plateau, which plays an important role for transferring lithospheric materials extruded from the Tibetan plateau to Southeast Asia. This region is characterized by several major fault systems, including the Longmenshan thrust fault (LMSF), the Xianshuihe (XSHF) – Anninghe – Zemuhe – Xiaojiang (XJF) sinistral strike-slip fault system, and the Red River dextral strike-slip fault (RRF) (Fig. 6). These major fault systems accommodate most of the large earthquakes in the region and separate the area into different tectonic units, such as the lens-shaped Chuandian block (bounded by XSHF, XJF, and RRF) (Kan *et al.* 1977; Wang *et al.* 1998), the Songpan-Ganzi block (SGB), the Yangtze craton (YZC), the Qiangtang block (QB), the Lhasa block (LB) and the Shan Thai block (STB) (Fig. 6).

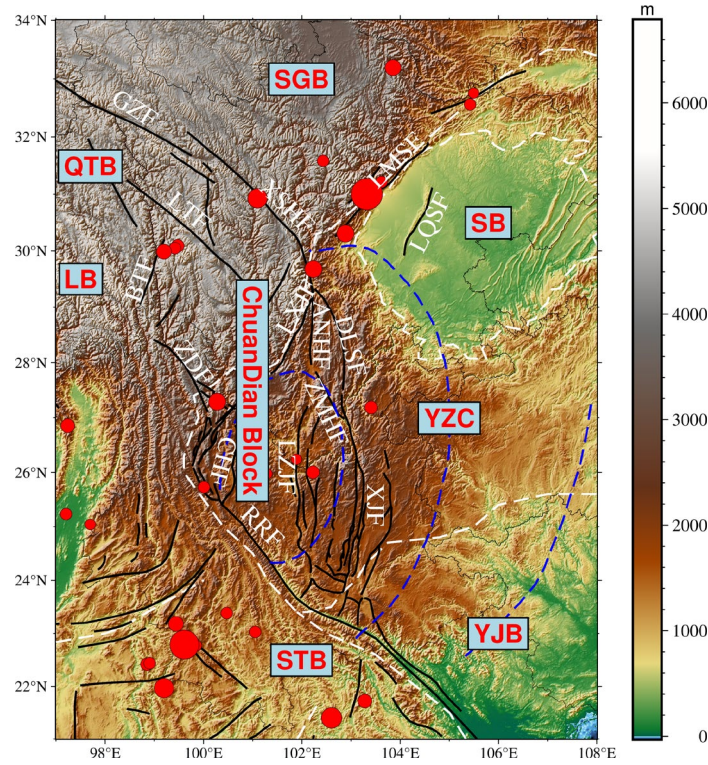


Figure 6. Tectonic elements and fault systems in the study region. The dashed white lines indicate tectonic block boundaries. LB, Lhasa block; QTB, Qiangtang block; SB, Sichuan basin; SGB, Songpan-Ganzi block; STB, Shan Thai block; YJB, Youjiang block; YZC, Yangtze craton. The thick solid black lines show major faults. ANHF, Anninghe fault; BTF, Batang fault; GZF, Ganzi fault; LTF, Litang fault; LJXJF, Lijiang-Xiaojinhe fault; LZJF, Lüzhijiang fault; RRF, Red River fault; XJF, Xiaojiang fault; XSHF, Xianshuihe fault; ZDF, Zhongdian fault; ZMHF, Zemuhe fault; DLSE, Daliangshan fault; CHF, Chenghai fault; LQSF, Longquanshan fault. The red filled circles show epicenters of earthquakes ($M_w \geq 6.0$) between 1980 and 2025. The blue dashed lines outline the boundary between the inner, intermediate, and outer zones of the Emeishan large igneous province (ELIP) (Xu *et al.* 2004).

A number of three-dimensional seismic velocity models have been proposed for southwestern China using various types of geophysical data. These include body wave tomography (Huang *et al.* 2015; Xin *et al.* 2019), surface wave tomography (Yao *et al.* 2010; Shen *et al.* 2016; Qiao *et al.* 2018), and joint inversion of surface wave dispersion and receiver functions (Liu *et al.* 2014; Bao *et al.* 2015). However, these studies did not provide reliable uncertainty estimates for velocity models, which introduces difficulties for interpretation. In this study, we apply the above variational inference surface wave tomography to construct shear wave velocity models for this region, together with accurate uncertainty estimates, so that reliability of the results can be better assessed.

4.1 Data and inversion details

We use 8100 Rayleigh-wave phase-velocity dispersion curves over periods ranging from 5 to 50 seconds extracted from ambient noise interferometry (Fig. 7b). The dataset integrates measurements

from three independent studies (Liu *et al.* 2021): 6910 Rayleigh wave phase-velocity dispersion curves (5–50 s) derived from ambient noise analysis at 124 permanent stations of the China National Seismic Network (Yang *et al.* 2020; Zheng *et al.* 2010); 544 curves (5–45 s) from 35 temporary stations deployed across southwest China by MIT and Lehigh University (Yao *et al.* 2010); and 640 curves (5–40 s) from 16 temporary stations in northern Vietnam and 49 permanent stations in Yunnan Province, installed by the Institute of Earth Sciences, Academia Sinica (Qiao *et al.* 2018; Nguyen *et al.* 2013).

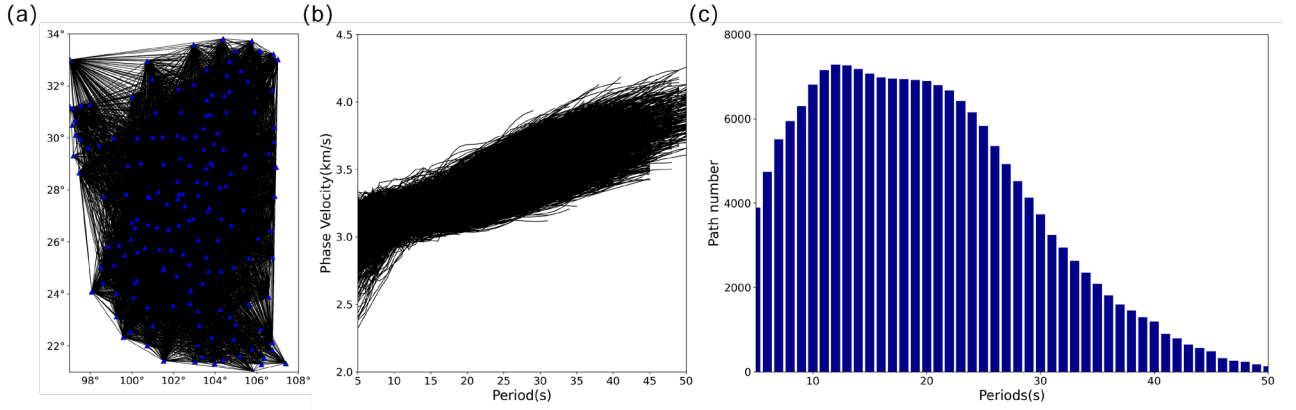


Figure 7. (a) The distribution of ray paths of surface-wave dispersion data. The blue triangles denote 224 stations used to extract interstation ambient noise Rayleigh-wave dispersion data. (b) Rayleigh-wave phase-velocity dispersion curves at periods from 5 s to 50 s. (c) Number of travel times (blue bars) at different periods.

We apply PSVI and sSVG D to infer the 3D shear-wave velocity (V_s) structure of the crust and upper mantle in southwest China, and compare their results. The velocity model is parameterized using a regular grid of $0.5^\circ \times 0.5^\circ$ in the horizontal direction. In the vertical direction, the grid spacings are set to be 2 km from the sea level down to 10 km depth, 5 km from 10 to 50 km depth, 10 km from 50 to 70 km depth, and 30 km from 70 to 100 km depth to take account of decreasing sensitivity of surface waves when depth increases. Above the sea level, the grid spacings are set to be the altitude at each geographic point. We set data noise level to be 2.036 s according to previous studies (Liu *et al.* 2021). The prior distribution at each grid node is defined as an independent uniform distribution centered on a smoothed 3-D V_s model from Yang *et al.* (2020), with intervals increasing linearly from 1.0 km/s to 2.0 km/s as a function of depth.

Similar as above, PSVI is initialized with a standard Gaussian distribution $N(\theta|0, I)$ in the unconstrained parameter space, and a correlation kernel of size $5 \times 5 \times 5$ is used to capture principal

spatial dependencies (the cells in red box, Fig. 1). The distribution is then updated for 2000 iterations using 8 Monte Carlo samples per iteration, from which we draw 1000 samples and transformed them back to the original space to compute summary statistics. For sSVGD, we generated 40 particles from the prior distribution and updated them using the algorithm for 2000 iterations including a burn-in period of 1000 iteration. Similarly as above, we only retain every second sample after the burn-in period to save memory and storage cost. This results in a total of 20,000 samples, which are used to calculate statistics of the estimated posterior pdf.

4.2 Results

Fig. 8 shows horizontal slices of the mean and standard deviation at depths of 5, 10, 20, 30, 40, and 50 km obtained using sSVGD. At 5 km depth, the mean velocity mainly reflects shallow geological features. For example, the Sichuan Basin generally shows lower velocity than the rest of areas in Tibet. Within the Sichuan Basin, there are two low velocity anomalies in the northwest and south, and a relatively higher velocity anomaly in the center which is likely related to the uplift zone in this region due to extrusion of Longquan Shan reverse fault (Zhang *et al.* 2022). Note that the value of the southern low velocity anomaly is higher than that of the northwestern anomaly, which probably indicates thicker sediment accumulation in the northwest. At 10 km depth, there are low velocity anomalies around the XSHF, the LTF and the CHF, which probably reflect activities of those faults. The Sichuan Basin displays similar low velocity anomalies in the northwest and south and high velocity anomalies in the central as at the 5 km depth. In addition, there are high velocity anomalies around several reverse faults (LMSF, LJXJF) and inside the inner ELIP. At 20-30 km depths, there are two low-velocity zones (LVZs): one located west of the LMSF-LJXJF system and the other situated near the Xiaojiang Fault (XJF), which indicate the mechanically weak zones in the middle to lower crust, and has also been reported in a range of previous studies (Yang *et al.* 2020; Liu *et al.* 2023). These LVZs are separated by high velocity anomalies in the region around the LZJF and the ANHF, which corresponds to the inner and intermediate zones of the ELIP. In the Sichuan Basin, in contrast to the heterogeneous low-velocity anomalies observed at 10 km depth, the results exhibit a coherent high-velocity anomaly. At 40–50 km depths, the structure of the mean velocity model primarily reflects variation of Moho depths. In the east the region has transitioned into the upper mantle at this depth (Wang *et al.* 2017), displaying a relatively high velocity anomaly, whereas in the west the results show lower velocity because the

lithology remains dominantly crustal. In addition, note that low velocity anomalies are observed across all depths beneath the Tengchong volcano (Shen *et al.* 2022).

Fig. 9 shows vertical profiles of the mean and standard deviation obtained using sSVGD at longitude of 31°N (AA'), 28°N (BB'), and 26°N (CC') respectively. The results reveal several LVZs between 10 and 30 km depth (regions contoured with 3.4 km/s), which have also been reported in a range of previous studies (Yang *et al.* 2020; Zhang *et al.* 2020; Liu *et al.* 2021, 2023; Qiao *et al.* 2018; Li *et al.* 2024). For example, in the northwest of the region the relatively large scale LVZ (LVZ1 in Fig. 8 and 9) can be clearly observed in profile AA' and BB'. In AA' the LVZ1 exhibits lateral discontinuity between the LTF and the XSHF, which is different from those found in previous studies (Yang *et al.* 2020; Liu *et al.* 2021, 2023) where the low velocity zone extends from the LTF to the XSHF. In profile BB' the LVZ1 is thin in the west and becomes broader between the ZDF and the LXJF, similarly to previous findings (Liu *et al.* 2021, 2023). In profile CC', apart from the LVZ1 in the west, the relatively small-scale LVZ (LVZ2 in Fig. 8) can be observed beneath the XJF. In addition, the model reveals prominent high-velocity anomalies beneath the LZJF zone (profile CC') and the ANHF zone (profile BB'), which indicates the more rigid crust from the inner zone of the ELIP to the region around the ANHF zone, and likely obstructs the southeastward expansion of crustal material from the Tibetan Plateau (Bao *et al.* 2015; Qiao *et al.* 2018). At greater depths (> 40 km) the mean velocity model shows higher velocity, indicative of upper mantle composition, and the transition depth between the low and high velocity matches the Moho depth from Chen *et al.* (2021).

Overall the standard deviation models show increasing uncertainties with depth because of decreasing sensitivity of surface waves when depth increases (Fig. 8 and 9). In addition, there are localized lower uncertainties associated with velocity anomalies. For example, at shallow depths (< 20 km), the results show lower standard deviations at the location of higher velocity anomalies around the LMSF and LXJF (Fig. 8a and 8b), which is probably because rays concentrate in these anomalies. Similar phenomena can also be observed at other depths. Note that in vertical profiles the uncertainty is high at both edges because of lower data coverage as one would expect.

In Fig. 10 and 11, we show the results obtained using PSVI. Overall, the mean velocity model exhibits very similar structures to those obtained using sSVGD, including the low and high velocity anomalies in the Sichuan Basin at shallow depths (< 20 km), the low velocity zones in the mid-to-

lower crust, and the velocity contrast between the western and eastern part at greater depths ($> 20\text{km}$). However, there are still small features that differ in the two results. For example, the shape and scale of the low velocity zones obtained using PSVI are different from those obtained using sSVGD (comparing Fig. 9 and Fig. 11). In addition, the standard deviation model obtained using PSVI show smoother structure and smaller magnitude than those from sSVGD, which probably reflects the limitation of PSVI as the method assumes Gaussian distribution with short-distance correlations. Similar observations have also been reported in earlier work (Ely *et al.* 2018; Zhang & Curtis 2020b; Wang *et al.* 2023; Zhao & Curtis 2024a). By contrast, sSVGD provides a more flexible approximation to the posterior distribution by using a set of samples, and therefore may produce more accurate results.

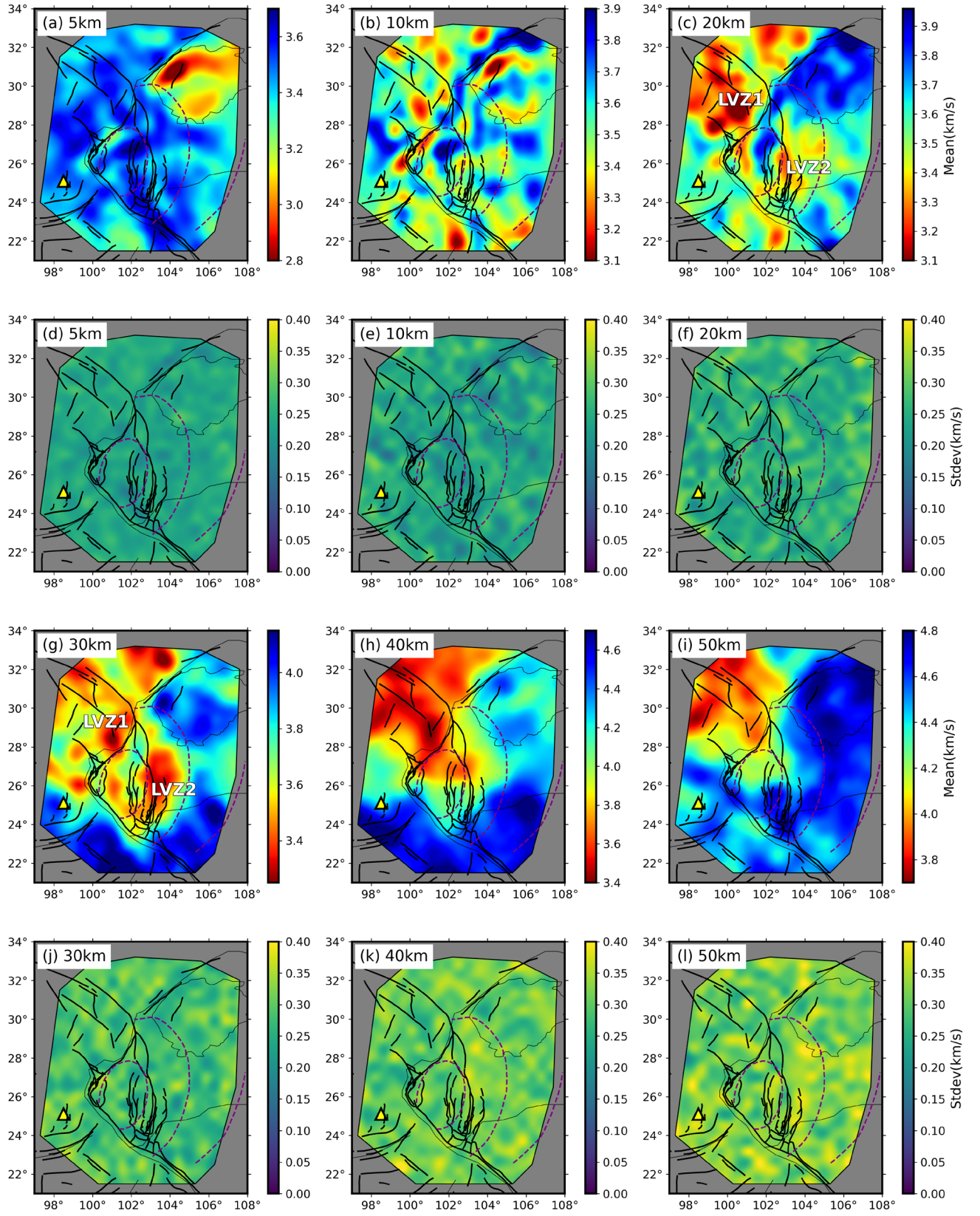


Figure 8. Horizontal slices of the mean and standard deviation of shear velocity obtained using sSVGd at depths of 5, 10, 20, 30, 40 and 50 km. The thin black lines outline the block boundaries, and the thick black lines represent major faults in this region. The purple dashed lines outline the boundaries between the inner, intermediate and outer zones of the ELIP. The yellow triangle denotes the location

of the Tengchong volcano.

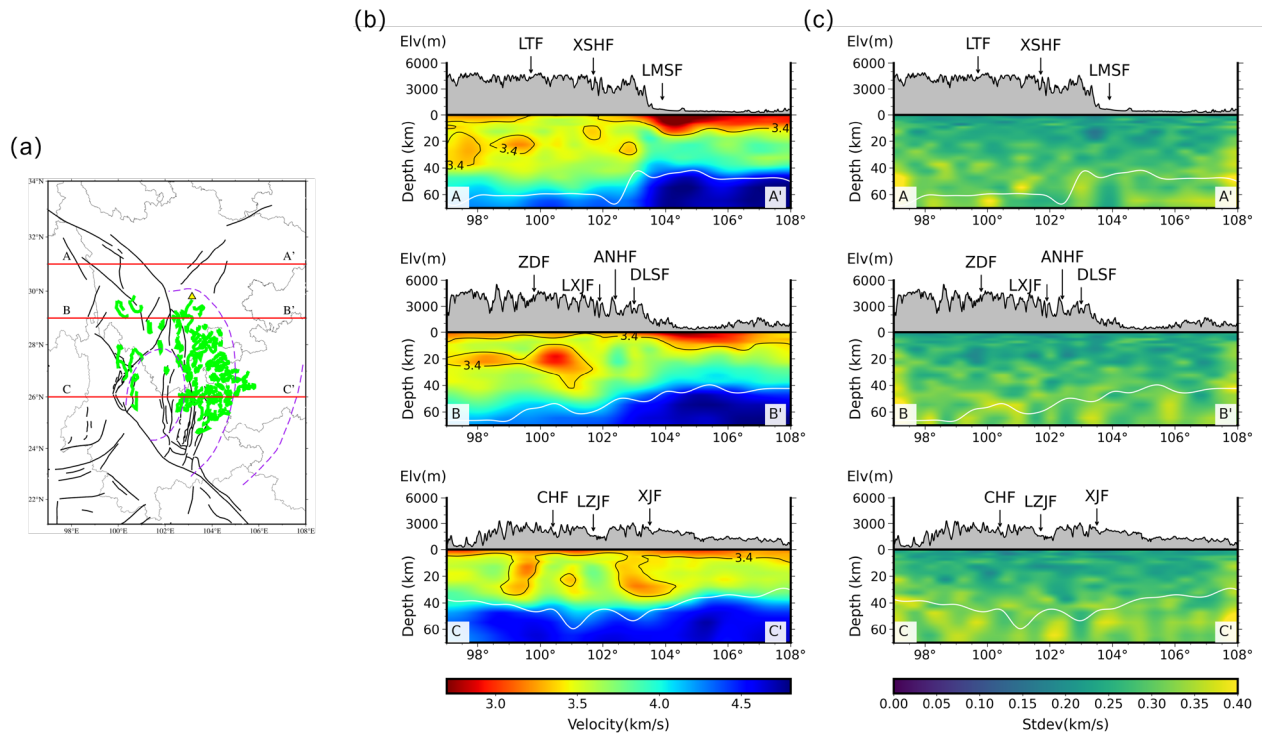


Figure 9. (a) Locations of the vertical profiles AA', BB' and CC' along the longitude direction. Green shaded areas represent the distribution of basalts. Purple dashed lines denote the inner, intermediate and outer boundary of the ELIP. (b) The mean and (c) standard deviation of shear wave velocity along the three vertical profiles obtained using sSVG. White solid lines show the Moho depth from Chen *et al.* (2021).

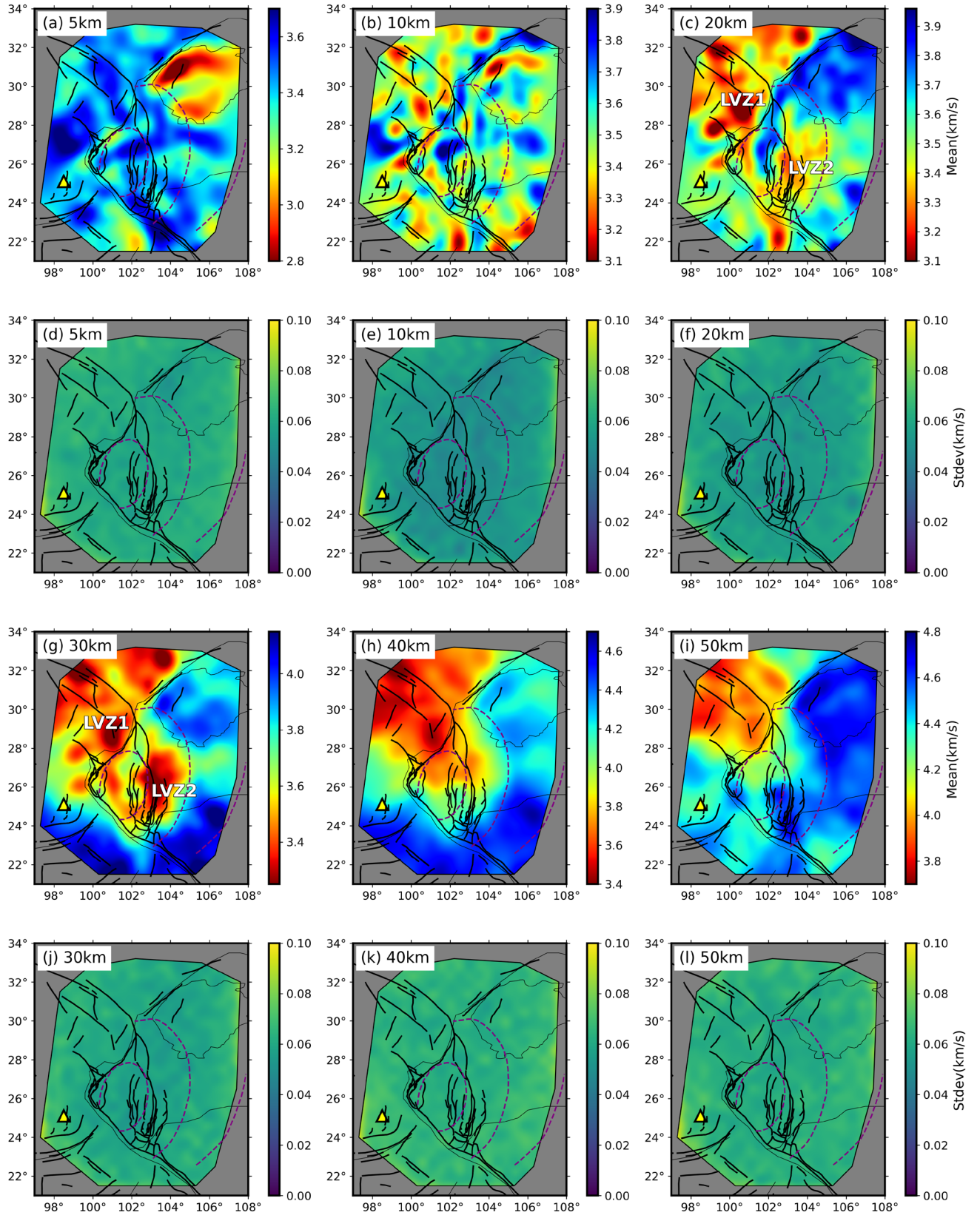


Figure 10. Horizontal slices of the mean and standard deviation obtained using PSVI. Keys as in Fig. 8.

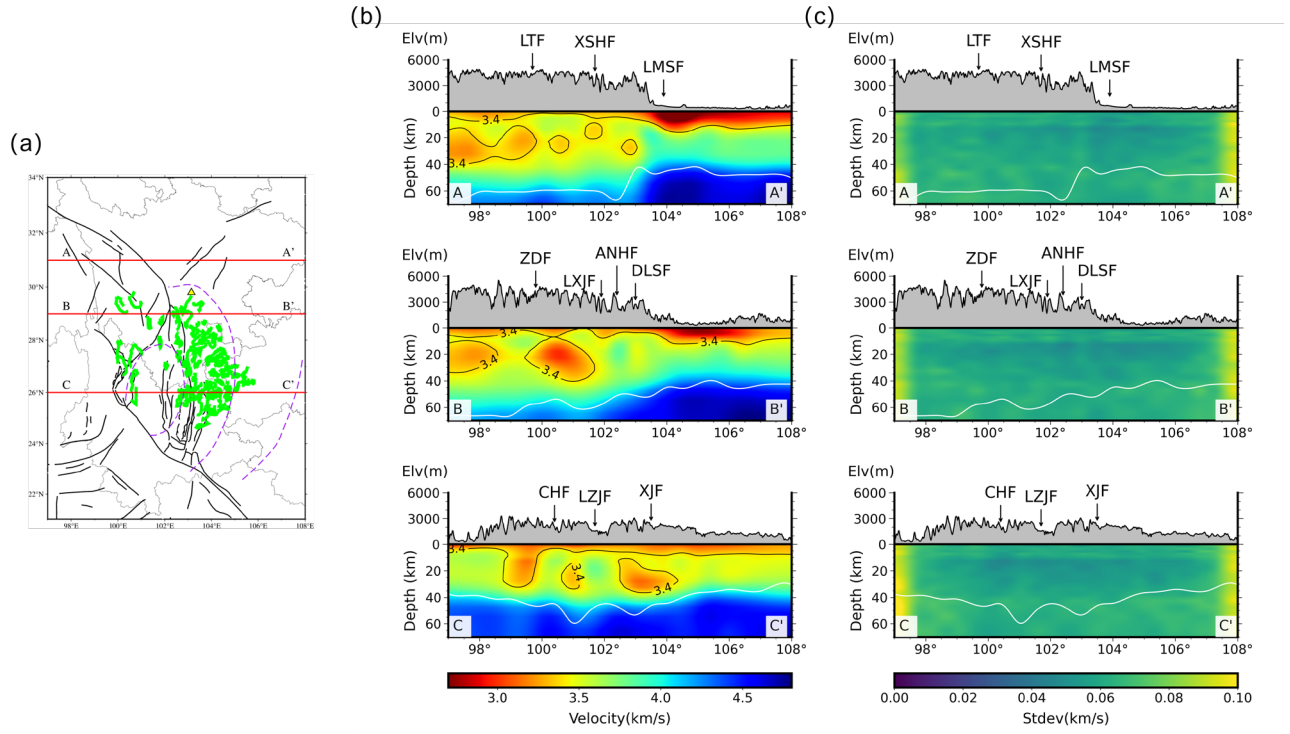


Figure 11. Vertical profiles of the mean and standard deviation obtained using PSVI. Kes as in Fig. 9.

4.3 Model comparison

We compare above models with two other velocity models obtained in previous studies using traditional methods: SWChinaCVM-1.0 (Liu *et al.* 2021) and SWChinaCVM-2.0 (Liu *et al.* 2023) from joint inversion of body and surface wave data (Fig. 12). Overall, those different models show similar features, for example, the low velocity anomalies in the northwest and south of the Sichuan basin at 10 km depth, the low velocity zones in middle-lower crustal, the high velocity anomaly within the inner zone of the ELIP at 20km depth, and the velocity contrast between the eastern and western region at 40 km depth. However, there still exists details that differ in those models. For example, the models obtained using sSVGD and PSVI show many low velocity anomalies around faults in regions outside the Sichuan Basin at 10 km depth, which cannot be clearly observed in those obtained using traditional methods. This discrepancy may arise from strong regularization constraints applied in the traditional inversions, which can suppress small-scale heterogeneities (Zhadow 2002; Yang *et al.* 2025). Within the Sichuan Basin, the models obtained using sSVGD and PSVI show stronger low velocity anomalies in the northwest and south compared to the SWChinaCVM-1.0 model, and are more similar to the SWChinaCVM-2.0 model which is obtained using more data (Liu *et al.* 2023). This demonstrates

that variational methods can probably obtain more accurate results than traditional methods. Similar phenomena can also be observed at 20 and 40 km depth. For example, the models obtained using sSVGD and PSVI show more detailed low velocity structures within the ELIP than those in the SWChinaCVM-1.0 model, which is also more similar to the SWChinaCVM-2.0 model. However, we note that these detailed features do not necessarily reflect real structures of the Earth, because these are the mean structures of a set of possible models, and also because the algorithms may not have converged sufficiently in such a high-dimensional problem. Nevertheless, the similarity between different models obtained using different methods provides confidence for these structures.

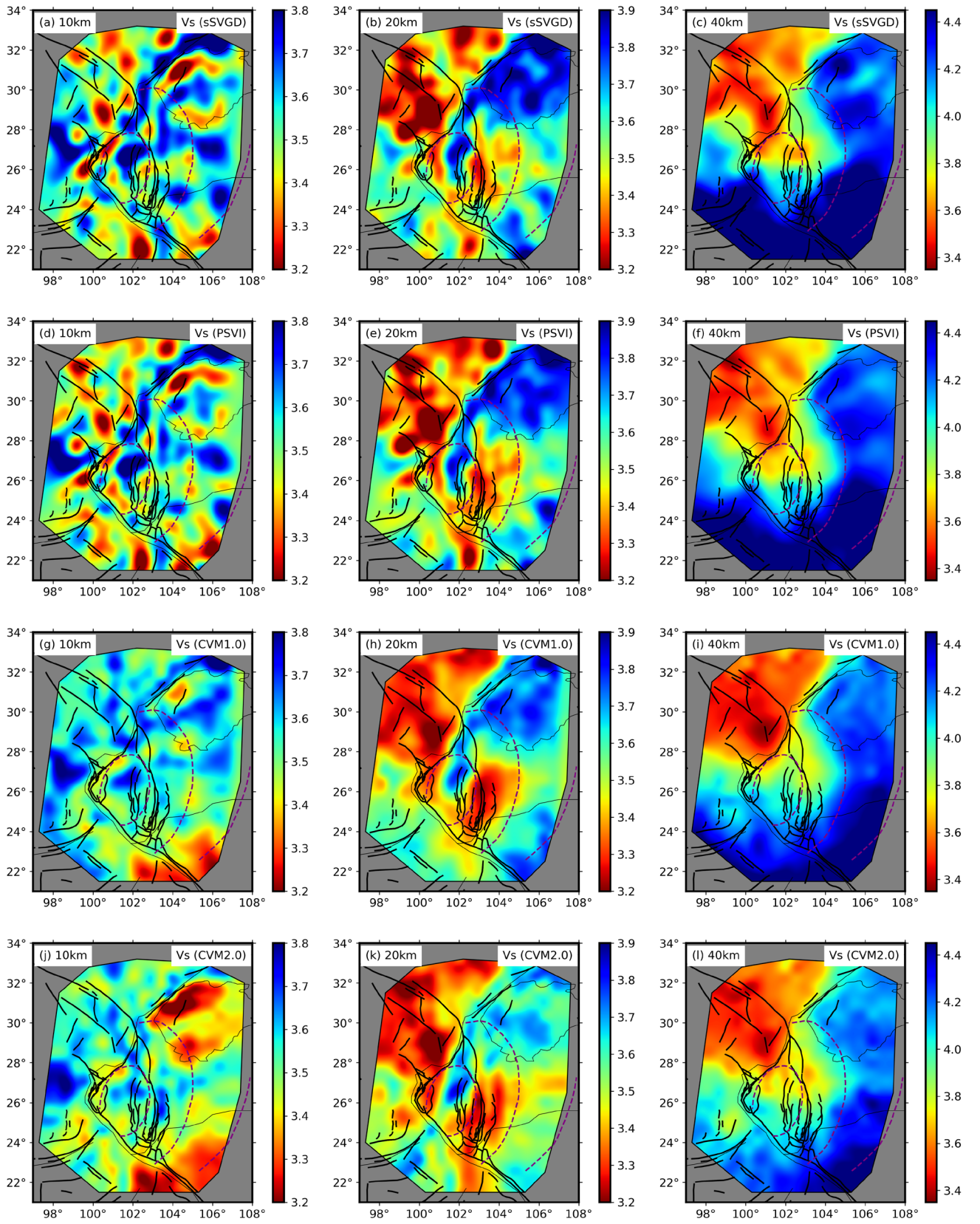


Figure 12. Horizontal slices of different V_s models at 10, 20 and 40 km depth. (a)–(c) V_s model from sSVGd. (d)–(f) V_s model from PSVI. (g)–(i) SWChinaCVM-1.0. (j)–(l) SWChinaCVM-2.0. Keys as in Figure 8.

5. Discussion

We demonstrated that variational inference methods can be effectively applied to 3D surface wave tomography. The mean-field ADVI method underestimates posterior uncertainties because the method neglects correlation information between parameters. PSVI improves the results accuracy by exploiting a sparse-structured covariance matrix which only includes correlation between parameters at proximal cells, while maintaining a computational cost comparable to mean-field ADVI. By contrast, sSVGD requires more computational cost than mean-field ADVI or PSVI because the method uses a set of particles to approximate the posterior distribution. However, sSVGD can generate more accurate results than mean-field ADVI or PSVI as the method can theoretically approximate the posterior distribution to arbitrary accuracy. To further improve efficiency of sSVGD, high order gradient information may be used, for example, using a Hessian matrix kernel (Wang *et al.* 2019) or the stochastic Stein variational Newton method (Leviyev *et al.* 2022). In addition, one may try to reduce dimensionality of the problem to reduce the number of required samples. For example, one may use other parameterizations that require fewer parameters to represent the velocity model, such as Voronoi cells (Bodin & Sambridge 2009; Zhang *et al.* 2018), wavelet parameterization (Hawkins & Sambridge 2015), discrete cosine transforms (Urozayev *et al.* 2022), or neural network reparameterization (Laloy *et al.* 2017; Mosser *et al.* 2020; Agata *et al.* 2023).

For the likelihood function we used a Gaussian distribution with a fixed data noise level in the variational methods. In practice this noise level should be determined from the data. For example, one can use the standard deviation of daily measurements as the noise level, or estimate it using the maximum likelihood method (Sambridge 2014). It may also be possible to estimate the noise level in the inversion process using a hierarchical Bayesian formulation (Malinverno & Briggs 2004; Ranganath *et al.* 2016). In addition, other non-Gaussian likelihood functions may be used to improve the results given that those likelihood functions represent the distribution of data uncertainty (Zhang *et al.* 2023).

In this study we used a smoothness constrained prior distribution, and assessed the influence of such constraints on the posterior results. However, in many practical applications it is challenging to define an appropriate prior probability distribution over smoothness levels. To reduce this issue, hierarchical Bayesian methods may be used to estimate this distribution a posteriori, but such

approaches can introduce physical inconsistencies into Bayesian inversion outcomes (Mosegaard & Curtis 2024). In practice where more knowledge about the subsurface is available, one can use a more informative prior distribution, for example, by incorporating information such as geological classifications, well logs, subsurface structures and expert knowledge into the prior distribution. In addition, other forms of prior distribution may be used, for example, Gaussian process (Ambikasaran *et al.* 2015; Liu *et al.* 2020), or neural network-based prior distribution that encode geological features into the prior distribution (Laloy *et al.* 2017; Mosser *et al.* 2020).

For computational efficiency we used a forward modelling method based on modal approximation and 2-D fast marching method. To further improve accuracy, full waveform modelling methods may be used to simulate surface wave data (Tromp *et al.* 2005; Fichtner *et al.* 2006). However, we note that this requires expensive computational cost (Zhang *et al.* 2023). In this study we only used surface wave dispersion data to constrain the subsurface. In future, other types of data, for example, body wave travel time data and receiver functions, can also be included to improve the results (Julia *et al.* 2000; Fang *et al.* 2016; Zhang *et al.* 2020).

6. Conclusions

In this study we applied a set of variational inference methods (mean-field ADVI, PSVI and sSVGD) to solve 3-D surface wave tomographic problems using both synthetic and real data. The results show that mean-field ADVI can provide rapid solutions but with systematically underestimated uncertainty. PSVI improves the results accuracy by incorporating structured correlations among parameters, while maintaining computational efficiency of mean-field ADVI. sSVGD appears to be the most expensive method, but provides the most accurate results with reasonable uncertainty estimates. In real data applications from Southwest China, the variational methods provide more detailed structures than those obtained using traditional methods, along with reliable uncertainty estimates. We therefore conclude that variational inference methods can be applied fruitfully to address real-world 3D surface wave tomographic problems.

Appendix A: Construction of the correlation matrix $\mathbf{L}$ in PSVI

In this section, we present a method to construct the three-dimensional correlation matrix $\mathbf{L}$ given a specific correlation length.

Assume a correlation length of two cells in the x -direction we can construct a lower-triangular sparse matrix with two sub-diagonals:

$$D_x = \begin{bmatrix} l_{0,1} & & & & & & & & & \\ l_{1,1} & l_{0,2} & & & & & & & & \\ l_{2,1} & l_{1,2} & l_{0,3} & & & & & & & \\ & l_{2,2} & l_{1,3} & \ddots & & & & & & \\ & & l_{2,3} & \ddots & \ddots & & & & & \\ & & & \ddots & \ddots & l_{0,nx-2} & & & & \\ & & & & \ddots & l_{1,nx-2} & l_{0,nx-1} & & & \\ & & & & & l_{2,nx-2} & l_{1,nx-1} & l_{0,nx} & & \end{bmatrix}$$

Similarly, we can obtain the D matrices in the y direction (D_y) and the z direction (D_z). The number of off-diagonal blocks in the D matrix can be controlled by the number of adjacent grid points considered in each spatial direction.

Given the above definition, the correlation matrix for parameters on a regular 3-D grid can be obtained by multiplying the D_x , D_y and D_z in sequence using the Kronecker product:

$$\mathbf{L} = D_x \otimes (D_y \otimes D_z)$$

where $\otimes$ represents the Kronecker product which is defined as:

$$D_y \otimes D_z = \begin{bmatrix} d_{1,1}D_z & \cdots & d_{1,ny}D_z \\ \vdots & \ddots & \vdots \\ d_{ny,1}D_z & \cdots & d_{ny,ny}D_z \end{bmatrix}$$

where $(d_{1,1}, \dots, d_{ny,ny})$ represents elements of D_y , and ny is the number of grid points in y direction. Note that the order of the matrix multiplication must correspond to the order of parameter expansion as the Kronecker product is non-commutative.

Appendix B: Synthetic tests for different prior hypotheses

In this section, we compare a smoothed prior distribution ($\sigma_x = \sigma_y = 0.2, \sigma_z = 1.0$) with a non-smoothed uniform prior distribution to demonstrate the effect of smoothness constraints. The inversions are performed in exactly the same way for different prior distributions.

Fig. B1 presents the results for a horizontal section at a depth of 20 km using both prior pdfs (smoothed and non-smoothed). The first two rows present the mean and standard deviation obtained without smoothness, while the last two rows show those with smoothness applied. From left to right, each column corresponds to the results obtained using mean-field ADVI, PSVI, and sSVGD, respectively.

Fig. B2 compares the posterior marginal pdfs at two vertical profiles (locations are indicated with white pluses in Fig. B1) obtained using non-smoothed and smoothed prior distributions with different methods. The results show broader distributions when using smoothed prior pdfs as one would expect. Note that in the case with non-smoothed prior pdfs, the results obtained using sSVGD show more complex distributions than those obtained using mean-field ADVI and PSVI, which may reflect limitation of mean-field ADVI and PSVI as they assume Gaussian assumptions.

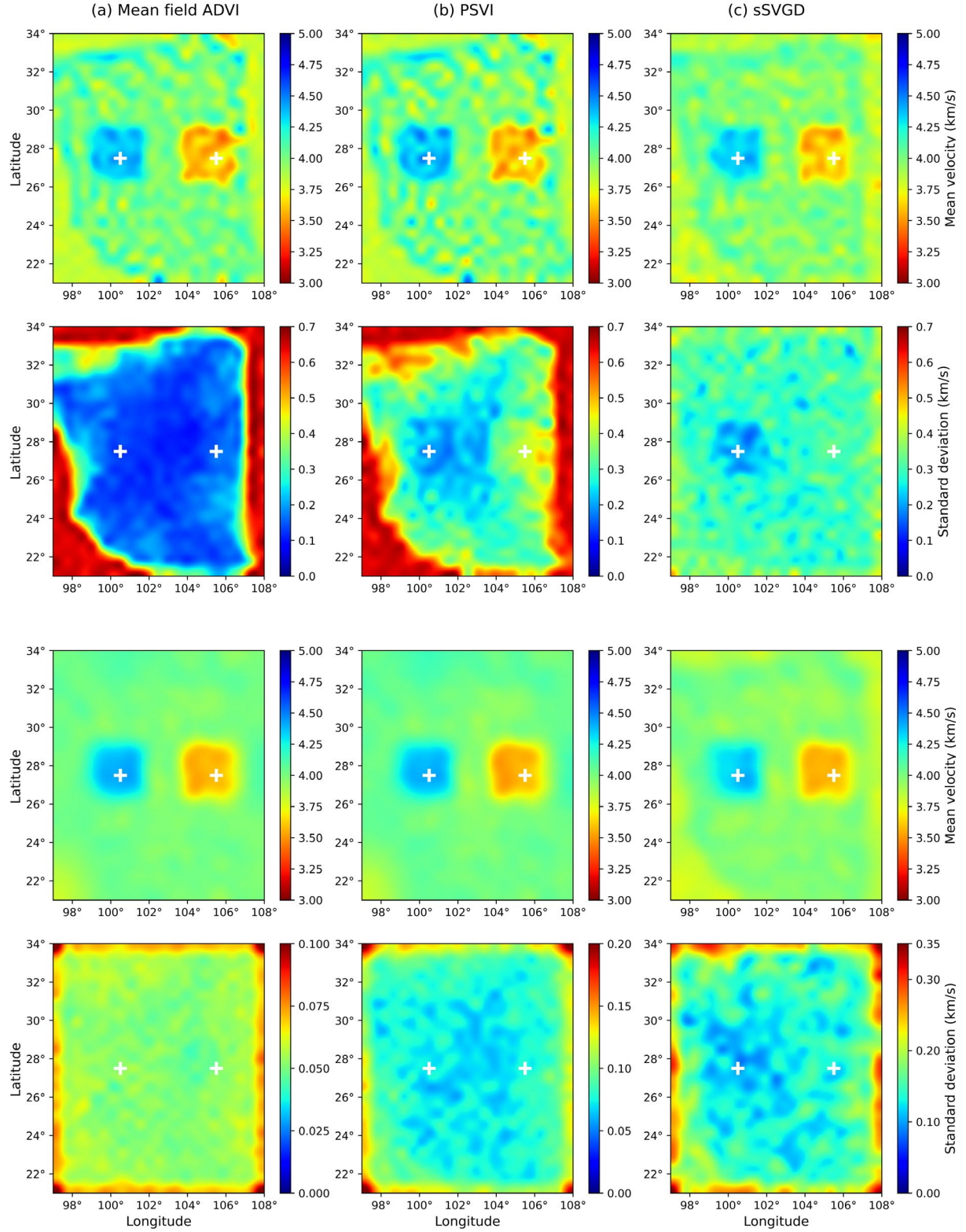


Figure B1. Inversion results on a horizontal section at 20 km depth obtained using (a) Mean Field ADVI, (b) PSVI, and (c) sSVGd. The top panels show the results without smoothness, and the lower panels show those with smoothness applied ($\sigma_x = \sigma_y = 0.2, \sigma_z = 1.0$).

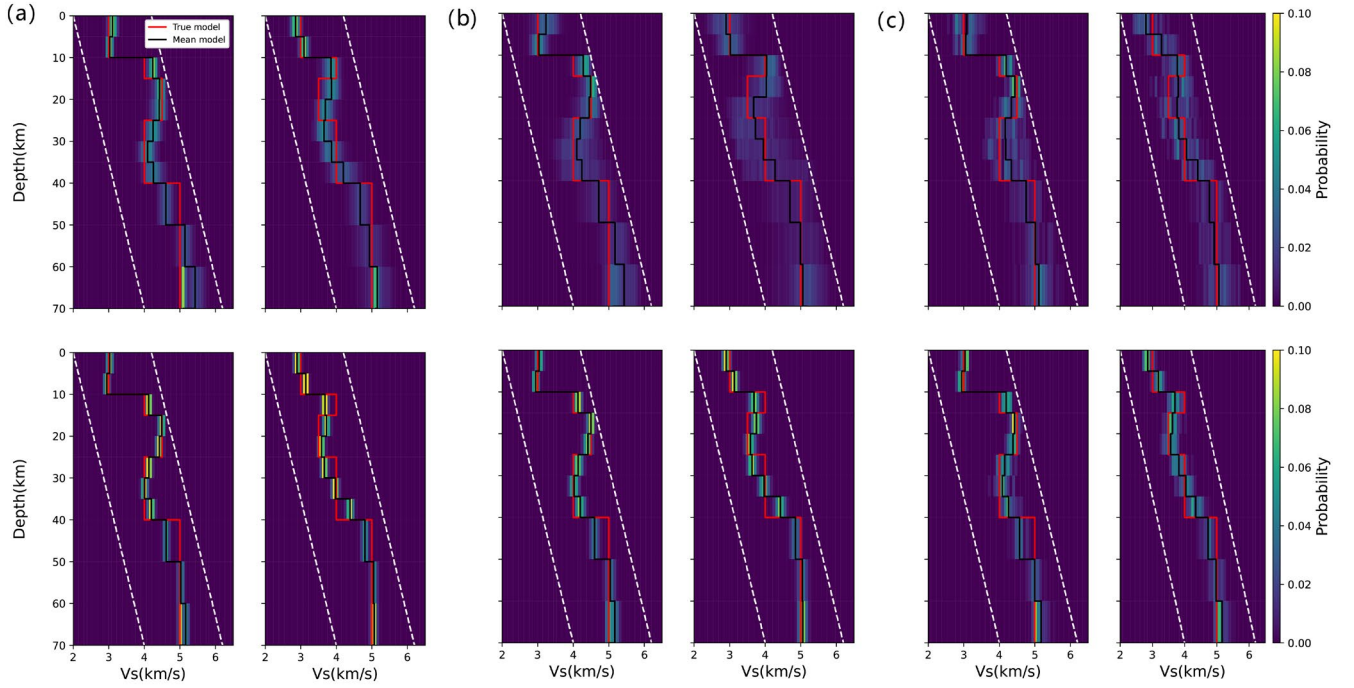


Figure B2. Posterior marginal probability density functions (pdfs) at two vertical locations, as indicated by the white pluses in Figure B1. (a) Results from mean-field ADVI. The top panels show the results obtained using the non-smoothed uniform prior pdf, and the low panels show those obtained using the smoothed prior pdf with parameters $\sigma_x = \sigma_y = 0.2$, and $\sigma_z = 1.0$. (b) and (c) present the posterior marginal pdfs obtained using PSVI and sSVGd, respectively.

Acknowledgments

This study is supported by the National Natural Science Foundation of China (Grant Nos. 42204055 and U23B20160), and the Fundamental Research Funds for the Central Universities of China University of Geosciences in Beijing. This work used the HPC service of China University of Geosciences Beijing (<https://hpc.cugb.edu.cn>).

References

- Agata, R., Shiraishi, K., & Fujie, G. (2023). Bayesian seismic tomography based on velocity-space Stein variational gradient descent for physics-informed neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1-17.
- Aki, K., & Lee, W. H. K. (1976). Determination of three - dimensional velocity anomalies under a seismic array using first P arrival times from local earthquakes: 1. A homogeneous initial model. *Journal of Geophysical research*, 81(23), 4381-4399.
- Aki, K., & Richards, P. G. (2002). *Quantitative seismology*.
- Aleardi, M., Salusti, A., & Pierini, S. (2020). Transdimensional and Hamiltonian Monte Carlo inversions of Rayleigh-wave dispersion curves: a comparison on synthetic datasets. *Near Surface Geophysics*, 18(5), 515-543.

- Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D. W., & O'Neil, M. (2015). Fast direct methods for Gaussian processes. *IEEE transactions on pattern analysis and machine intelligence*, 38(2), 252-265.
- Andrieu, C., & Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and computing*, 18, 343-373.
- 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*.
- Bao, X., Sun, X., Xu, M., Eaton, D. W., Song, X., Wang, L., ... & Wang, P. (2015). Two crustal low-velocity channels beneath SE Tibet revealed by joint inversion of Rayleigh wave dispersion and receiver functions. *Earth and Planetary Science Letters*, 415, 16-24.
- Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4, p. 738). New York: springer.
- Blei, D. M., & Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures.
- 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.
- Bodin, T., & Sambridge, M. (2009). Seismic tomography with the reversible jump algorithm. *Geophysical Journal International*, 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. *Journal of geophysical research: solid earth*, 117(B2).
- Burdick, S., & Lekić, V. (2017). Velocity variations and uncertainty from transdimensional P-wave tomography of North America. *Geophysical Journal International*, 209(2), 1337-1351.
- Chen, L., Wang, W., & Zhang, L. (2021). Crustal thickness in southeast Tibet based on the SWChinaCVM-1.0 model. *Earthquake Science*, 34(3), 246-260.
- Chen, T., Fox, E., & Guestrin, C. (2014, June). Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning* (pp. 1683-1691). PMLR.
- Chib, S., & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4), 327-335.
- Curtis, A., Trampert, J., Snieder, R., & Dost, B. (1998). Eurasian fundamental mode surface wave phase velocities and their relationship with tectonic structures. *Journal of Geophysical Research: Solid Earth*, 103(B11), 26919-26947.
- Dębski, W. (2004). Application of Monte Carlo Techniques for Solving Selected Seismological Inverse Problems: Monographic Volume. Instytut Geofizyki Polskiej Akademii Nauk.
- Durmus, A., Majewski, S., & Miasojedow, B. (2019). Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20(73), 1-46.
- Dziewonski, A. M., & Woodhouse, J. H. (1987). Global images of the Earth's interior. *Science*, 236(4797), 37-48.
- 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.
- Fang, H., Yao, H., Zhang, H., Huang, Y. C., & van der Hilst, R. D. (2015). Direct inversion of surface wave dispersion for three-dimensional shallow crustal structure based on ray tracing: methodology and application. *Geophysical Journal International*, 201(3), 1251-1263.
- Fang, H., Zhang, H., Yao, H., Allam, A., Zigone, D., Ben-Zion, Y., ... & van der Hilst, R. D. (2016). A new algorithm for three-dimensional joint inversion of body wave and surface wave data and its application to the Southern California plate boundary region. *Journal of Geophysical Research: Solid Earth*, 121(5), 3557-3569.
- Fichtner, A., Bunge, H. P., & Igel, H. (2006). The adjoint method in seismology: I. Theory. *Physics of the Earth and Planetary Interiors*, 157(1-2), 86-104.

- Fichtner, A., Zunino, A., & Gebraad, L. (2019). Hamiltonian Monte Carlo solution of tomographic inverse problems. *Geophysical Journal International*, 216(2), 1344-1363.
- Fu, X., & Innanen, K. A. (2021, October). An MCMC-based approach to time-lapse full-waveform inversion. In SEG International Exposition and Annual Meeting (p. D011S147R001). SEG.
- 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. *Geophysical Journal International*, 208(1), 36-58.
- Galetti, E., Curtis, A., Meles, G. A., & Baptie, B. (2015). Uncertainty loops in travel-time tomography from nonlinear wave physics. *Physical review letters*, 114(14), 148501.
- Gallego, V., & Insua, D. R. (2018). Stochastic gradient MCMC with repulsive forces. *arXiv preprint arXiv:1812.00071*.
- Gao, C., & Lekić, V. (2018). Consequences of parametrization choices in surface wave inversion: Insights from transdimensional Bayesian methods. *Geophysical Journal International*, 215(2), 1037-1063.
- Gao, H., & Shen, Y. (2014). Upper mantle structure of the Cascades from full-wave ambient noise tomography: Evidence for 3D mantle upwelling in the back-arc. *Earth and Planetary Science Letters*, 390, 222-233.
- 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.
- Geyer, C. J. (2011). Introduction to markov chain monte carlo. *Handbook of markov chain monte carlo*, 20116022(45), 22.
- Golub, G. H., Hansen, P. C., & O'Leary, D. P. (1999). Tikhonov regularization and total least squares. *SIAM journal on matrix analysis and applications*, 21(1), 185-194.
- Gong, C., Peng, J., & Liu, Q. (2019, May). Quantile stein variational gradient descent for batch Bayesian optimization. In *International Conference on machine learning* (pp. 2347-2356). PMLR.
- Gosselin, J. M., Dosso, S. E., Cassidy, J. F., Quijano, J. E., Molnar, S., & Dettmer, J. (2017). A gradient-based model parametrization using Bernstein polynomials in Bayesian inversion of surface wave dispersion. *Geophysical Journal International*, 211(1), 528-540.
- Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4), 711-732.
- Hawkins, R., & Sambridge, M. (2015). Geophysical imaging using trans-dimensional trees. *Geophysical Journal International*, 203(2), 972-1000.
- Herrmann, R. B. (2013). Computer programs in seismology: An evolving tool for instruction and research. *Seismological Research Letters*, 84(6), 1081-1088.
- Huang, X., Eikrem, K. S., Jakobsen, M., & Nævdal, G. (2020). Bayesian full-waveform inversion in anisotropic elastic media using the iterated extended Kalman filter. *Geophysics*, 85(4), C125-C139.
- Huang, Z., Wang, P., Xu, M., Wang, L., Ding, Z., Wu, Y., ... & Li, H. (2015). Mantle structure and dynamics beneath SE Tibet revealed by new seismic images. *Earth and Planetary Science Letters*, 411, 100-111.
- Iyer, H. M., & Hirahara, K. (Eds.). (1993). *Seismic tomography: Theory and practice*. Springer Science & Business Media.
- Izzatullah, M., Van Leeuwen, T., & Peter, D. (2021). Bayesian seismic inversion: a fast sampling Langevin dynamics Markov chain Monte Carlo method. *Geophysical Journal International*, 227(3), 1523-1553.
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37, 183-233.
- Julia, J., Ammon, C. J., Herrmann, R. B., & Correig, A. M. (2000). Joint inversion of receiver function and surface wave dispersion observations. *Geophysical Journal International*, 143(1), 99-112.

- Kan, R. J., Zhang, S. L., & Yan, F. T. (1977). Study on the current tectonic stress field and the characteristics of current tectonic activity in southwest China. *Chin. J. Geophys.*, 20(2), 96-107.
- Kästle, E. D., Tilmann, F., & AlpArray and Swath-D Working Groups. (2024). Anisotropic Reversible-Jump McMC Shear-Velocity Tomography of the Eastern Alpine Crust. *Geochemistry, Geophysics, Geosystems*, 25(3), e2023GC011238.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kotsi, M., Malcolm, A., & Ely, G. (2020). Uncertainty quantification in time-lapse seismic imaging: A full-waveform approach. *Geophysical Journal International*, 222(2), 1245-1263.
- Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of machine learning research*, 18(14), 1-45.
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1), 79-86.
- Kurita, T. (1973). Regional variations in the structure of the crust in the central United States from P-wave spectra. *Bulletin of the Seismological Society of America*, 63(5), 1663-1687.
- 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. *Advances in water resources*, 110, 387-405.
- Leviyev, A., Chen, J., Wang, Y., Ghattas, O., & Zimmerman, A. (2022). A stochastic stein variational newton method. *arXiv preprint arXiv:2204.09039*.
- Li, C., Yu, C., Hillenbrand, I., & Zhang, W. (2024). Three-dimensional crustal channel flows beneath the southeastern Tibetan Plateau revealed by full-waveform ambient noise tomography. *Geophysical Research Letters*, 51(22), e2024GL110704.
- Liu, H., Ong, Y. S., Shen, X., & Cai, J. (2020). When Gaussian process meets big data: A review of scalable GPs. *IEEE transactions on neural networks and learning systems*, 31(11), 4405-4423.
- Liu, Q. Y., Van Der Hilst, R. D., Li, Y., Yao, H. J., Chen, J. H., Guo, B., ... & Li, S. C. (2014). Eastward expansion of the Tibetan Plateau by crustal flow and strain partitioning across faults. *Nature Geoscience*, 7(5), 361-365.
- Liu, Q., & Wang, D. (2016). Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29.
- Liu, Y., Yao, H., Zhang, H., & Fang, H. (2021). The community velocity model V. 1.0 of southwest China, constructed from joint body-and surface-wave travel-time tomography. *Seismological Society of America*, 92(5), 2972-2987.
- Liu, Y., Yu, Z., Zhang, Z., Yao, H., Wang, W., Zhang, H., ... & Fang, L. (2023). The high-resolution community velocity model V2. 0 of southwest China, constructed by joint body and surface wave tomography of data recorded at temporary dense arrays. *Science China Earth Sciences*, 66(10), 2368-2385.
- Ma, Y. A., Chen, T., & Fox, E. (2015). A complete recipe for stochastic gradient MCMC. *Advances in neural information processing systems*, 28.
- Malinverno, A., & Briggs, V. A. (2004). Expanded uncertainty quantification in inverse problems: Hierarchical Bayes and empirical Bayes. *Geophysics*, 69(4), 1005-1016.
- 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 Journal on Scientific Computing*, 34(3), A1460-A1487.
- Meier, U., Curtis, A., & Trampert, J. (2007a). Global crustal thickness from neural network inversion of surface wave data. *Geophysical Journal International*, 169(2), 706-722.
- Meier, U., Curtis, A., & Trampert, J. (2007b). A global crustal model constrained by nonlinearised inversion of

fundamental mode surface waves. *Geophysical Research Letters*, 34, L16304.

Molnar, S., Dosso, S. E., & Cassidy, J. F. (2010). Bayesian inversion of microtremor array dispersion data in southwestern British Columbia. *Geophysical Journal International*, 183(2), 923-940.

Mosegaard, K., & Curtis, A. (2024). Inconsistency and Acausality in Bayesian Inference for Physical Problems. *arXiv preprint arXiv:2411.13570*.

Mosegaard, K., & Tarantola, A. (1995). Monte Carlo sampling of solutions to inverse problems. *Journal of Geophysical Research: Solid Earth*, 100(B7), 12431-12447.

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

Nakanishi, I., & Anderson, D. L. (1983). Measurement of mantle wave velocities and inversion for lateral heterogeneity and anisotropy: 1. Analysis of great circle phase velocities. *Journal of Geophysical Research: Solid Earth*, 88(B12), 10267-10283.

Nawaz, M. A., & Curtis, A. (2018). Variational Bayesian inversion (VBI) of quasi-localized seismic attributes for the spatial distribution of geological facies. *Geophysical Journal International*, 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. *Journal of Geophysical Research: Solid Earth*, 124(6), 5867-5887.

Nawaz, M. A., Curtis, A., Shahrane, M. S., & Gere, C. (2020). Variational Bayesian inversion of seismic attributes jointly for geologic facies and petrophysical rock properties. *Geophysics*, 85(4), MR213-MR233.

Nguyen, V. D., Huang, B. S., Le, T. S., Dinh, V. T., Zhu, L., & Wen, K. L. (2013). Constraints on the crustal structure of northern Vietnam based on analysis of teleseismic converted waves. *Tectonophysics*, 601, 87-97.

Oksendal, B. (2013). *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media.

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

Qiao, L., Yao, H., Lai, Y. C., Huang, B. S., & Zhang, P. (2018). Crustal structure of southwest China and northern Vietnam from ambient noise tomography: Implication for the large-scale material transport model in SE Tibet. *Tectonics*, 37(5), 1492-1506.

Rahimi Dalkhani, A., Ágústsðóttir, T., Gudnason, E. A., Hersir, G. P., Zhang, X., & Weemstra, C. (2024). Transdimensional ambient-noise surface wave tomography of the Reykjanes Peninsula, SW Iceland. *Geophysical Journal International*, 236(1), 621-643.

Ramgraber, M., Weatherl, R., Blumensaat, F., & Schirmer, M. (2021). Non-Gaussian parameter inference for hydrogeological models using Stein variational gradient descent. *Water Resources Research*, 57(4), e2020WR029339.

Ranganath, R., Tran, D., & Blei, D. (2016, June). Hierarchical variational models. In *International conference on machine learning* (pp. 324-333). PMLR.

Rawlinson, N., & Sambridge, M. (2004). Multiple reflection and transmission phases in complex layered media using a multistage fast marching method. *Geophysics*, 69(5), 1338-1350.

Ritzwoller, M. H., Shapiro, N. M., Barmin, M. P., & Levshin, A. L. (2002). Global surface wave diffraction tomography. *Journal of Geophysical Research: Solid Earth*, 107(B12), ESE-4.

Sambridge, M. (2014). A parallel tempering algorithm for probabilistic sampling and multimodal optimization. *Geophysical Journal International*, 196(1), 357-374.

Sambridge, M., & Mosegaard, K. (2002). Monte Carlo methods in geophysical inverse problems. *Reviews of Geophysics*, 40(3), 3-1.

Shapiro, N. M., & Ritzwoller, M. H. (2002). Monte-Carlo inversion for a global shear-velocity model of the crust and upper mantle. *Geophysical Journal International*, 151(1), 88-105.

- Shapiro, N. M., Campillo, M., Stehly, L., & Ritzwoller, M. H. (2005). High-resolution surface-wave tomography from ambient seismic noise. *Science*, 307(5715), 1615-1618.
- Shen, W., Liu, S., Yang, D., Wang, W., Xu, X., & Yang, S. (2022). The crustal and uppermost mantle dynamics of the Tengchong–Baoshan region revealed by P-wave velocity and azimuthal anisotropic tomography. *Geophysical Journal International*, 230(2), 1092-1105.
- Shen, W., Ritzwoller, M. H., Kang, D., Kim, Y., Lin, F. C., Ning, J., ... & Zhou, L. (2016). A seismic reference model for the crust and uppermost mantle beneath China from surface wave dispersion. *Geophysical Journal International*, 206(2), 954-979.
- Simons, F. J., Van Der Hilst, R. D., Montagner, J. P., & Zielhuis, A. (2002). Multimode Rayleigh wave inversion for heterogeneity and azimuthal anisotropy of the Australian upper mantle. *Geophysical Journal International*, 151(3), 738-754.
- 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.
- Tarantola, A. (2005). Inverse problem theory and methods for model parameter estimation. Society for industrial and applied mathematics.
- Team, S. D. (2016). Stan modeling language users guide and reference manual. *Team SD*.
- Tilmann, F. J., Sadeghisorkhani, H., & Mauerberger, A. (2020). Another look at the treatment of data uncertainty in Markov chain Monte Carlo inversion and other probabilistic methods. *Geophysical Journal International*, 222(1), 388-405.
- Trampert, J., & Woodhouse, J. H. (1995). Global phase velocity maps of Love and Rayleigh waves between 40 and 150 seconds. *Geophysical Journal International*, 122(2), 675-690.
- Tromp, J., Tape, C., & Liu, Q. (2005). Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels. *Geophysical Journal International*, 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. *Geophysical Journal International*, 229(2), 838-852.
- Wang, D., Tang, Z., Bajaj, C., & Liu, Q. (2019). Stein variational gradient descent with matrix-valued kernels. *Advances in neural information processing systems*, 32.
- Wang, E. (Ed.). (1998). *Late cenozoic xianshuihe-xiaojiang, Red River, and dali fault systems of southwestern sichuan and central yunnan, China* (Vol. 327). Geological Society of America.
- Wang, W., McMechan, G. A., & Ma, J. (2023). Reweighted variational full-waveform inversions. *Geophysics*, 88(4), R499-R512.
- Wang, W., Wu, J., Fang, L., Lai, G., & Cai, Y. (2017). Crustal thickness and Poisson's ratio in southwest China based on data from dense seismic arrays. *Journal of Geophysical Research: Solid Earth*, 122(9), 7219-7235.
- 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).
- Xin, H., Zhang, H., Kang, M., He, R., Gao, L., & Gao, J. (2019). High-resolution lithospheric velocity structure of continental China by double-difference seismic travel-time tomography. *Seismological Research Letters*, 90(1), 229-241.
- Xu, Y. G., He, B., Chung, S. L., Menzies, M. A., & Frey, F. A. (2004). Geologic, geochemical, and geophysical consequences of plume involvement in the Emeishan flood-basalt province. *Geology*, 32(10), 917-920.
- Yang, H., Zhang, X., & Zhang, H. (2025). 3D Variational Inference-Based Double-Difference Seismic Tomography Method and Application to the SAFOD Site, California. *Geophysical Journal International*, ggaf040.
- Yang, W., & Zhang, X. (2025). 3D Bayesian Variational Surface Wave Tomography and Application to the Southwest China [Data set]. Zenodo.

- Yang, Y., Yao, H., Wu, H., Zhang, P., & Wang, M. (2020). A new crustal shear-velocity model in Southwest China from joint seismological inversion and its implications for regional crustal dynamics. *Geophysical Journal International*, 220(2), 1379-1393.
- Yao, H., van Der Hilst, R. D., & Montagner, J. P. (2010). Heterogeneity and anisotropy of the lithosphere of SE Tibet from surface wave array tomography. *Journal of Geophysical Research: Solid Earth*, 115(B12).
- 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.
- Zhang, C., & Chen, T. (2022). Bayesian slip inversion with automatic differentiation variational inference. *Geophysical Journal International*, 229(1), 546-565.
- Zhang, J. J., Li, H. B., Zhang, G. Z., Gu, Y. P., & Liu, Z. F. (2021). Rock physics inversion based on an optimized MCMC method. *Applied Geophysics*, 18(3), 288-298.
- Zhang, X., & Curtis, A. (2020a). Seismic tomography using variational inference methods. *Journal of Geophysical Research: Solid Earth*, 125(4), e2019JB018589.
- Zhang, X., & Curtis, A. (2020b). Variational full-waveform inversion. *Geophysical Journal International*, 222(1), 406-411.
- Zhang, X., & Curtis, A. (2021). Bayesian full-waveform inversion with realistic priors. *Geophysics*, 86(5), A45-A49.
- Zhang, X., & Curtis, A. (2023). VIP--Variational Inversion Package with example implementations of Bayesian tomographic imaging. arXiv preprint arXiv:2310.13325.
- Zhang, X., Curtis, A., Galetti, E., & De Ridder, S. (2018). 3-D Monte Carlo surface wave tomography. *Geophysical Journal International*, 215(3), 1644-1658.
- Zhang, X., Hansteen, F., & Curtis, A. (2019, June). Fully 3D Monte Carlo ambient noise tomography over Grane field. In *81st EAGE Conference and Exhibition 2019* (Vol. 2019, No. 1, pp. 1-5). European Association of Geoscientists & Engineers.
- Zhang, X., Hansteen, F., Curtis, A., & De Ridder, S. (2020a). 1-D, 2-D, and 3-D Monte Carlo Ambient Noise Tomography Using a Dense Passive Seismic Array Installed on the North Sea Seabed. *Journal of Geophysical Research: Solid Earth*, 125(2), e2019JB018552.
- 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.
- Zhang, X., Nawaz, M. A., Zhao, X., & Curtis, A. (2021). An introduction to variational inference in geophysical inverse problems. In *Advances in geophysics* (Vol. 62, pp. 73-140). Elsevier.
- Zhang, X., Roy, C., Curtis, A., Nowacki, A., & Baptie, B. (2020). 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. *Geophysical Journal International*, 222(3), 1639-1655.
- 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. *Geophysical Journal International*, 222(3), 1639-1655.
- Zhang, Z., Yao, H., & Yang, Y. (2020). Shear wave velocity structure of the crust and upper mantle in Southeastern Tibet and its geodynamic implications. *Science China Earth Sciences*, 63, 1278-1293.
- Zhang, Z., Yao, H., Wang, W., & Liu, C. (2022). 3-D crustal azimuthal anisotropy reveals multi-stage deformation processes of the Sichuan basin and its adjacent area, SW China. *Journal of Geophysical Research: Solid Earth*, 127(1), e2021JB023289.
- Zhao, X., & Curtis, A. (2024a). Bayesian inversion, uncertainty analysis and interrogation using boosting variational inference. *Journal of Geophysical Research: Solid Earth*, 129(1), e2023JB027789.

- [Zhao, X.](#), & Curtis, A. (2024b). Physically structured variational inference for Bayesian full waveform inversion. *Journal of Geophysical Research: Solid Earth*, 129(11), e2024JB029557.
- [Zhao, X.](#), & Curtis, A. (2025). Efficient bayesian full waveform inversion and analysis of prior hypotheses in 3D. *Geophysics*, 90(6), 1-58.
- [Zhao, Y.](#), Guo, Z., & Wang, Y. (2022). Asthenospheric Flow From East Asia to the Philippine Sea Plate Revealed by Rj-MCMC Inversion of Surface Waves. *Geochemistry, Geophysics, Geosystems*, 23(9), e2022GC010342.
- [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.
- [Zhdanov, M. S.](#) (2002). *Geophysical inverse theory and regularization problems* (Vol. 36). Elsevier.
- [Zheng, X. F.](#), Yao, Z. X., Liang, J. H., & Zheng, J. (2010). The role played and opportunities provided by IGP DMC of China National Seismic Network in Wenchuan earthquake disaster relief and researches. *Bulletin of the Seismological Society of America*, 100(5B), 2866-2872.
- [Zhu, H.](#) (2018). Seismogram registration via Markov chain Monte Carlo optimization and its applications in full waveform inversion. *Geophysical Journal International*, 212(2), 976-987.
- [Zielhuis, A.](#), & Nolet, G. (1994). Deep seismic expression of an ancient plate boundary in Europe. *Science*, 265(5168), 79-81.