Geophysical Journal International

Geophys. J. Int. (2025) **241,** 1088–1113
Advance Access publication 2025 March 12
GJI Seismology

https://doi.org/10.1093/gji/ggaf096

Annealed stein variational gradient descent for improved uncertainty estimation in full-waveform inversion

Miguel Corrales [1], Sean Berti,[2] Bertrand Denel,[3] Paul Williamson,[3] Mattia Aleardi[2] and Matteo Ravasi[1]

[1] *Physical Sciences and Engineering Division, KAUST, Thuwal 23955, Kingdom of Saudi Arabia. E-mail: miguel.corrales@kaust.edu.sa*
[2] *Earth Sciences Department, University of Pisa, I-56126 Pisa, Italy*
[3] *OneTech R&D, TotalEnergies, 64000 Pau, France*

Accepted 2025 March 11. Received 2025 January 21; in original form 2024 October 22

SUMMARY

In recent years, full-waveform inversion (FWI) has been extensively used to derive high-resolution subsurface velocity models from seismic data. However, due to the nonlinearity and ill-posed nature of the problem, FWI requires a good starting model to avoid producing non-physical solutions (i.e. being trapped in local minima). Moreover, traditional optimization methods often struggle to effectively quantify the uncertainty associated with the recovered solution, which is critical for decision-making processes. Bayesian inference offers an alternative approach as it directly or indirectly evaluates the posterior probability density function using Bayes' theorem. For example, Markov Chain Monte Carlo (MCMC) methods generate multiple sample chains to characterize the solution's uncertainty. Despite their ability to theoretically handle any form of distribution, MCMC methods require many sampling steps; this limits their usage in high-dimensional problems with computationally intensive forward modelling, as is the FWI case. Variational inference (VI), on the other hand, approximates the posterior distribution in the form of a parametric or non-parametric proposal distribution. Among the various algorithms used in VI, Stein Variational Gradient Descent (SVGD) is characterized for its ability to iteratively refine a set of samples (commonly referred to as particles) to approximate the target distribution through an optimization process. However, mode and variance-collapse issues affect SVGD in high-dimensional inverse problems. In this study, we propose to improve the performance of SVGD within the context of FWI by combining an annealed variant of the SVGD algorithm with a multiscale strategy, a common practice in deterministic FWI settings. Additionally, we demonstrate that principal component analysis (PCA) can help us to evaluate the performance of the optimization process and gain insights into the behaviour of the output particles and their overall distribution. Clustering techniques are also employed to provide more rigorous and meaningful statistical analysis of the particles in the presence of multimodal distributions (as is usually the case in FWI). Numerical tests, performed on a portion of the acoustic Marmousi model using both single and multiple frequency ranges, reveal the benefits of annealed SVGD compared to vanilla SVGD to enhance uncertainty estimation using a limited number of particles and thus address the challenges of dimensionality and computational constraints.

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

1 INTRODUCTION

Full-waveform inversion (FWI) is a high-resolution imaging technique for estimating subsurface parameters from recorded seismic waveform data. Unlike methods that rely solely on the kinematic component of the recorded seismic waveforms (i.e. traveltimes), FWI exploits the entire wavefield information to invert for detailed subsurface models (Virieux & Operto 2009). However, the complex and nonlinear relationship between model parameters and seismic data—coupled with the oscillatory nature of the seismic waveforms, incomplete data coverage and noise in the data—renders FWI an ill-posed inverse problem with a non-unique solution. In other words, many sets of model parameters can fit the data equally well within their inherent uncertainties; therefore, it is crucial to quantify the

range of possible solutions, to ultimately assess the confidence in the inverted models (Fernandez-Martinez *et al.* 2013).

FWI is typically addressed by minimizing a misfit function (e.g. L_2 norm) between the observed and simulated seismograms (Lailly & Santosa 1984; Tarantola 1984; Virieux & Operto 2009). The computational cost of the forward problem has historically imposed local optimization methods, but due to the highly nonlinear nature of the problem and the multimodal landscape of the objective function, local optimization algorithms often get trapped in local minima. This challenge can be mitigated by enforcing specific requirements on the observed data, such as the presence of low frequencies or long offset, and/or a good starting model. Gauthier *et al.* (1986), Bozdağ *et al.* (2011) and Guo *et al.* (2020) have shown that a poor starting model can easily lead the inversion into a local minimum of the objective function, compromising the inversion outcome. Various alternative misfit functions have been proposed to reduce this dependence (Luo & Schuster 1991; Sambridge & Mosegaard 2002; Brossier *et al.* 2010; Warner & Guasch 2014; Métivier *et al.* 2016). Finally, when deterministic optimization algorithms are used alongside these misfit functions, local approximations of the problem uncertainty around the maximum *a posteriori* (MAP) solution can be derived by estimating the inverse of the Hessian matrix (Tarantola 2005; Rawlinson *et al.* 2014; Liu & Peter 2019; Liu *et al.* 2021).

Bayesian inference extends deterministic optimization methods since it aims to quantify the uncertainty of the inverted models (Mosegaard & Tarantola 2002; Sambridge & Mosegaard 2002). Bayesian methods embody Bayes' theorem to update our prior knowledge about the model parameters with new information obtained from the observed data. In this case, the posterior probability distribution (PPD), or the number of samples taken from such a distribution, represents the solution of the inversion process, thus Bayesian methods offer a more comprehensive solution describing all parameter values consistent with the observed data and quantifying their relative probabilities. Markov Chain Monte Carlo (MCMC) methods are commonly employed to characterize the PPD by constructing multiple chains of successive samples from the posterior distribution through structured random walks in the parameter space. These samples form the basis for inferring valuable statistics of the PPD, and thereby enable the estimation of uncertainties that affect the recovered solution. One such method, the random walk Metropolis algorithm, has been applied across various geophysical problems, including electrical resistivity inversion (Malinverno 2002), traveltime tomography (Bodin & Sambridge 2009) and gravity inversion (Mosegaard & Tarantola 2002). However, this algorithm faces significant computational challenges, as the curse of dimensionality (Curtis & Lomax 2001) restricts its applicability in high-dimensional problems with computationally expensive forward modelling operators, such as those encountered in FWI.

Over the past decade, by leveraging rapid advances in computing capabilities, researchers have revisited sampling-based methods to solve Bayesian FWI problems, developing sophisticated algorithms that aim to improve the efficiency for large-scale inversions. However, while more computationally feasible than standard MCMC approaches, these methods may also risk losing information due to implicit undersampling in high-dimensional spaces. These include Hamiltonian Monte Carlo (HMC) (Fichtner *et al.* 2019; Gebraad *et al.* 2020), stochastic Newton MCMC (Martin *et al.* 2012), parallel tempering (Sambridge 2014) and gradient-based MCMC (Aleardi 2021; Zhao & Sen 2021; Berti *et al.* 2024a, b). Finally, trans-dimensional MCMC represent another class of MCMC techniques in which the number of model parameters is treated as an additional unknown (Bodin & Sambridge 2009; Ray *et al.* 2016; Sen & Biswas 2017; Guo *et al.* 2020). Despite their robustness and efficiency, when applied to problems such as FWI, MCMC methods typically require a large number of sampling steps and a long burn-in period to achieve accurate uncertainty estimations.

Variational inference (VI) has emerged as an appealing alternative as it offers greater adaptability for approximating the posterior distributions with significantly lower computational effort than MCMC (Jordan *et al.* 1998; Blei *et al.* 2017; Zhang *et al.* 2019). In VI, a set of simple probability distributions is defined (often called the variational family), and an optimal member of such a family is sought to approximate the true PPD. The Kullback–Leibler (KL) divergence measures the disparity between the two distributions, which enables potentially efficient and parallelizable optimization processes with well-understood convergence criteria. This can be achieved by either directly estimating the free parameters of the chosen distribution that best approximate the true PPD (Kingma *et al.* 2016; Kucukelbir *et al.* 2017) or by deterministically modifying a set of samples from the proposal distribution to match the PPD (Liu & Wang 2016; Gallego & Insua 2020). Variational approaches have been applied to various problems in geophysics, including traveltime tomography (Zhang & Curtis 2020a; Zhao *et al.* 2022), seismic denoising (Siahkoohi *et al.* 2021), seismic interpolation (Ravasi 2023), earthquake hypocentre inversion (Smith *et al.* 2022), 2-D FWI (Zhang & Curtis 2020b; Urozayev *et al.* 2022) and 3-D FWI (Lomas *et al.* 2023; Zhang *et al.* 2023).

Recently, particle-based methods have emerged to bridge the gap between parametric VI and MCMC techniques. These methods utilize a specific number of samples, or particles, to represent the approximate distribution, akin to MCMC, whilst updating these particles through an optimization process similar to VI. This hybrid approach offers greater flexibility than parametric VI and is more particle-efficient than MCMC, as it fully leverages particle interactions. A notable example of this category is the Stein Variational Gradient Descent (SVGD) method (Liu & Wang 2016), which has already been applied to post-stack seismic inversion (Izzatullah *et al.* 2024b), petrophysical inversion (Corrales *et al.* 2022) and FWI (Zhang *et al.* 2023; Izzatullah *et al.* 2024a; Berti et al. 2025). This sampling algorithm iteratively minimizes the Kullback–Leibler (KL) divergence between the chosen approximate distribution and the target density (Kullback & Leibler 1951) to ensure that the final set of particles is distributed according to the desired posterior distribution. Despite the empirical successes of SVGD, its application to high-dimensional problems remains challenging, as it becomes computationally demanding to sample more particles than there are unknowns. SVGD can suffer from mode- and variance-collapse issues as the dimensionality of the problem increases. More specifically, variance collapse refers to the scenario in which the variance estimated by SVGD is significantly smaller than the true variance of the target distribution (Zhuo *et al.* 2018). This is undesirable because underestimation of the variance leads to a failure in explaining the uncertainty of the model predictions, which is a key benefit of Bayesian inference.

This study aims to enhance the performance of SVGD within the FWI framework by replacing the standard SVGD algorithm with an annealed variant. We conduct numerical experiments on a portion of the Marmousi model using both single- and multiple frequency approaches to evaluate the effectiveness of these methods to improve uncertainty estimation when working with a limited number of particles (where the number of particles is less

than the number of model parameters). Furthermore, we propose a number of additional techniques to augment our analysis of the particles after the SVGD process. First, principal component analysis (PCA) is used to evaluate the performance of SVGD and gain deeper insights into the behaviour and distribution of the particles. In addition to reducing the dimensionality of the solution space, PCA highlights the least confident model combinations, enabling more precise analysis of model uncertainties compared to simply examining the parameter variances (the diagonal of the covariance matrix). Additionally, we employ clustering techniques to identify whether particles converge to distinct modes, allowing for more rigorous and meaningful statistical insights by grouping particles into geological and non-geological clusters. Overall, we show that these techniques may help to bridge the gap between theoretical potential and practical applications for high-dimensional, and computationally intensive problems to enable better-informed decisions.

2 THEORETICAL FRAMEWORK

FWI aims to estimate subsurface model parameters, such as *P*-wave velocity, represented by $\mathbf{m} \in \mathbb{R}^m$, from observed seismic data $\mathbf{d} \in \mathbb{R}^d$, where m and d denote the dimensions of model and data spaces, respectively. In order to capture the uncertainties inherent in this estimation process, any Bayesian estimation algorithm formulates this inverse problem using Bayes' rule:

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

where the probability density function (PDF) of the posterior, $p(\mathbf{m}|\mathbf{d})$, is determined by the likelihood $p(\mathbf{d}|\mathbf{m})$, describing the conditional probability of successfully modelling the seismic data given a seismic velocity model, and by our prior knowledge $p(\mathbf{m})$ of the model parameters, which reflects our initial confidence in the unknown model based on any available prior information. Lastly, a normalization constant, also known as the evidence, $p(\mathbf{d})$, ensures that the posterior distribution properly integrates to one over the entire parameter space; however, this is often computationally intractable.

2.1 Variational inference

At the core of variational inference lies the idea of approximating this posterior distribution with a simpler, surrogate distribution, denoted as $q(\mathbf{m})$. This distribution should be selected from a family (called variational family) that is easy to sample and evaluate; a common choice is therefore the Gaussian distribution. The essence of this optimization process lies in minimizing the Kullback–Leibler (KL) divergence (Kullback & Leibler 1951; Blei *et al.* 2017), which quantifies the discrepancy between the surrogate (approximate) distribution and the target (posterior) distribution. The KL divergence is expressed as:

$$\mathrm{KL}(q(\mathbf{m})|p(\mathbf{m}|\mathbf{d})) = \mathrm{E}_{\mathbf{m}\sim q}[-\log p(\mathbf{m}|\mathbf{d}) + \log q(\mathbf{m})]. \tag{2}$$

By minimizing eq. (2), we are able to compute the expectation of the model parameters sampled from the surrogate distribution. Consequently, VI achieves an approximation of the posterior distribution through the minimization outlined below:

$$q^* = \underset{q}{\operatorname{argmin}}\ \mathrm{KL}(q(\mathbf{m})|p(\mathbf{m}|\mathbf{d})). \tag{3}$$

2.2 Stein variational gradient descent

Unlike most VI techniques, SVGD is a deterministic, particle-based inference algorithm that iteratively minimizes the KL divergence between the chosen approximate distribution and the target density. This innovative method leverages the concept of functional gradients to effectively transport a pre-defined set of particles towards the target distribution. The transport occurs within the Reproducing Kernel Hilbert Space (RKHS), guided by the gradient of the KL divergence [for a detailed derivation of the SVGD formulation, we refer the reader to Liu & Wang (2016)].

Given a collection of particles, the optimal update direction ϕ^* of eq. (3), for each particle, is given by:

$$\phi^*(\cdot) = \mathrm{E}_{\mathbf{m}\sim q}\underbrace{[k(\mathbf{m},\cdot)\nabla_{\mathbf{m}}\log p(\mathbf{m}|\mathbf{d})}_{\text{driving force}} + \underbrace{\nabla_{\mathbf{m}}k(\mathbf{m},\cdot)]}_{\text{repulsive force}}, \tag{4}$$

where $k(\cdot,\cdot)$ is called the kernel function that quantifies the distance between different particles. If we denote the particles that we are using to represent q as $\{\mathbf{m}_i\}_{n=1}^N$, the expectation in eq. (4) can be approximated using the sample mean over the particles. Thus, the KL divergence can be iteratively minimized as follows:

$$\phi^*_{q_l,p}(\mathbf{m}) = \frac{1}{N}\sum_{j=1}^{N}[k(\mathbf{m}^l_j,\mathbf{m})\nabla_{\mathbf{m}^l_j}\log p(\mathbf{m}^l_j|\mathbf{d}) + \nabla_{\mathbf{m}^l_j}k(\mathbf{m}^l_j,\mathbf{m})] \tag{5}$$

$$\mathbf{m}^{l+1}_i = \mathbf{m}^l_i + \epsilon_l\phi^*_{q_l,p}(\mathbf{m}^l_i),$$

where l denotes the current iteration, N is the number of particles and ϵ_l is the step size. Assuming the step size to be sufficiently small, the process asymptotically converges to the target posterior as the number of particles tends to infinity.

The RHS of eq. (4) comprises two distinct terms: the *driving force* and the *repulsive force*. The driving force aims to direct the particles towards higher probability regions. Conversely, the repulsive force has the crucial role of maintaining particle diversity, and actively prevents particle collapse by dispersing particles across the parameter space. This balance enables a comprehensive exploration and characterization of the target distribution.

Various types of kernel functions have been proposed over the last few years (Liu & Wang 2016; Gorham & Mackey 2017; Zhang & Curtis 2020b). In our study, we consider the two most commonly employed kernels, namely the radial basis function (RBF) and the inverse multi-quadratic (IMQ) (Wang *et al.* 2019). The RBF kernel is defined as:

$$k(\mathbf{m},\mathbf{m}') = \exp\left(-\frac{||\mathbf{m}-\mathbf{m}'||^2}{2h^2}\right), \tag{6}$$

where h is the bandwidth, a scaling factor that controls the strength of the interaction between different particles based on their distances. As suggested by Liu & Wang (2016), we set $h = \tilde{d}^2/\log N$, where $\tilde{d}$ is the median of pairwise distances between all particles. It is worth noting that this parameter is recalculated at each iteration with limited heuristic justification, and there is some evidence that this nonlinearity can generate instability. The IMQ kernel, instead, is defined as:

$$k(\mathbf{m},\mathbf{m}') = \left(c^2 + \frac{||\mathbf{m}-\mathbf{m}'||^2}{2h^2}\right)^{\beta}, \tag{7}$$

where c and β are two user-defined parameters. We set c and β to 1 and $-\frac{1}{2}$, respectively, as suggested by Gorham & Mackey (2017).

2.3 Annealed stein variational gradient descent

Despite the empirical success of SVGD, convergence guarantees are absent—except in the mean-field limit (where the number of particles $N \to \infty$, while the dimension d is kept fixed). Zhuo *et al.* (2018) showed that SVGD encounters degeneracy issues under finite particle conditions, which cause the particles to collapse into a small number of modes—the so-called mode collapse issue. On the other hand, as dimensionality increases (such that $d > N$), the variance estimated by SVGD may significantly underestimate the variance of the target distribution—a phenomenon known as variance collapse; this is an important concern in high-dimensional problems like FWI, which also affects other particle-driven approaches such as ensemble-based methods (Liu & Grana 2018; Thurin *et al.* 2019).

Ba *et al.* (2021) compared the SVGD update to the application of gradient descent to a maximum mean discrepancy (MMD) objective function. In a high-dimensional example, they found that SVGD and MMD descent differ primarily in the driving force term that becomes increasingly problematic in higher dimensions. They have empirically demonstrated that removing the bias introduced by the deterministic update present in the driving force leads to more accurate estimation of the variance. Consequently, Ba *et al.* (2021) proposed modifying the driving force term of SVGD with a damped version, resulting in the damped SVGD approach. At a high level, this modification mirrors the approach taken by D'Angelo & Fortuin (2021) under the name of Annealed SVGD (A-SVGD), where a heuristic temperature parameter $\alpha(l) \in [0, 1]$ is introduced to adjust the intensity of the driving force. The updated rule is then simply expressed as follows:

$$\phi^*_{ql,p}(\mathbf{m}) = \frac{1}{N}\sum_{j=1}^{N}[\alpha(l)k(\mathbf{m}^l_j, \mathbf{m})\nabla_{\mathbf{m}^l_j} \log p(\mathbf{m}^l_j|\mathbf{d}) + \nabla_{\mathbf{m}^l_j} k(\mathbf{m}^l_j, \mathbf{m})]. \tag{8}$$

Varying the α parameter within the [0,1] range induces two distinct phases. A first exploratory phase, dominated by a strong repulsive force (α close to 0) that disperses the particles from their initial positions, facilitating broad coverage of the target distribution. This is followed by a second exploitative phase, where the driving force dominates (α close to 1) and concentrates the particle distribution around different modes. The selection of the temperature parameter $\alpha(l)$ is crucial to maintain the convergence properties of SVGD, in order to ensure that the final iterations operate effectively on the target density, that is, $\lim_{l\to\infty} \alpha(l) = 1$. In this work we will employ and compare the performances of two different annealing schedules proposed by D'Angelo & Fortuin (2021). The first schedule is the hyperbolic tangent, defined as:

$$\alpha(l) = \tanh\left[\left(1.3\frac{l}{T}\right)^p\right], \tag{9}$$

where l is the current iteration, T is the total number of iterations and p is a user-defined parameter that controls the rate of transition between the two phases. The second is the cyclic schedule, which allows a sequence of exploratory and converging phases and can be defined as:

$$\alpha(l) = \left(\frac{\mathrm{mod}(l, T/C)}{T/C}\right)^p, \tag{10}$$

where C is the number of cycles, where the α value ranges from 0 to 1.

2.4 Principal component analysis

To enhance our analysis of the particles produced during the optimization process, we propose to use PCA and clustering techniques. PCA is a dimensionality reduction technique, particularly effective when dealing with high-dimensional and highly correlated data. The main objective of PCA is to find a smaller set of features that can accurately represent the original data in a lower dimensional space, while preserving as much information as possible (Hotelling 1933). PCA can be summarized as follows.

First, let us consider a set of velocity particles $\mathbf{X}$, expressed as a $n \times m$ matrix, where n is the number of particles (observations) and m is the number of dimensions (variables). PCA begins by computing the mean vector $\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i$, which contains the mean of each column and allows us to standardize the data to have a zero mean:

$$\mathbf{Z} = \mathbf{X} - \mathbf{I}\bar{\mathbf{X}}. \tag{11}$$

Next, we compute the covariance matrix $\mathbf{C}$ of the mean-centred data $\mathbf{Z}$:

$$\mathbf{C} = \frac{1}{n-1}\mathbf{Z}^T\mathbf{Z}. \tag{12}$$

We then perform eigenvalue decomposition on the covariance matrix $\mathbf{C}$:

$$\mathbf{CV} = \mathbf{V}\Lambda, \tag{13}$$

where $\mathbf{V}$ is the $m \times m$ matrix of eigenvectors (principal components) and Λ is the $m \times m$ diagonal matrix of eigenvalues. Finally, we sort the eigenvalues in descending order and reorder the eigenvectors accordingly to project the mean-centred data onto the new principal components:

$$\mathbf{Y} = \mathbf{ZV}, \tag{14}$$

where $\mathbf{Y}$ is the $n \times m$ matrix of the transformed data.

In this study, we apply PCA to the entire set of particles collected per iteration (after SVGD is performed) for a number of components equal to $\min(m, n-1)$, where n is the number of particles. We aim to obtain the explained variance per component, which quantifies the proportion of the total variability in the particles that each principal component captures. The explained variance for each component is defined by:

$$\text{Explained variance of the } i\text{-th component} = \frac{\lambda_i}{\sum_{j=1}^{n}\lambda_j}, \tag{15}$$

where λ_i represents the eigenvalue of the i-th component, and $\sum_{j=1}^{n}\lambda_j$ denotes the sum of all eigenvalues, which accounts for the total variance. Analysing the explained variance per component helps us understanding the behaviour and distribution of the particles.

2.5 Clustering

Given the highly nonlinear nature of FWI, the particles obtained at the end of the SVGD iterations may have converged to different modes (i.e. become trapped in various local minima). Whilst the particles that have reached the global minimum are likely to be geologically meaningful, this may not be the case for those that reached other basins of attraction. Consequently, it is important to identify clusters of particles within our high-dimensional model space and distinguish those that are geologically meaningful from the other ones.

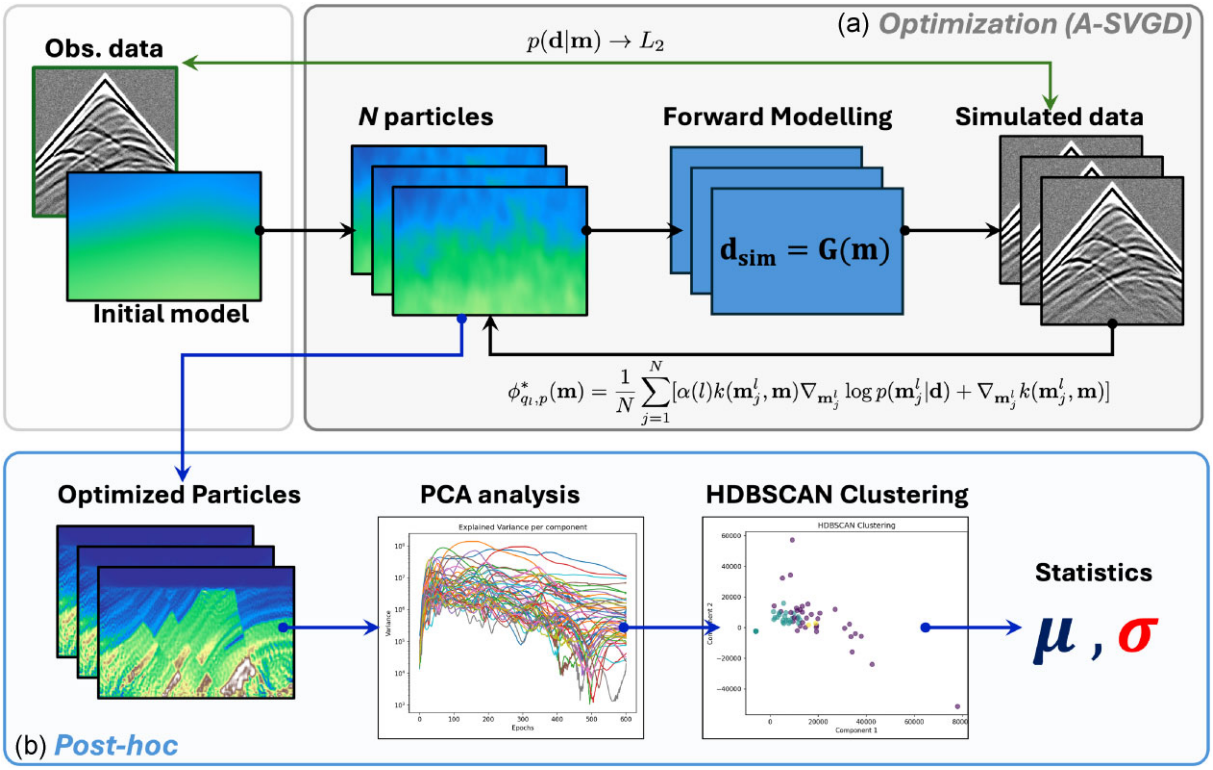

Figure 1. Schematic representation of the workflow used in this study, divided into two main stages: (a) Optimization and (b) Post-hoc analysis.

In order to do so, we have employed the hierarchical density-based spatial clustering of applications with noise (HDBSCAN) to perform such a clustering operation (Campello *et al.* 2013). HDBSCAN is an advanced clustering algorithm that extends density-based spatial clustering of applications with noise (DBSCAN; Ester *et al.* 1996). In short, HDBSCAN uses the mutual reachability distance, a metric that combines a density-based measure with pairwise distances to facilitate meaningful clustering. The algorithm creates a minimum spanning tree (MST) from the mutual reachability distances, assembling the basis of the hierarchical cluster tree. Through condensed clustering, HDBSCAN extracts significant clusters by systematically removing edges in the MST. Finally, stability-based clustering selects the most stable clusters from the hierarchical tree. Unlike other clustering methods, HDBSCAN is well-suited for high-dimensional data and does not require the user to specify the number of clusters in advance. This flexibility makes HDBSCAN an ideal candidate for our analysis, as it allows for robust identification of clusters and noise.

2.6 Workflow

Fig. 1 provides a schematic representation of the workflow utilized in this study. The workflow is divided into two main stages: optimization (Fig. 1a) and post-hoc analysis (Fig. 1b).

In the optimization stage (Fig. 1a), the starting point is a smooth velocity model used as the initial guess for FWI. Gaussian random field (GRF) perturbations are applied to this initial model to generate the different particles. For each particle, forward modelling is conducted to simulate the corresponding data, which is then used to compute the likelihood function. Subsequently, SVGD or its annealed variant is employed. These methods iteratively minimize the KL divergence between the particle and posterior distributions. The optimization process concludes when the particles converge to represent the posterior distribution.

The post-hoc analysis stage (Fig. 1b) begins with applying PCA to the optimized particles'; this provides a more comprehensive analysis than simply examining the parameter variances (the diagonal elements of the covariance matrix). In addition, the optimized particles are subjected to clustering using HDBSCAN. This clustering analysis identifies whether the particles converge to distinct modes, enabling a more precise differentiation between geological and non-geological features. By grouping particles into these modes, this analysis provides robust and meaningful statistical insights, such as estimates of each cluster's mean and standard deviation.

3 NUMERICAL EXAMPLES

3.1 Synthetic data

In this section, we present a series of numerical experiments using a portion of the Marmousi model (Brougois *et al.* 1990) to evaluate the capability of the previously described SVGD approaches to estimate the uncertainty associated with FWI when working with a limited number of particles. Synthetic data are modelled using Deepwave (Richardson 2023) in a rectangular uniform grid with dimensions of $n_z = 81$ and $n_x = 216$ and spacing of 20 m in both directions, using five shots evenly distributed along the horizontal axis and recorded by 201 receivers placed at 1-m intervals. The seismic source signature is a Ricker wavelet with a peak frequency of 7 Hz. The recording time is set to 3 s and the sample interval is 1 ms. To simulate more realistic conditions, Gaussian noise is added to the synthetic data, resulting in a signal-to-noise ratio (SNR) of approximately 17 dB (see Fig. 2).

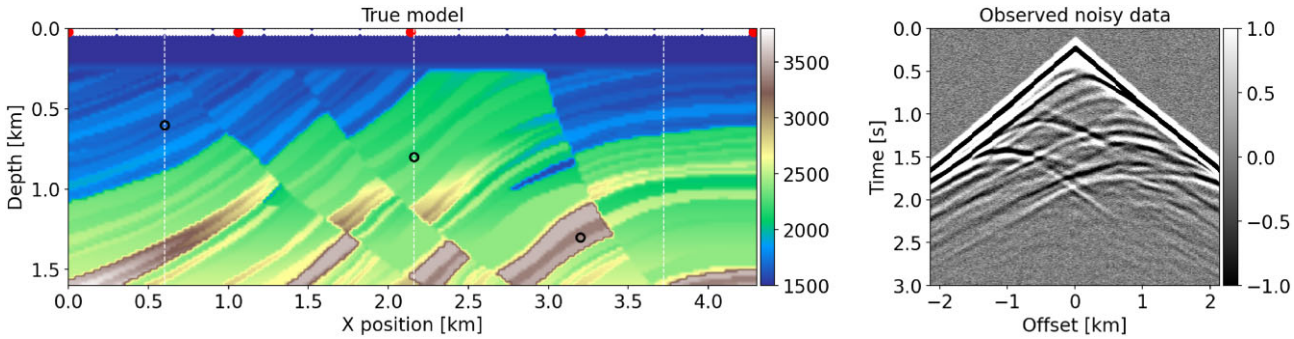

Figure 2. Portion of the Marmousi model used in our numerical experiments. The left panel illustrates the velocity model in meters per second ($m\,s^{-1}$), with red dots indicating shot locations and white dots representing receivers. Vertical dashed lines show the positions of the pseudo-well logs, and black circles indicate pixel locations for marginal plots discussed in the results section and appendix. The right panel shows the shot gather for the source in the middle of the model.

3.2 Selection of hyperparameters

The primary objective of this work is to investigate the uncertainty associated with the modelling operator (the likelihood is assumed to be Gaussian) by deliberately excluding any influence of our prior knowledge, rather than imposing a uniform bound limits on velocities. Since the uncertainty associated with the FWI problem is expected to encapsulate both the scattering (high wavenumber) and transmission (low wavenumber) components of the model, the choice of perturbations of the initial particles is crucial. Drawing inspiration from Izzatullah *et al.* (2024a), we aim to explore the uncertainties of these components by generating Gaussian random field (GRF) perturbations that introduce variability in both amplitude and scale.

In terms of the hyperparameters employed in this study, we opt for the Gaussian RBF and IMQ kernels. Bandwidth selection is performed using both the median trick and a fixed constant value. The constant value of the bandwidth is determined after a meticulous analysis of the bandwidth evolution in the scenario where the median trick is used. Finally, we use the Adam optimizer to update the particles at each iteration, given their gradient in eq. (5), with a constant learning rate of 100 and a fixed number of iterations equal to 600.

We conduct two sets of experiments: one with a small number of particles (50) and the other with a larger number (200). We use both the vanilla and annealed SVGD algorithms to assess the impact of annealing on the severity of mode and variance collapse in each case. For the annealed SVGD, we consider two temperature schedules, namely hyperbolic and cyclic. For the hyperbolic formulation (eq. 9), we set $p = 3$ and maintained $\alpha = 1$ for the last 20 per cent of iterations. For the cyclic formulation (eq. 10), we set $p = 2$ and $C = 8$, ensuring that the temperature remains at $\alpha = 1$ during the last two cycles.

The overall performance in terms of data misfit (L_2 norm) and SNR for the various experiments with different hyperparameters is presented in Figs A1 and A2. The subsequent results and discussion focus on the vanilla and annealed SVGD (tanh) formulations, utilizing the RBF kernel with the median trick and a set of 200 particles, as these configurations showed the best performance.

3.3 Single-scale experiments

In the first set of experiments, we perform FWI using a single frequency band with a peak frequency of 7 Hz. Initial particles are generated by applying GRF perturbations with different variances to a highly smoothed version of the true Marmousi model. During the inversion process, we impose lower and upper-velocity bounds, allowing velocity values between 1500 and 4370 $m\,s^{-1}$.

Fig. 3 illustrates the performances of the different methods in terms of data misfit and SNR with respect to the true model, computed from the mean over iterations for the various experiments. The vanilla and annealed SVGD with hyperbolic tangent demonstrate superior performance compared to the other methods (see Appendix A). Notably, the annealed SVGD using the cyclic formulation results in poorer data misfit and SNR; this may be due to its design, which aims to explore the parameter space better and identify widely separated high-probability regions. As a result, the mean model may be perturbed, fitting the data less precisely compared to a more compact, but less comprehensive, posterior distribution.

A more detailed comparison of the mean and standard deviation for the 200-particle experiments is presented in Figs A3 and A4, respectively. Fig. 4 shows the mean obtained after 600 iterations for the vanilla and annealed versions for the 200-particle experiment, along with the mean of the initial distribution. All predicted models are relatively similar in the shallower part (<1 km) and reproduce most of the main features of the true model. However, the model reconstruction is poorer in the deeper part, likely due to a lack of illumination (especially near the edges of the model). As stated before, the results obtained using the vanilla SVGD and the annealed SVGD with the hyperbolic tangent formulation (both with the RBF kernel) exhibit higher SNR and are less affected by artefacts. In contrast, the models predicted using other approaches show errors in terms of velocity magnitudes, particularly in the last 500 m of depth.

In addition, Fig. 4 presents the standard deviation maps (expressed in $m\,s^{-1}$) obtained after 600 iterations using 200 particles. Both maps exhibit a similar expected pattern, with very low values in the shallower parts where illumination is greater and the values increasing towards the deeper portions of the model. These deeper regions, along with the lateral edges of the model, are expected to have higher uncertainties due to limitations in acquisition geometry and the physics of wave propagation within the subsurface. Higher uncertainties are also observed in areas of high velocity and near the main velocity contrasts.

The standard deviation associated with the initial distribution is relatively low ($< 150\ m\,s^{-1}$) across the entire model. We select a narrow proposal for the initial samples to prevent SVGD from rapidly repelling particles into undesirable modes at the early stages (particles converging to suboptimal local minima). By maintaining

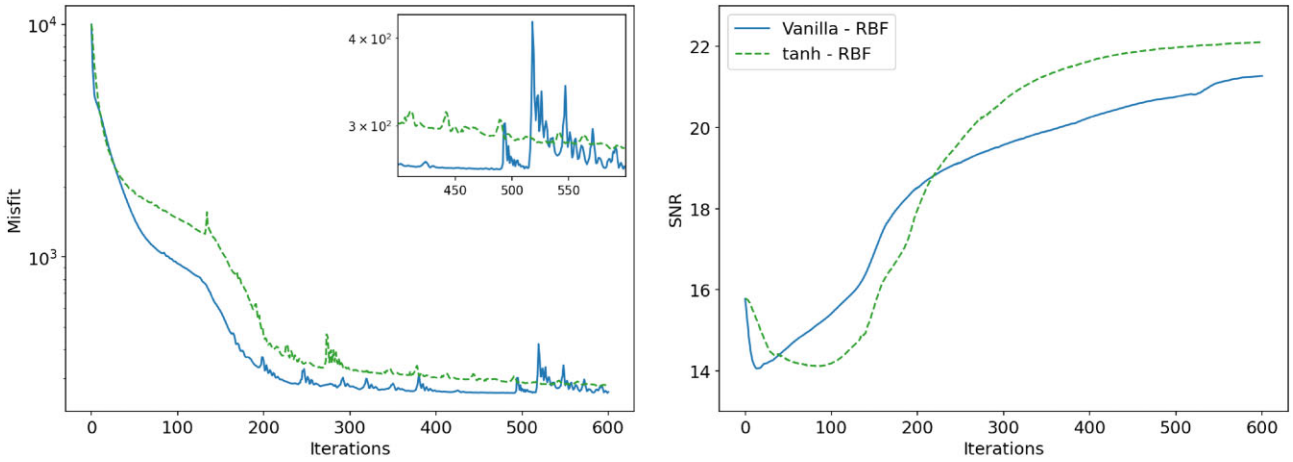

Figure 3. Data misfit (left) and SNR with respect to the true model (right) for vanilla and annealed (tanh) SVGD in the single-frequency scenario using 200 particles. The zoom window provides a clearer misfit comparison for the final 150 iterations.

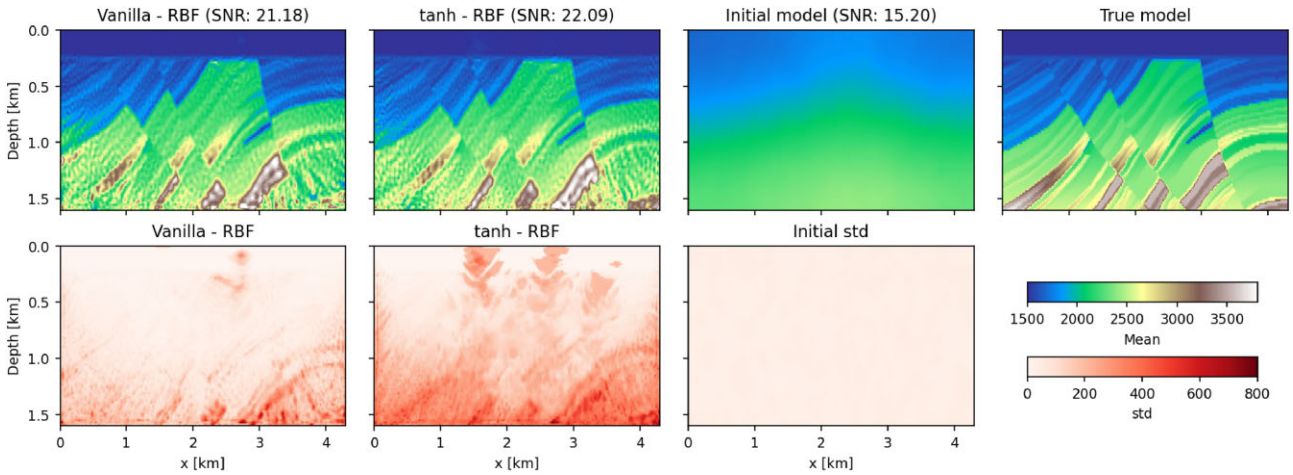

Figure 4. Mean and standard deviation comparison of the experiments using 200 particles for vanilla SVGD and annealed SVGD using RBF kernel and median trick after 600 iterations in the single-scale scenario. The velocity values are expressed in m s^{-1}.

an initial low standard deviation, we aim to direct the convergence towards fewer and more geologically consistent local minima. In contrast, the values associated with the predicted models are substantially higher, reaching over 800 m s^{-1}. Notably, the standard deviation maps associated with the annealed version of SVGD show significantly higher values throughout the model compared to those associated with the vanilla SVGD. This indicates that the annealed approach allows for better exploration of the model space and reduces the variance collapse phenomenon affecting the vanilla SVGD.

Moreover, to better highlight the differences between the vanilla and annealed approaches, in Figs 5 (left) and 5 (right) we present three pseudo well logs corresponding to three distinct spatial locations, as indicated by the white vertical lines in Fig. 2. For each position, we illustrate the true velocity varying with depth, the velocities extracted from the mean model obtained using both approaches, and three confidence intervals (corresponding to one, two and three standard deviations) based on the standard deviation maps shown in Fig. A4. We observe that in both cases, the width of the confidence intervals increases with depth, indicating larger uncertainty at greater depths, as expected. Within the first 800 m of depth, the mean model closely resembles the true model, with standard deviation values close to zero. The main differences between the two approaches become more evident at greater depths, where discrepancies between the predicted and true models are more significant. Specifically, the logs associated with the annealed SVGD show larger confidence intervals, with the values extracted from the true model falling almost entirely within these bounds. In contrast, the vanilla SVGD approach yields smaller confidence intervals, with the true values occasionally falling outside these bounds (e.g. the pseudo well at the spatial position of 0.6 km and around 1.4 km of depth). This comparison indicates that, while the annealed approach captures a broader range of uncertainty, the vanilla approach may sometimes underestimate the true variability at greater depths. Similar results are also obtained at three different pixel locations for vanilla (see Fig. 6) and annealed formulations (see Fig. 7).

To further investigate the behaviour and distribution of particles during the SVGD optimization process, we performed PCA. This analysis helps us understand the explained variance of the components, providing insights into how the particles evolve throughout the optimization. By examining the explained variance for $N-1$ components (where N is the number of particles), it is possible to identify the dominant directions in which the particle positions vary the most. This allows us to understand the main variances captured by components, the structural dynamics of the particle distribution

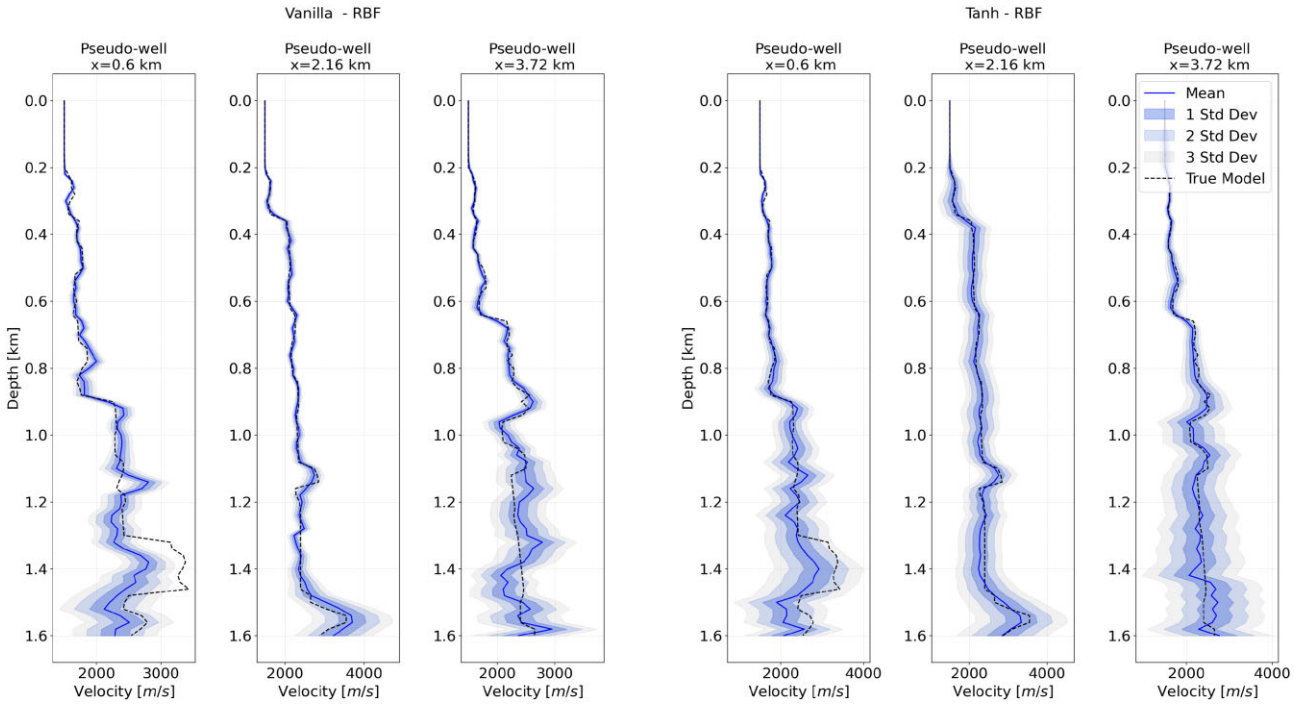

Figure 5. Pseudo-well marginal distributions showing mean and three confidence intervals for the experiment using vanilla SVGD (left) and annealed SVGD (right), using the RBF kernel with median trick and 200 particles in the single-frequency scenario.

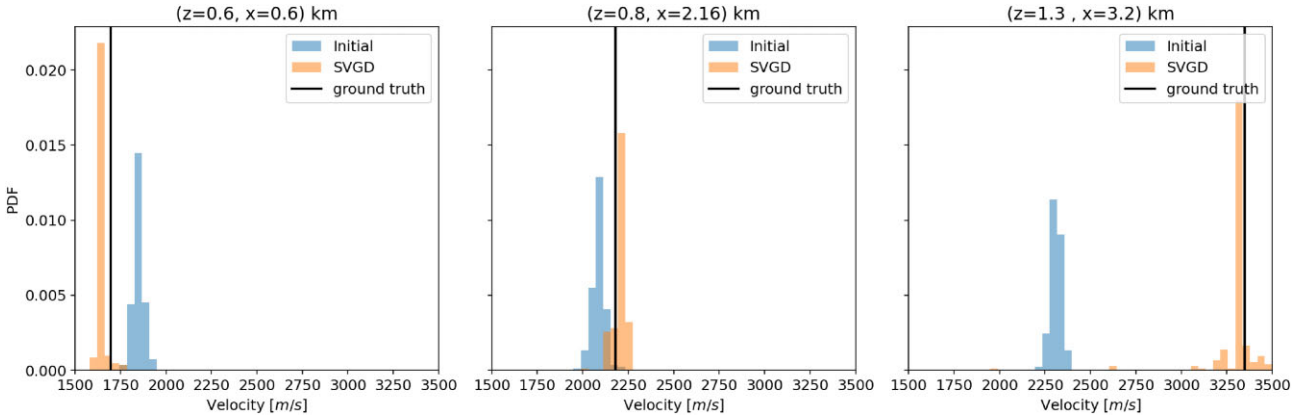

Figure 6. Single-scale scenario: point-wise marginals at three different locations for vanilla SVGD with RBF kernel and 200 particles.

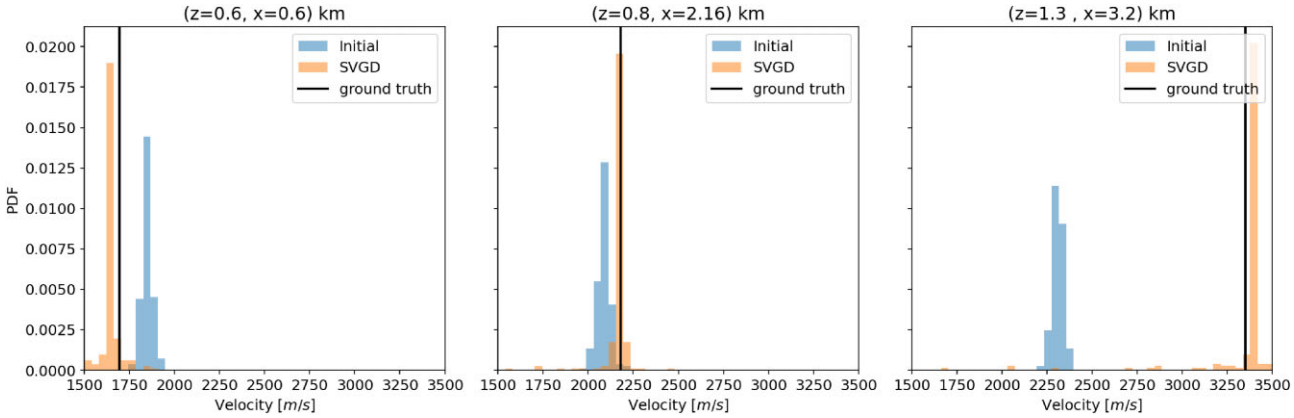

Figure 7. Single-scale scenario: point-wise marginals at three different locations for annealed SVGD (tanh) with RBF kernel, and 200 particles.

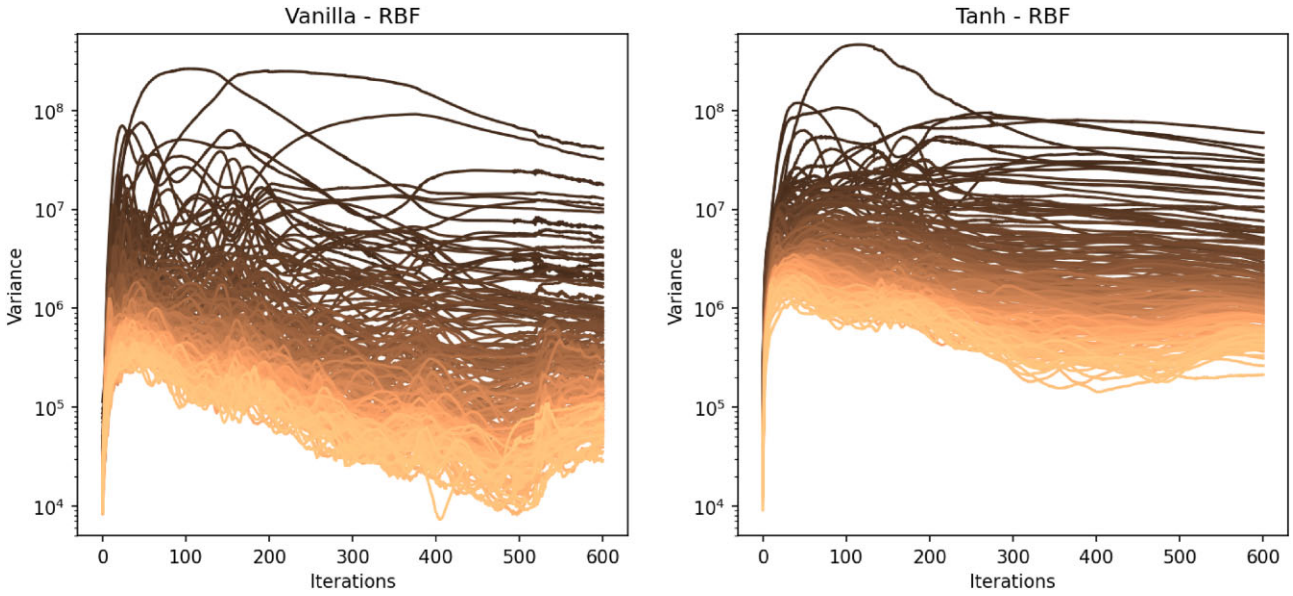

Figure 8. Explained variance (without normalization) per component (using $N-1$ components, where N is the particle number) for the experiments using (left) vanilla SVGD with the RBF kernel and (right) annealed SVGD with the tanh formulation and RBF kernel with 200 particles.

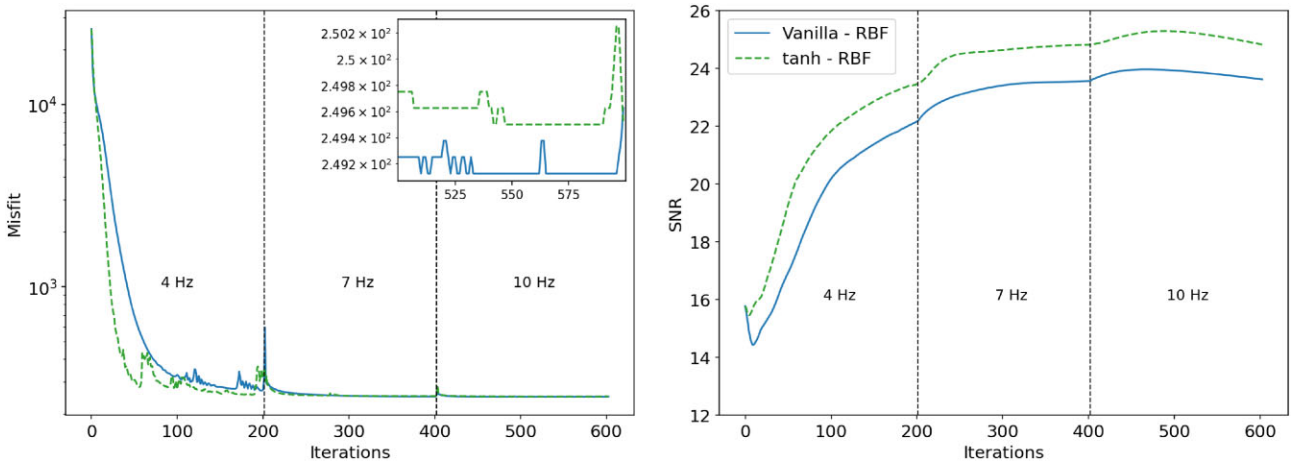

Figure 9. Data misfit (left) and SNR (right) across different experiments for 200 particles in the multiscale scenario. The close-up window provide a clearer comparison of the final 150 iterations.

and take into account the spatial correlations between (nearby) pixel values, which are ignored by the simple pixel-wise std plots.

Fig. 8 (left) illustrates the PCA variances for the vanilla SVGD case with 200 particles using the RBF kernel. We observe a relatively uniform spread of the components over the variance but erratic convergence, with a significant portion of the components converging towards mid- and low-variance values. In contrast, Fig. 8 (right) shows the PCA for the annealed SVGD with the tanh formulation, which demonstrates a more stable convergence pattern and a more balanced distribution of variance between high and low variance directions. These PCA results highlight the effectiveness of the annealed SVGD to provide a more stable and informative representation of the particle dynamics compared to the vanilla SVGD approach.

3.4 Multiscale experiments

In this section, we present results for the multiscale FWI experiments, focusing exclusively on the Vanilla and Annealed variants of SVGD (see Figs 9 and 10), as these have demonstrated superior performance in our single-frequency experiments. The multiscale approach [introduced by Bunks *et al.* (1995)] has become a standard practice in FWI due to its ability to improve the inversion's convergence and accuracy, mitigating the cycle-skipping issue. It begins with the lowest frequencies available in the observed data and progressively incorporates higher frequencies, thereby mitigating the nonlinearity of the inversion process by initially targeting large-scale features, which are more sensitive to low frequencies and then refining the model with higher frequencies to capture higher resolution details. Specifically, the inversion process is conducted in three stages: 200 iterations at a peak frequency of 4 Hz, 200 iterations at 7 Hz and 200 iterations at 10 Hz, with learning rates of 100, 10 and 10, respectively. This step-wise frequency escalation ensures that each scale of the model is accurately resolved before moving on to the next, providing a robust framework for the inversion process.

As in the single-frequency experiment, the multiscale approach using annealed SVGD demonstrates superior performance compared to the vanilla formulation in terms of SNR with respect to the true model (see Fig. 9). It is essential to highlight that multiscale SVGD not only achieves higher SNR values but also yields more

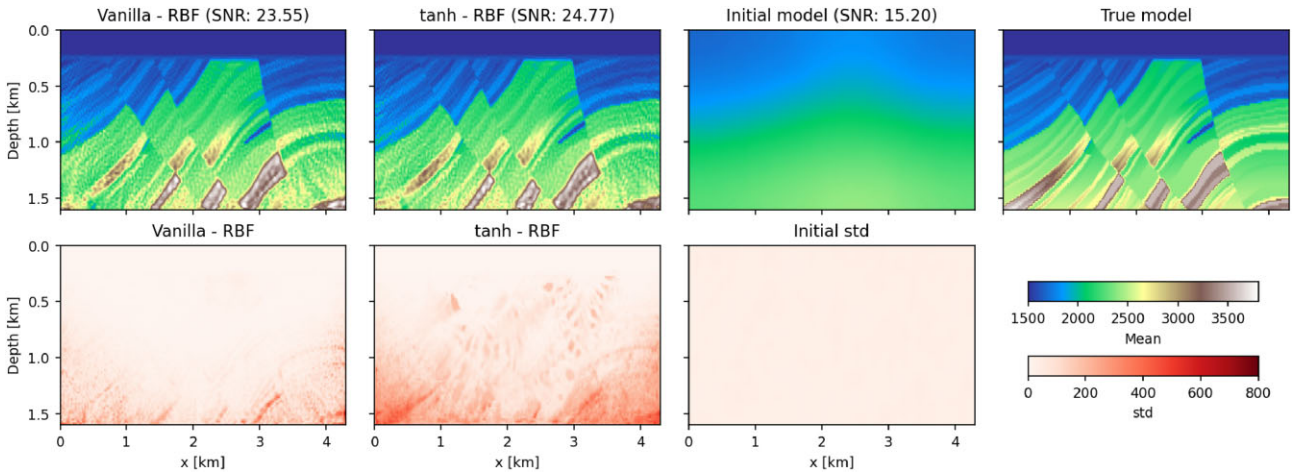

Figure 10. Mean and standard deviation comparison of the experiments using 200 particles for vanilla SVGD and annealed SVGD using RBF kernel and median trick after 600 iterations in the multiscale scenario. The velocity values are expressed in m s^{-1}.

meaningful and representative statistics compared to the single-frequency approach. This improvement is primarily due to the enhanced control over the particle refinement process, which prevents some particles from diverging and causing high standard deviation values. Consequently, the optimization process is more likely to converge toward geologically representative models, leading to more accurate and reliable results.

Appendices A and B provide supplementary results for both the single-scale and multiscale approaches, respectively, which further strengthen the evidence supporting our findings.

3.5 Cluster analysis and statistical evaluation

Our single-frequency experiments yield several key observations. A small number of principal components explain the majority of the variance (Fig. 8, left). Consequently, many components capture small variances rather than meaningful patterns in the data. This is further supported by the presence of abnormal individual particles (see Appendix A) and artefacts in the shallower parts of the particles, which are evident in the respective mean and standard deviation maps (Figs A3 and A4). These observations suggest that the particles tend to converge to different modes, some of which may not be geologically meaningful (since we have not added extra *a priori* information in our objective function). Although these particles fit the data term, they do not accurately represent the subsurface structure. Therefore, we propose to perform clustering to identify the presence of different modes and conduct statistical analysis within each cluster, rather than assuming all samples have converged to the same global minima.

To illustrate the importance of identifying clusters in the final particles and using such information for any subsequent statistical analysis, we consider the experiment with annealed SVGD employing a tanh temperature schedule and RBF kernel for a set of 200 particles. After the optimization process (600 iterations), we apply HDBSCAN to the final particles. In this case, the clustering algorithm produces three distinct groups, labelled as -1, 0 and 1, of sizes 35, 163 and 2, respectively. For visualization purposes, we plot the particles in the 2-D space defined by the first two components of the PCA and colour-code the particles to indicate which cluster they belong to (Fig. 11 left). We then compute the SNR for each individual particle in the different clusters and display their distribution in Fig. 11 (right). Overall, cluster 0 and cluster 1 contain particles with higher SNR values, although some of the particles in cluster 0 converge to similar SNR values. More importantly, cluster -1 is identified as the noisy cluster, representing particles not assigned to any other cluster, to be considered as outliers and geologically implausible (Fig. 12).

Further analysis of the mean and standard deviation for each cluster (Fig. 13) reveals that while the mean of cluster -1 appears reasonable, the individual particles are not representative, which leads to significant variations in the shallow parts of the model and, therefore, high uncertainty. This observation correlates with the artefacts detected in the shallow regions of the results in subsection 3.3. In contrast, clusters 0 and 1 display more consistent and plausible mean and standard deviation patterns, similar to those obtained with the multiscale approach. This underscores the importance of performing clustering analysis of the particles produced by SVGD in order to discard non-representative particles and thereby obtain more accurate and geologically plausible outcomes.

4 DISCUSSION

In the context of FWI, uncertainty quantification using SVGD (or other particle-based methods) presents significant computational challenges. Given the high-dimensionality of our model space, we operate in a regime where the number of particles is much smaller than the number of unknown parameters, and therefore, it is only possible to provide a low-rank approximation of the posterior covariance, with the rank limited to at most N (particles) – 1. Variance collapse occurs when the actual rank falls below this theoretical limit. This limitation prevents us from fully capturing uncertainty across all directions, thus our analyses produce only relative—but still meaningful—uncertainty estimates.

The SVGD algorithm stands out due to its flexibility, enabling optimization problems to be solved through standard gradient descent for a certain number of particles, whilst introducing interparticle communication. However, it is crucial—yet laborious—to address the complexities of hyperparameter tuning. In our study, we experimented with two variants of SVGD (vanilla and annealed), GRF perturbations to build initial particles, different kernel functions (RBF and IMQ), different bandwidth selection strategies (median trick and constant value), a constant learning rate and varying particle counts. Moreover, our experimental evaluations were confined

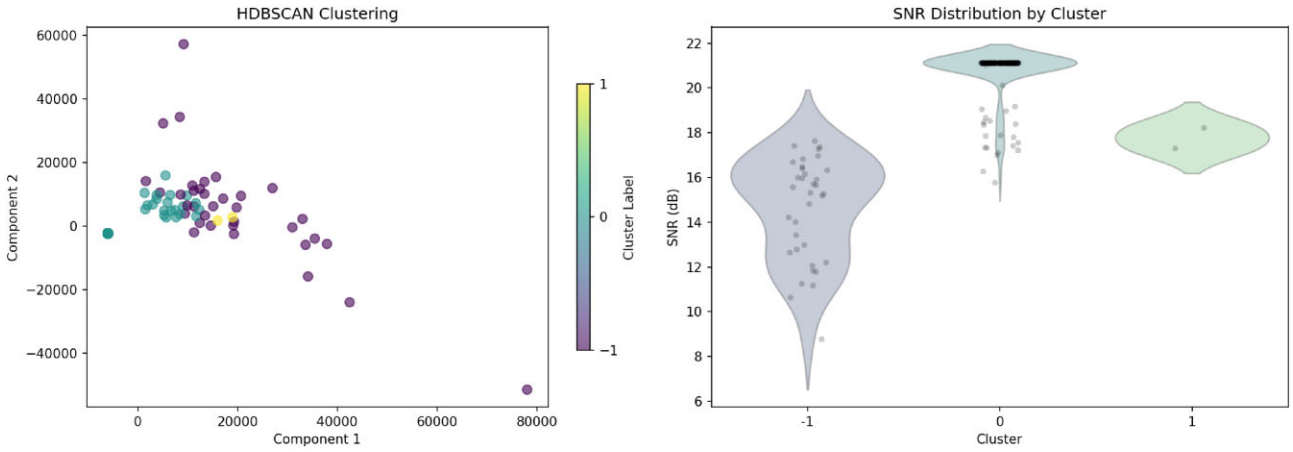

Figure 11. HDBSCAN clustering for the experiment using annealed SVGD with tanh formulation and RBF kernel with 200 particles. The left panel displays clusters obtained in high-dimensional data, plotted after dimensionality reduction to two components. The right panel shows the distribution of each cluster after computing the SNR.

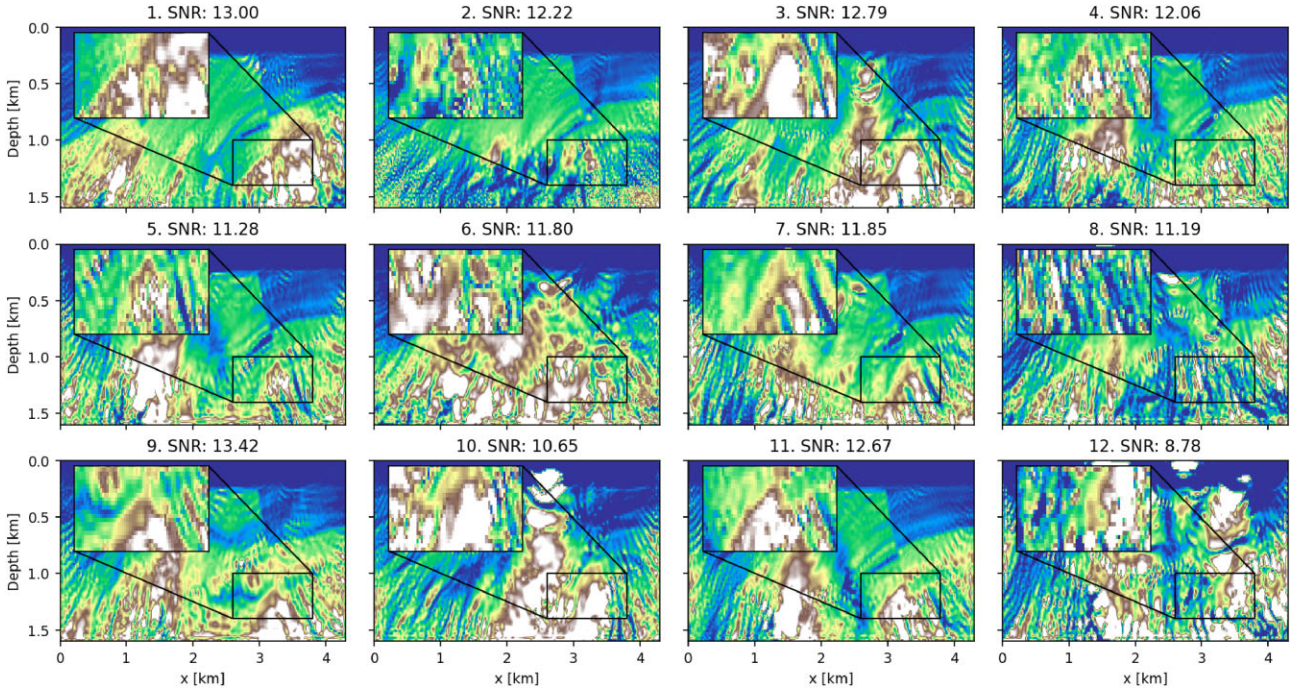

Figure 12. Noisy particles corresponding to cluster -1 for the experiment with annealed SVGD using the tanh formulation and RBF kernel with 200 particles.

to single-scale and multiscale scenarios and only assessed the uncertainties associated with the data misfit term (modelling operator). This methodology has the potential to produce nuanced standard deviation maps of velocities, though the integration of more informative prior information remains a subject for future exploration.

Our primary motivation was to apply annealed SVGD to mitigate mode- and variance-collapse issues that affect the vanilla SVGD approach. In the single-scale scenario, our findings reveal that annealed SVGD with the tanh formulation provide better model estimates (i.e. higher SNR, lower data misfit and overall more meaningful uncertainty estimates) than the vanilla formulation. Higher standard deviation values are associated with high-velocity layers and areas of poor coverage due to the limited acquisition geometry. The vanilla approach yields smaller confidence intervals, sometimes following a different trend than the true model. The annealed approach with the tanh formulation captures a broader range of uncertainty and increases the standard deviation values throughout the model. The single-scale outcomes present undesirable artefacts in the shallower parts, which should theoretically be well-illuminated areas. Under such conditions, there is no guarantee that all particles will converge to the neighbourhood of a unique local minimum. When projecting the particles onto a subspace spanned by $N-1$ dimensions or less, we observe that the annealed version recovers higher variances per component, confirming the capability of mitigating variance collapse. Visualization of individual particles shows some particles fitting the data term but not representing the subsurface.

In the multiscale scenario, both convergence and exploration are significantly enhanced. The mean and standard deviation estimates are improved and the shallower artefacts observed in the single-scale

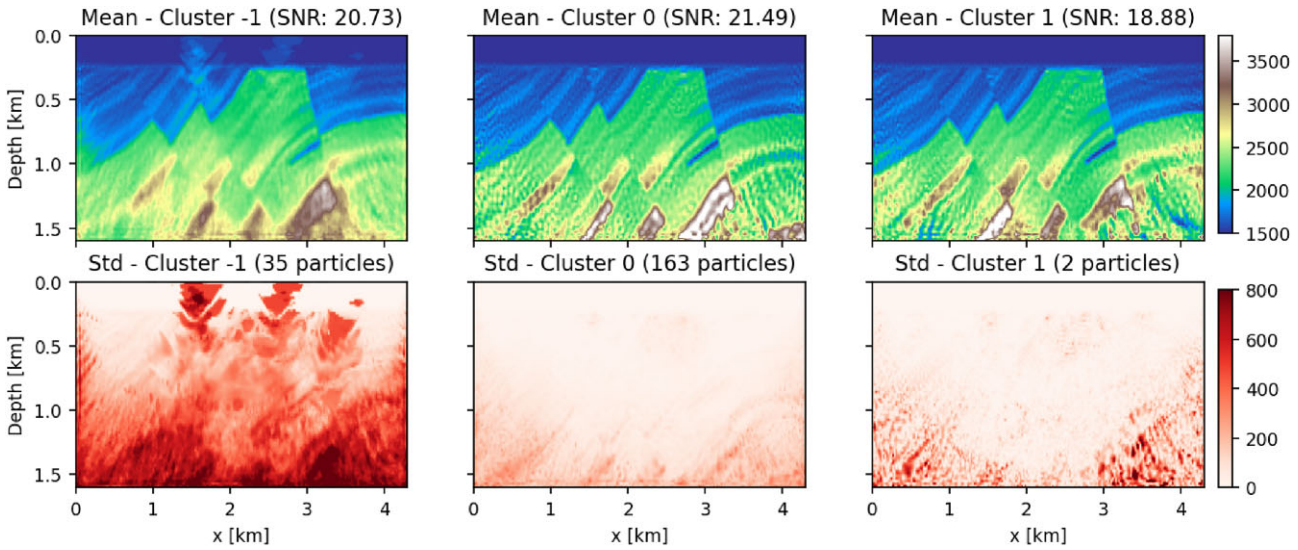

Figure 13. Cluster statistics for the experiment with annealed SVGD using the tanh formulation and RBF kernel with 200 particles.

scenario are absent. This suggests that a sequential approach from low to high frequencies may mitigate, though not entirely eliminate, the mode-collapse issue. For what concerns the variance-collapse issue, the annealed version with the tanh formulation yields more reasonable standard deviation maps, as indicated by a more significant number of components explaining the majority of the data variance.

Given the nonlinearity and high-dimensionality of the problem, it is challenging to ensure that all particles belong to a single mode. Therefore, we propose to perform clustering analysis, regardless of the scenario. We opted for HDBSCAN due to its applicability in high dimensions. HDBSCAN enables the easy discovery of a noisy cluster composed of non-geological particles. Independent statistical analysis per cluster in the single-scale scenario produced mean and standard deviation maps which confirmed that unexpected features of the std maps are generated by the members of the noisy cluster. The primary goal of using clustering analysis on the final set of particles as a *post hoc* technique is to quickly identify different modes and geologically meaningful particles. We prioritize this approach over incorporating a prior term, which can be mathematically challenging to formulate for filtering out geologically implausible features. A promising future direction could involve training a generative adversarial network (GAN) or variational autoencoder (VAE) to incorporate more informative priors into the optimization process (Corrales *et al.* 2022).

While annealed SVGD does not entirely solve the problem of variance collapse, it seems a promising method for FWI, where gradient computations are usually computationally expensive. Furthermore, when annealed SVGD is combined with a multiscale FWI scenario, reasonable estimates could be obtained with fewer gradient evaluations (iterations). This study highlights the potential of advanced SVGD methods to improve the reliability of FWI.

One promising direction for future work involves applying SVGD in a reduced or projected space (Chen & Ghattas 2020; Liu *et al.* 2022) to decrease the number of unknowns in our inverse problem and assess the impact on computational efficiency, convergence rates and quality of the inversion results. This approach would align with the ideal conditions for SVGD, that is, when the number of particles equals or exceeds the number of unknowns. For example, a potential solution could be to run the SVGD FWI algorithms in a discrete cosine transform (DCT) compressed domain, following the works of Aleardi (2021) and Berti *et al.* (2024a, b), who found that model compression through DCT effectively reduces the ill-conditioning of the problem. Also, we could use variational autoencoders to compress both the model and data and apply SVGD in the compressed (latent) domain (Sun & Williamson 2024). This exploration holds the potential to significantly advance the application of SVGD in FWI, which could pave the way for more accurate and computationally feasible seismic imaging techniques.

5 CONCLUSIONS

This study demonstrates that annealed SVGD can significantly improve convergence and performance compared to vanilla SVGD in FWI applications, in scenarios where the number of particles is much smaller than the number of unknown parameters. Specifically, the annealed SVGD with the tanh formulation enhances the accuracy of mean estimates, leading to higher SNR, lower data misfit and more reasonable standard deviation maps and thereby mitigating—but not eliminating—the variance-collapse issue. Additionally, applying multiscale FWI with SVGD yields better mean and standard deviation estimates compared to the single-frequency scenario, which is further enhanced by using annealed SVGD. Also, combining multiscale FWI with annealed SVGD yields superior performance. The use of PCA to explain variance by component provides valuable insights into the behaviour of the samples during the optimization process. This method provides additional understanding of how different components contribute to the overall variance to give a clearer picture of the sample distribution and convergence patterns. Finally, given the inherent complexity and high dimensionality of the problem, it is crucial to account for the possibility of multiple modes in the solution space. Consequently, we use HDBSCAN to analyse the final set of particles and identify clusters which have non-geological features, which could be excluded from the final statistics, to compensate for the absence of a comprehensive prior PDF. These strategies collectively offer a more robust and insightful approach to uncertainty analysis in FWI as they enhance the reliability of the results, provide a deeper understanding of the subsurface and ultimately aid more informed decision-making in industrial applications.

ACKNOWLEDGMENTS

This research was supported by King Abdullah University of Science and Technology (KAUST) and the DeepWave consortium. MC gratefully acknowledges TotalEnergies for the opportunity to conduct an internship in Pau, France and SB extends thanks to KAUST for hosting an internship in Saudi Arabia. MC and SB equally contributed for this work. Similarly, the authors extend their gratitude to the editors, as well as to the reviewers for their insightful feedback and suggestions.

DATA AVAILABILITY

The Marmousi data set utilized in this study is publicly accessible through the SEG Wiki page, available at https://wiki.seg.org/wiki/Open_data. The framework and experiments conducted in this research will be available on GitHub after publication at https://github.com/DeepWave-KAUST/AnnealedSVGD_FWI.

REFERENCES

Aleardi, M., 2021. A gradient-based Markov chain Monte Carlo algorithm for elastic pre-stack inversion with data and model space reduction, *Geophys. Prospect.,* **69**(5), 926–948.

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*.International Conference on Learning Representations.

Berti, S., Aleardi, M. & Stucchi, E., 2024a. A computationally efficient Bayesian approach to full-waveform inversion, *Geophys. Prospect.,* **72**(2), 580–603.

Berti, S., Aleardi, M. & Stucchi, E., 2024b. A Bayesian approach to elastic full-waveform inversion: application to two synthetic near surface models, *Bull. Geophys. Oceanogr.,* **65**(2), 291–308.

Berti, Sean., Ravasi, Matteo., Aleardi, Mattia. & Stucchi, Eusebio., 2025. Bayesian full waveform inversion of surface waves with annealed stein variational gradient descent, *Geophysical Journal International,* **241** (1), 641.

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

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

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

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

Brougois, A., Bourget, M., Lailly, P., Poulet, M., Ricarte, P. & Versteeg, R., 1990, *Marmousi, Model and Data,* European Association of Geoscientists & Engineers.

Bunks, C., Saleck, F., Zaleski, S. & Chavent, G., 1995. Multiscale seismic waveform inversion, *Geophysics,* **60**(5), 1457–1473.

Campello, R. J. G.B., Moulavi, D. & Sander, J., 2013. Density-based clustering based on hierarchical density estimates, in *Advances in Knowledge Discovery and Data Mining,* pp. 160–172, Springer.

Chen, P. & Ghattas, O., 2020. Projected stein variational gradient descent, in *Advances in Neural Information Processing Systems,* Vol. **33,** pp. 1947–1958, Curran Associates, Inc.

Corrales, M., Izzatullah, M., Ravasi, M. & Hoteit, H., 2022. *Bayesian RockAVO: Direct petrophysical inversion with hierarchical conditional GANs.SEG International Exposition and Annual Meeting*. https://doi.org/10.1190/image2022-3745255.1.

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

D'Angelo, F. & Fortuin, V., 2021. *Annealed Stein Variational Gradient Descent*. arXiv preprint. https://doi.org/10.48550/arXiv.2101.09815.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. *et al.*, 1996. A density-based algorithm for discovering clusters in large spatial databases with noise, in *kdd,* Vol. **96,** pp. 226–231.

Fernandez-Martinez, J., Fernandez-Muniz, Z., Pallero, J. & Pedruelo-Gonzalez, L., 2013. From Bayes to Tarantola: new insights to understand uncertainty in inverse problems, *J. Appl. Geophys.,* **98,** 62–72.

Fichtner, A., Zunino, A. & Gebraad, L., 2019. Hamiltonian Monte Carlo solution of tomographic inverse problems, *J. geophys. Int.,* **216**(2), 1344–1363.

Gallego, V. & Insua, D.R., 2020. *Stochastic Gradient MCMC with Repulsive Forces*. arXiv preprint . https://doi.org/10.48550/arXiv.1812.00071.

Gauthier, O., Virieux, J. & Tarantola, A., 1986. Two dimensional nonlinear inversion of seismic waveform; numerical results, *Geophysics,* **51,** 1387–1403.

Gebraad, L., Boehm, C. & Fichtner, A., 2020. Bayesian elastic full-waveform inversion using Hamiltonian Monte Carlo, *J. geophys. Res.: Solid Earth,* **125**(3), e2019JB018428.

Gorham, J. & Mackey, L., 2017. Measuring Sample Quality with Kernels, in *Proceedings of the 34th International Conference on Machine Learning,* PMLR. pp. 1292–1301.

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

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components, *J. Educat. Psychol.,* **24**(6), 417–441.

Izzatullah, M., Alali, A., Ravasi, M. & Alkhalifah, T., 2024a. Physics-reliable frugal local uncertainty analysis for full waveform inversion, *Geophys. Prospect.,* **73,** 2718–2738.

Izzatullah, M., Alkhalifah, T., Romero, J., Corrales, M., Luiken, N. & Ravasi, M., 2024b. Posterior sampling with convolutional neural network-based plug-and-play regularization with applications to poststack seismic inversion, *Geophysics,* **89**(2), R137–R153.

Jordan, M.I., Ghahramani, Z., Jaakkola, T.S. & Saul, L.K., 1998. An Introduction to Variational Methods for Graphical Models, in *Learning in Graphical Models,* pp. 105–161, ed. Jordan, M.I., Springer Netherlands.

Kingma, D.P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I. & Welling, M., 2016. Improved Variational Inference with Inverse Autoregressive Flow, in *Advances in Neural Information Processing Systems,* Vol. **29,** Curran Associates, Inc.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A. & Blei, D.M., 2017. Automatic differentiation variational inference, *J. Mach. Learn. Res.,* **18**(14), 1–45.

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

Lailly, P. & Santosa, F., 1984. Migration methods: partial but efficient solutions to the seismic inverse problem, in *Inverse problems of acoustic and elastic waves,* Vol. **51,** SIAM Philadelphia, pp. 1387–1403.

Liu, M. & Grana, D., 2018. Stochastic nonlinear of seismic data for the estimation of petroelastic properties using the ensemble smoother and data reparameterization, *Geophysics,* **83,** M25–M39.

Liu, Q. & Peter, D., 2019. Square-root variable metric based elastic full-waveform inversion–part 2: uncertainty estimation, *J. geophys. Int.,* **218,** 1100–1120.

Liu, Q. & Wang, D., 2016. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm, in *Advances in Neural Information Processing Systems,* Vol. **29,** Curran Associates, Inc.

Liu, Q., Beller, S., Lei, W., Peter, D. & Tromp, J., 2021. Pre-conditioned bfgs-based uncertainty quantification in elastic full-waveform inversion, *J. geophys. Int.,* **228,** 796–815.

Liu, X., Zhu, H., Ton, J.-F., Wynne, G. & Duncan, A., 2022. *Grassmann Stein Variational Gradient Descent*. arXiv preprint, https://doi.org/10.48550/arXiv.2202.03297.

Lomas, A., Luo, S., Irakarama, M., Johnston, R., Vyas, M. & Shen, X., 2023. 3D Probabilistic full waveform inversion: application to Gulf of Mexico field data, in *84th EAGE Annual Conference & Exhibition,* pp. 1–5, European Association of Geoscientists & Engineers, Vienna, Austria.

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

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

Martin, J., Wilcox, L.C., Burstedde, C. & Ghattas, O., 2012. A Stochastic Newton MCMC Method for Large-Scale Statistical Inverse Problems with Application to Seismic Inversion, *SIAM J. Sci. Comput.,* **34**(3), A1460–A1487.

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

Mosegaard, K. & Tarantola, A., 2002. 16 - Probabilistic Approach to Inverse Problems, in *International Geophysics, Vol. 81 of International Handbook of Earthquake and Engineering Seismology, Part A,* pp. 237–265, eds Lee, W. H.K., Kanamori, H., Jennings, P.C. & Kisslinger, C., Academic Press.

Ravasi, M., 2023. *Multi-realization seismic data processing with deep variational preconditioners*. SEG International Exposition and Annual Meeting. https://doi.org/10.1190/image2023-3904212.1.

Rawlinson, N., Fichtner, A., Sambridge, M. & Young, M.K., 2014. Chapter one—seismic tomography and the assessment of uncertainty, *Elsevier,* **55,** 1–76.

Ray, A., Sekar, A., Hoversten, G.M. & Albertin, U., 2016. Frequency domain full waveform elastic inversion of marine seismic data from the Alba field using a Bayesian trans-dimensional algorithm, *J. geophys. Int.,* **205**(2), 915–937.

Richardson, A., 2023. *Deepwave*.

Sambridge, M. & Mosegaard, K., 2002. Monte Carlo methods in geophysical inverse problems, *Rev. Geophys.,* **40**(3), 3–1-3-29.

Sambridge, M., 2014. A Parallel Tempering algorithm for probabilistic sampling and multimodal optimization, *J. geophys. Int.,* **196**(1), 357–374.

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

Siahkoohi, A., Rizzuti, G., Louboutin, M., Witte, P.A. & Herrmann, F.J., 2021. *Preconditioned training of normalizing flows for variational inference in inverse problems*. arXiv preprint. https://doi.org/10.48550/arXiv.2101.03709.

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

Sun, Y. & Williamson, P., 2024. Invertible neural networks for uncertainty quantification in refraction tomography, *Leading Edge,* **43**(6), 358–366.

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

Tarantola, A., 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation,* Society for Industrial and Applied Mathematics.

Thurin, J., Brossier, R. & Metivier, L., 2019. Ensemble-based uncertainty estimation in full-waveform inversion, *J. geophys. Int.,* **219,** 1613–1635.

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

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

Wang, D., Tang, Z., Bajaj, C. & Liu, Q., 2019. Stein variational gradient descent with matrix-valued kernels, in *Advances in Neural Information Processing Systems,* Vol. **32,** Curran Associates, Inc.

Warner, M. & Guasch, L., 2014. *Adaptive Waveform Inversion—FWI Without Cycle Skipping—Theory,* Vol. **2014,** European Association of Geoscientists & Engineers, pp. 1–5.

Zhang, C., Bütepage, J., Kjellström, H. & Mandt, S., 2019. Advances in variational inference, *IEEE Trans. Pattern Anal. Mach. Intell.,* **41**(8), 2008–2026.

Zhang, X. & Curtis, A., 2020a. Seismic tomography using variational inference methods, *J. geophys. Res.: Solid Earth,* **125**(4), e2019JB018589.

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

Zhang, X., Lomas, A., Zhou, M., Zheng, Y. & Curtis, A., 2023. 3-D Bayesian variational full waveform inversion, *J. geophys. Int.,* **234**(1), 546–561.

Zhao, X., Curtis, A. & Zhang, X., 2022. Bayesian seismic tomography using normalizing flows, *J. geophys. Int.,* **228**(1), 213–239.

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.

Zhuo, J., Liu, C., Shi, J., Zhu, J., Chen, N. & Zhang, B., 2018. Message passing stein variational gradient descent, in *Proceedings of the 35th International Conference on Machine Learning,* pp. 6018–6027, PMLR, Stockholm, Sweden

APPENDIX A: SUPPLEMENTARY RESULTS FOR SINGLE-SCALE EXPERIMENTS

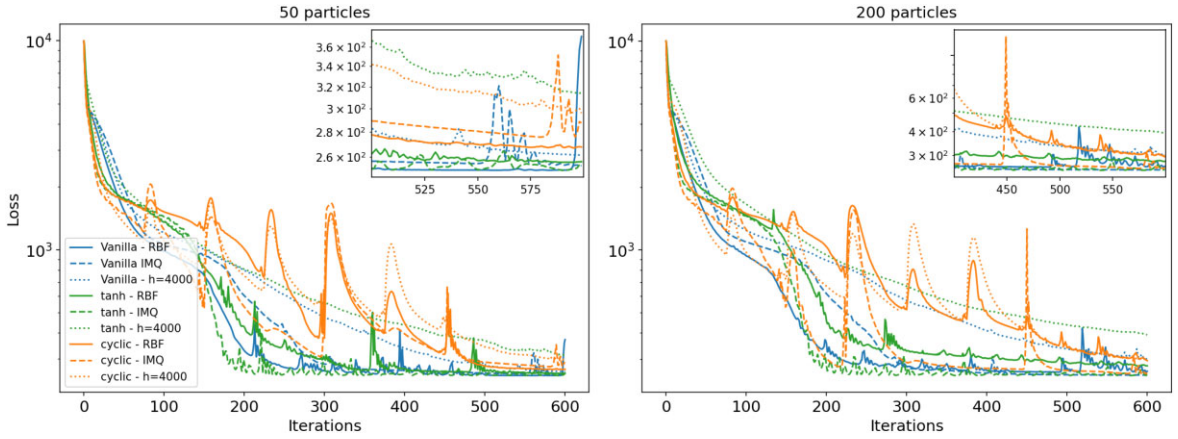

Figure A1. Data misfit for 50 particles (left) and 200 particles (right) across different experiments in the single-frequency scenario. The zoom window provides a clearer misfit comparison of the final 150 iterations.

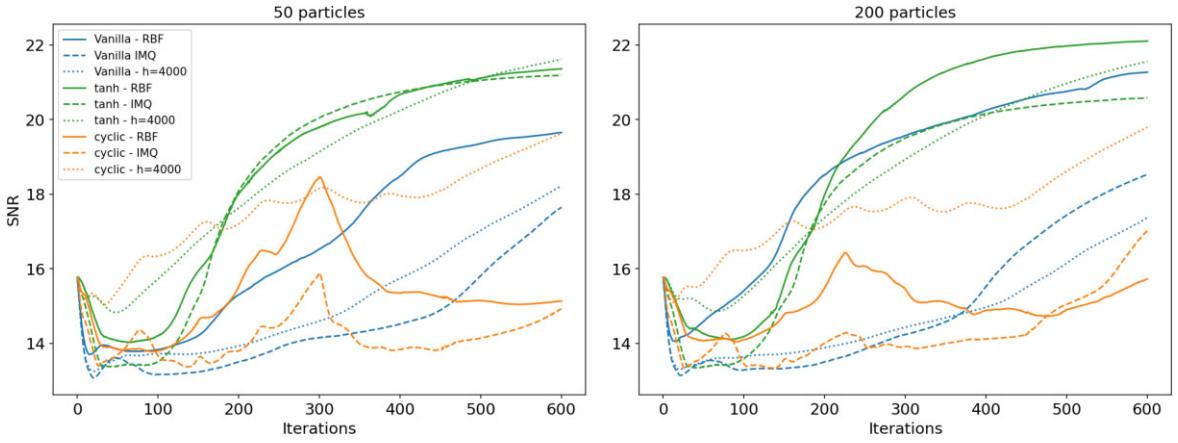

Figure A2. SNR for 50 particles (left) and 200 particles (right) across different experiments in the single-frequency scenario.

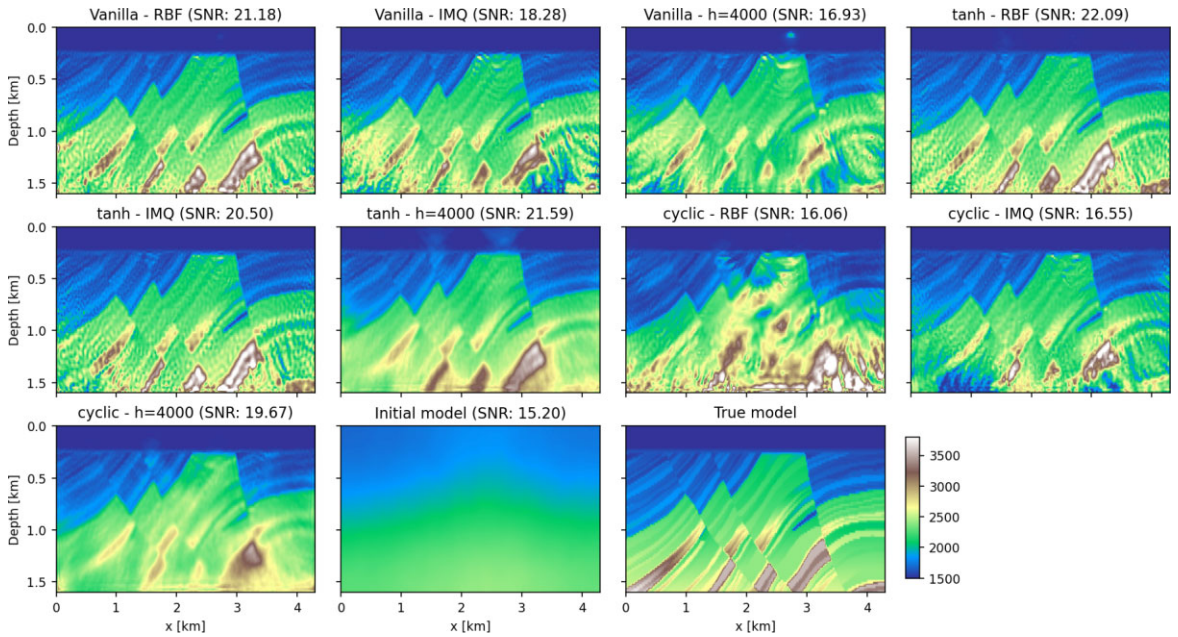

Figure A3. Comparison of the mean models from different experiments using 200 particles, highlighting various SVGD variants and hyperparameters after 600 iterations in the single-frequency scenario. The velocity values are expressed in $\mathrm{m\,s^{-1}}$.

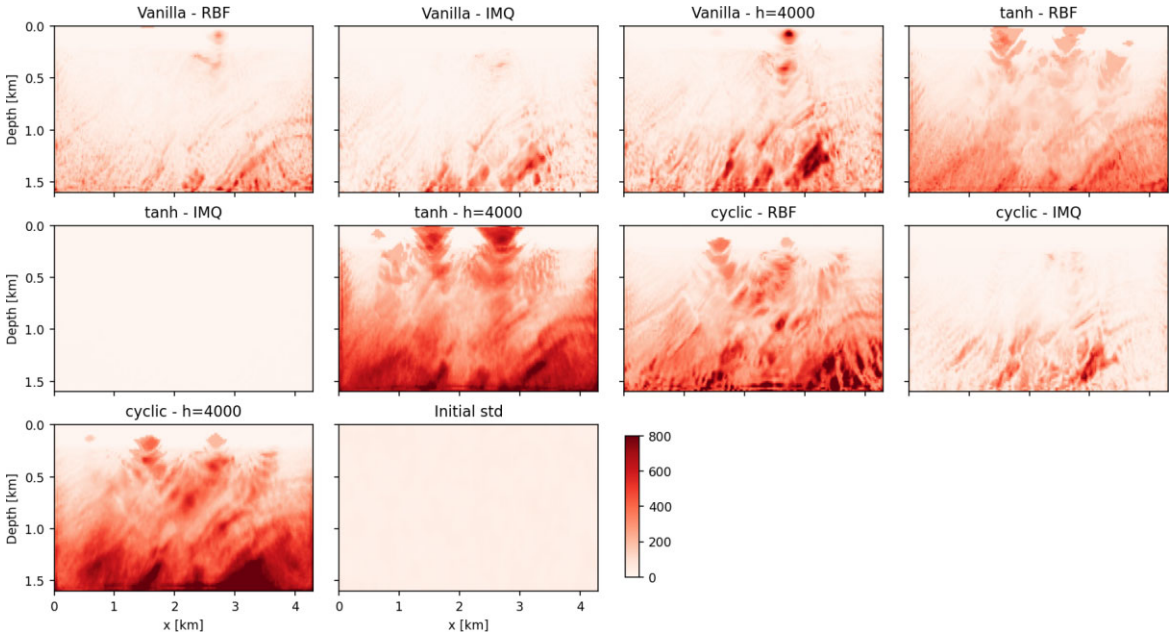

Figure A4. Comparison of standard deviation from different experiments using 200 particles, highlighting various SVGD variants and hyperparameters after 600 iterations in the single-frequency scenario. The velocity values are expressed in $\mathrm{m\,s^{-1}}$.

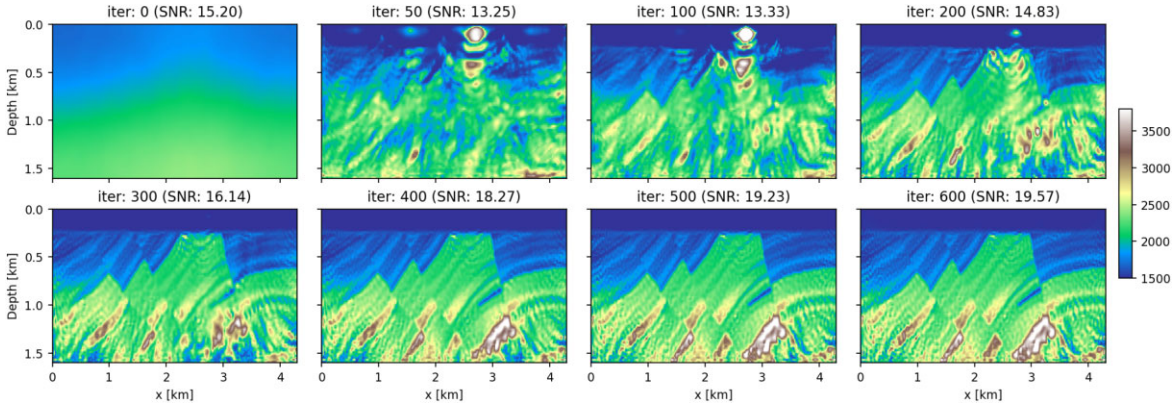

Figure A5. Evolution of the mean for the 200-particle experiment using Vanilla SVGD with RBF kernel.

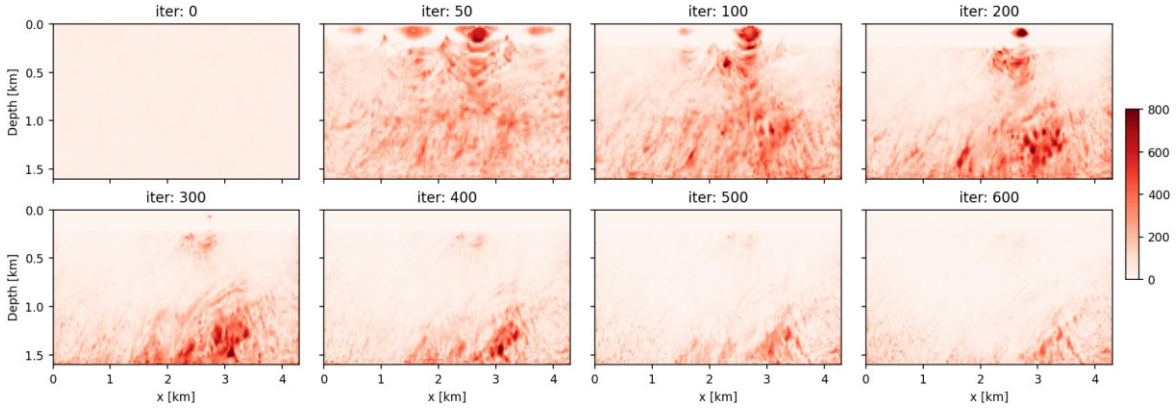

Figure A6. Evolution of the standard deviation for the 200-particle experiment using Vanilla SVGD with RBF kernel.

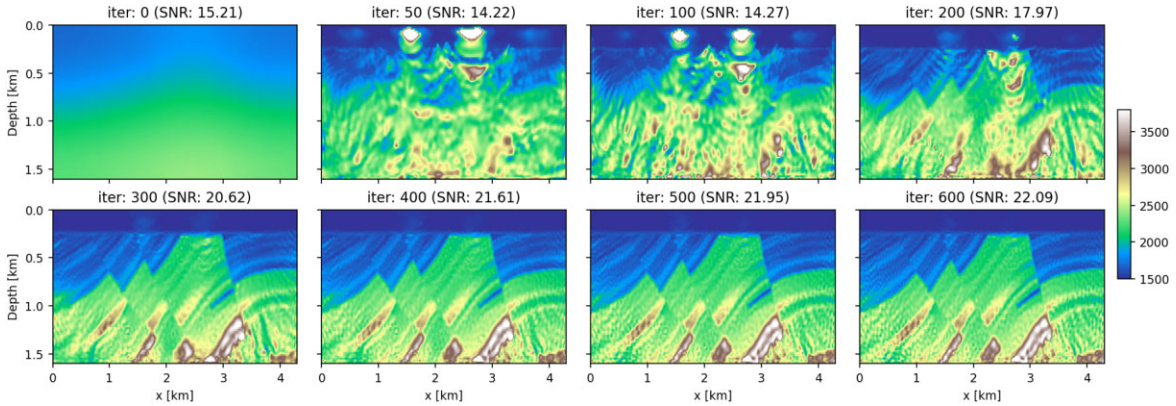

Figure A7. Evolution of the mean for the 200-particle experiment using annealed SVGD (tanh) with RBF kernel.

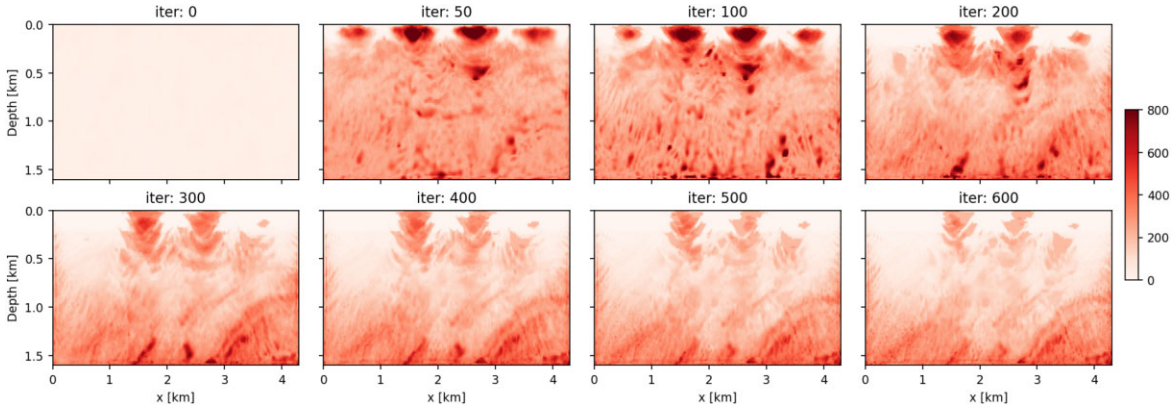

Figure A8. Evolution of the standard deviation for the 200-particle experiment using annealed SVGD (tanh) with RBF kernel.

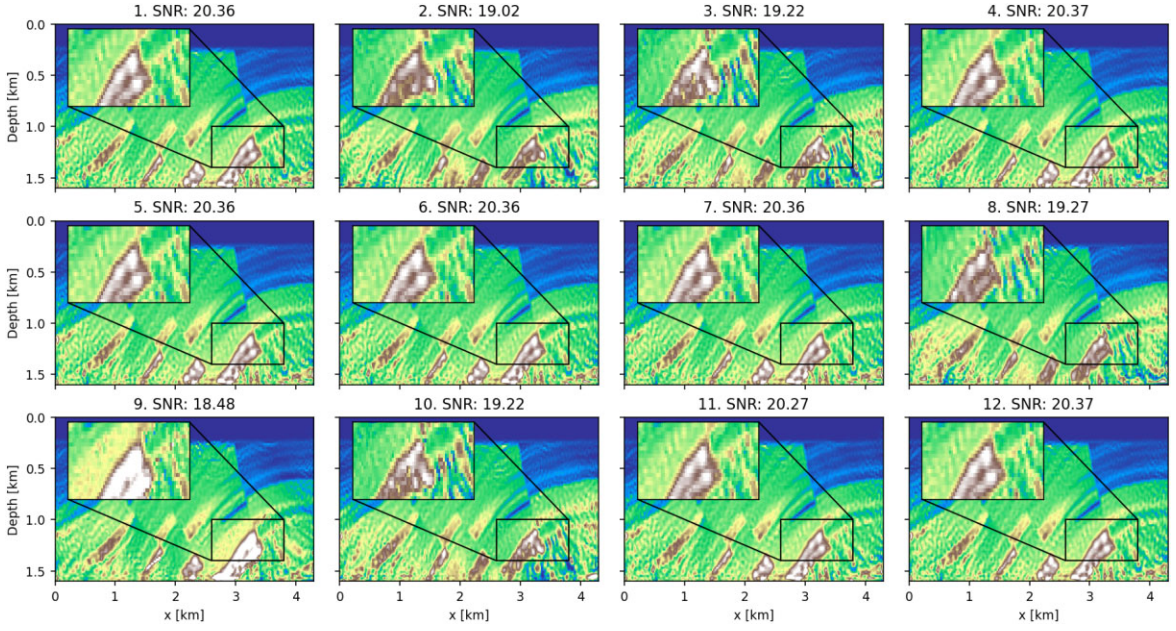

Figure A9. Single-frequency scenario: visualization of 12 particles from a 200-particle experiment using vanilla SVGD with RBF Kernel after 600 iterations.

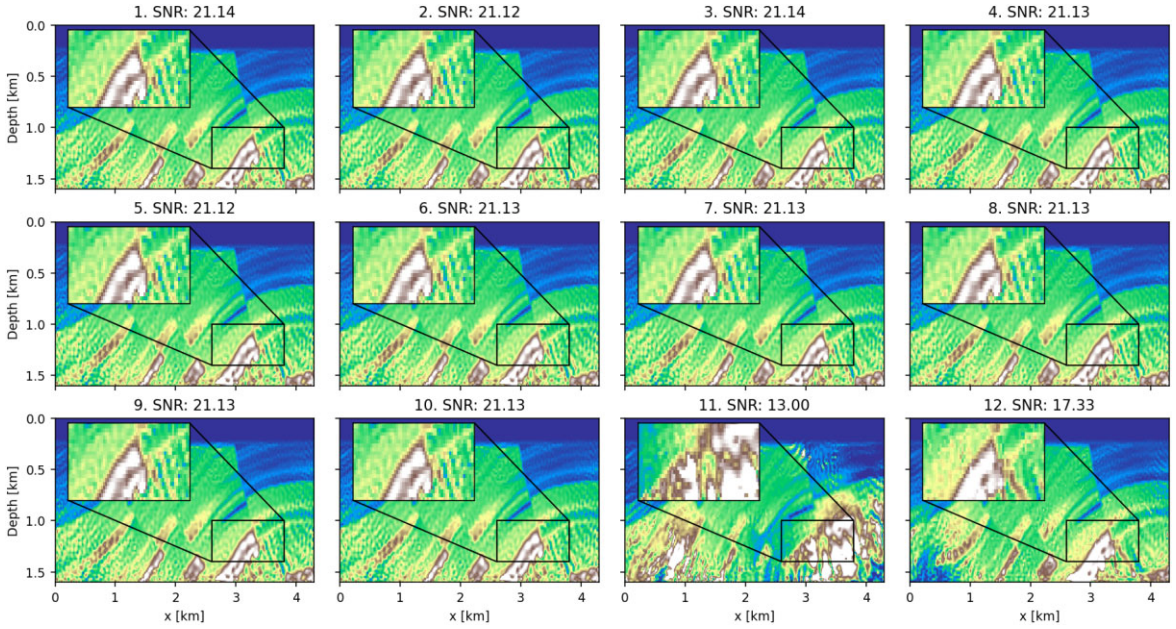

Figure A10. Single-frequency scenario: visualization of 12 particles from a 200-particle experiment using annealed SVGD (tanh) with RBF Kernel after 600 iterations.

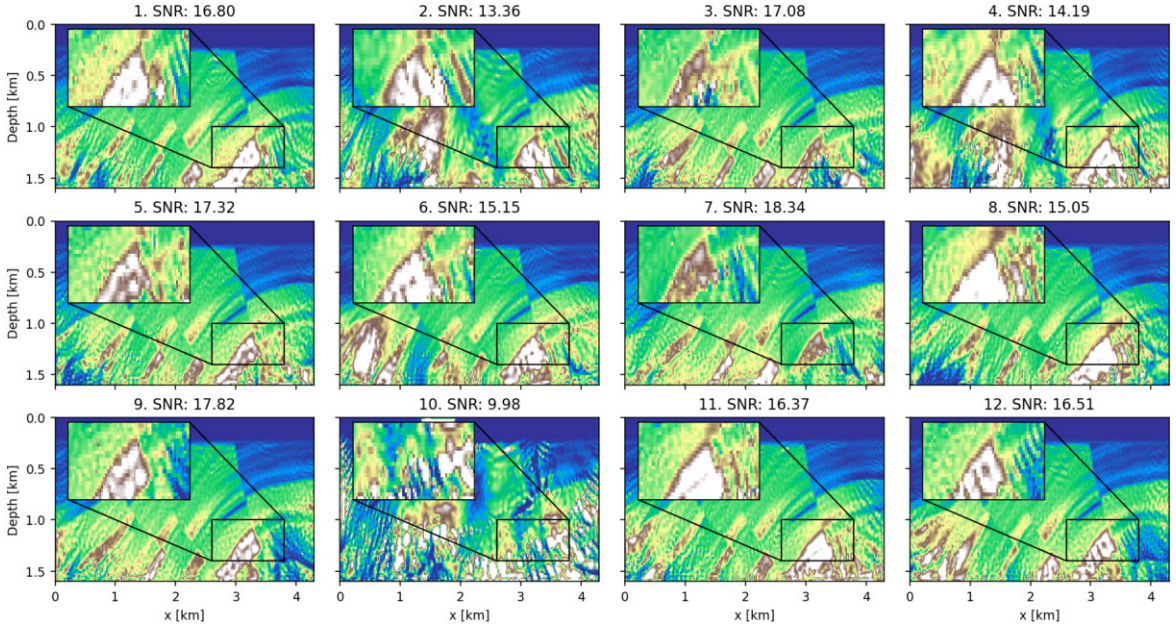

Figure A11. Single-frequency scenario: Visualization of 12 particles from a 200-particle experiment using annealed SVGD, (tanh) with RBF kernel and constant bandwidth ($h = 4000$) after 600 iterations.

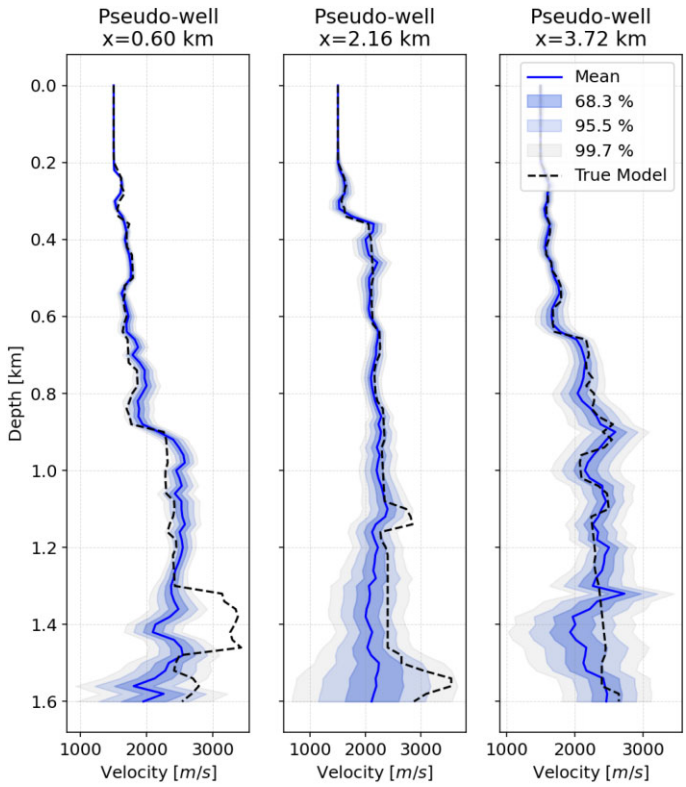

Figure A12. Single frequency scenario: pseudo-wells marginals from a 200-particle experiment using vanilla SVGD with RBF kernel and constant bandwidth ($h = 4000$) 200 particles.

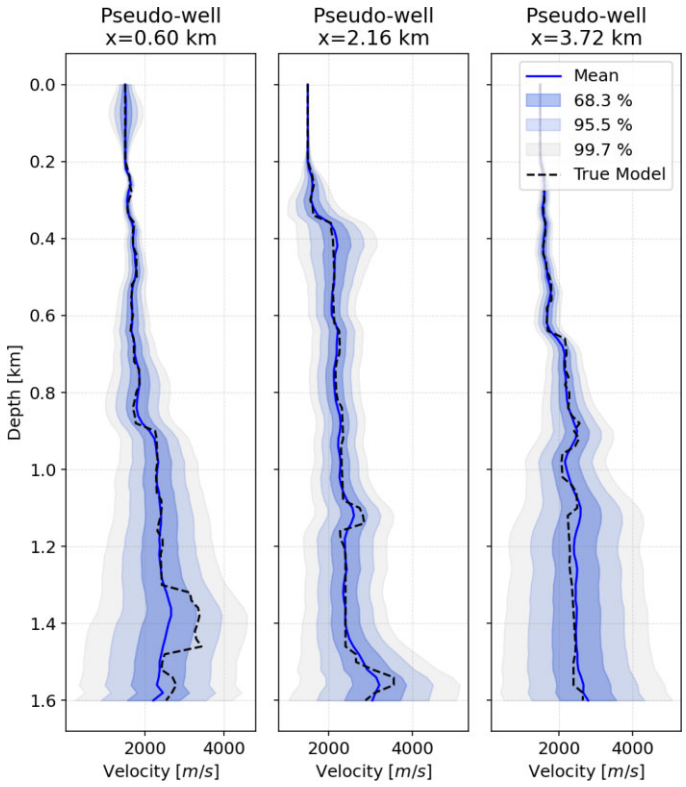

Figure A13. Single frequency scenario: pseudo-wells marginals from a 200-particle experiment using annealed SVGD (tanh) with RBF kernel and constant bandwidth ($h = 4000$) 200 particles.

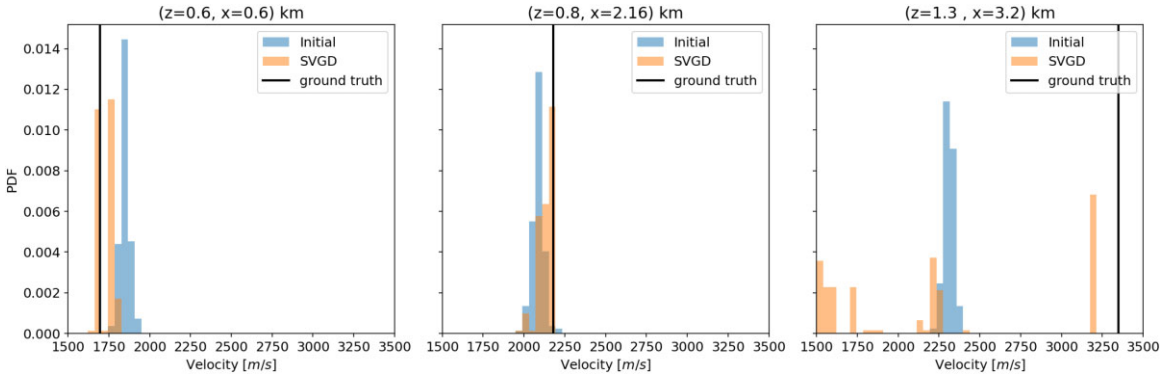

Figure A14. Single frequency scenario: pixels marginals for vanilla SVGD with RBF kernel and constant bandwidth ($h = 4000$), and 200 particles.

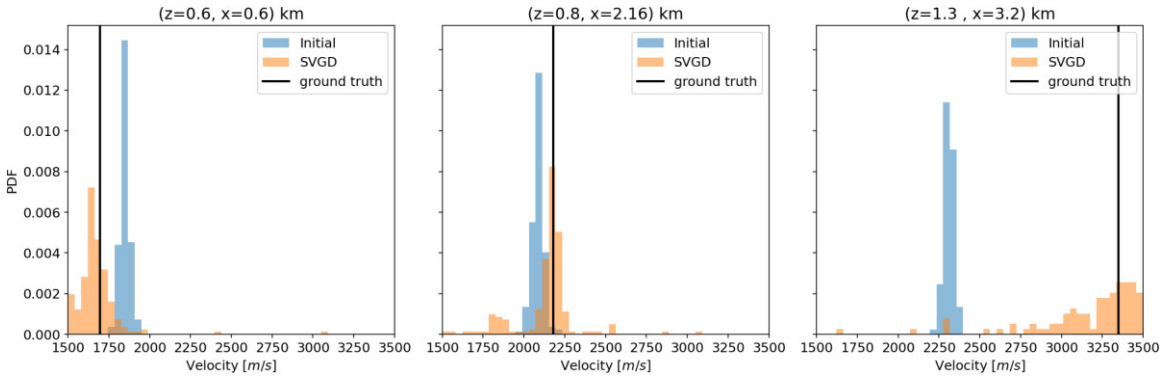

Figure A15. Single frequency scenario: pixels marginals for annealed SVGD (tanh) with RBF kernel and constant bandwidth ($h = 4000$), and 200 particles.

APPENDIX B: SUPPLEMENTARY RESULTS FOR MULTISCALE EXPERIMENTS

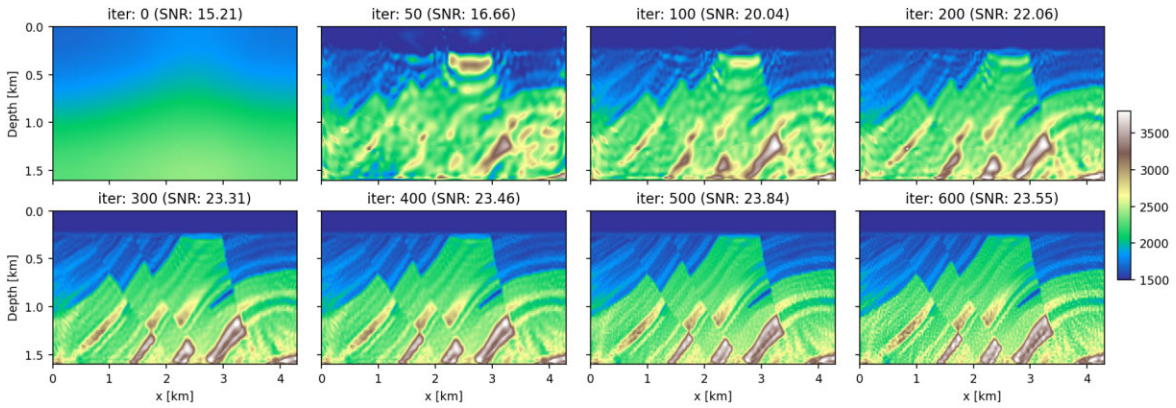

Figure B1. Multiscale scenario: Mean evolution for the 200-particle experiment using vanilla SVGD with RBF kernel.

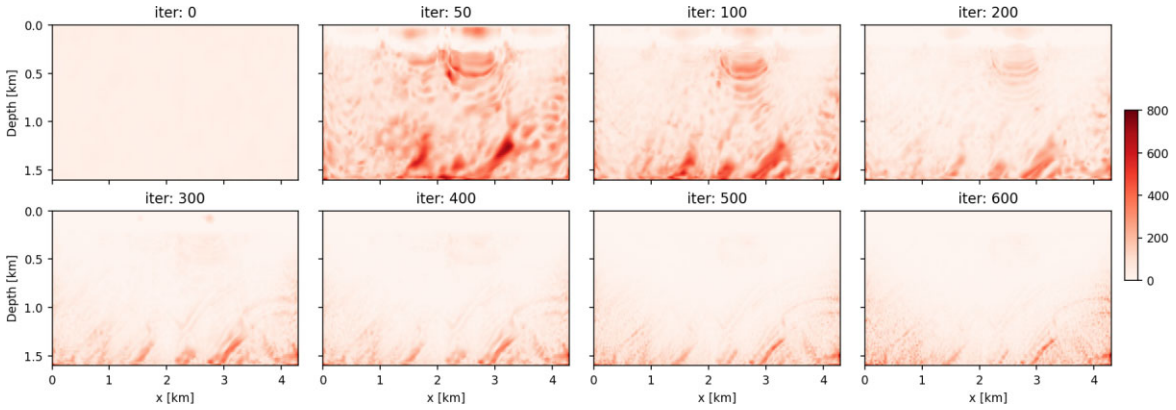

Figure B2. Multiscale scenario: Standard deviation evolution for the 200-particle experiment using vanilla SVGD with RBF kernel.

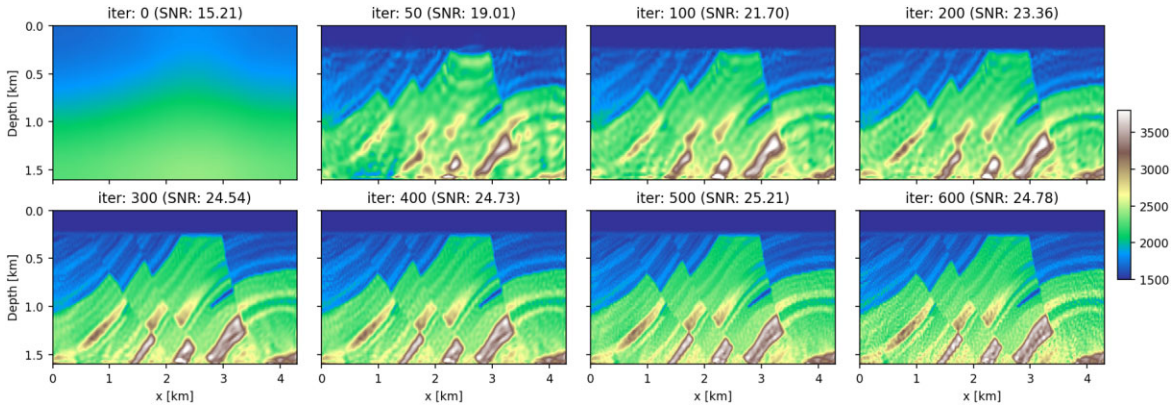

Figure B3. Multiscale scenario: Mean evolution for the 200-particle experiment using annealed SVGD (tanh) and RBF kernel.

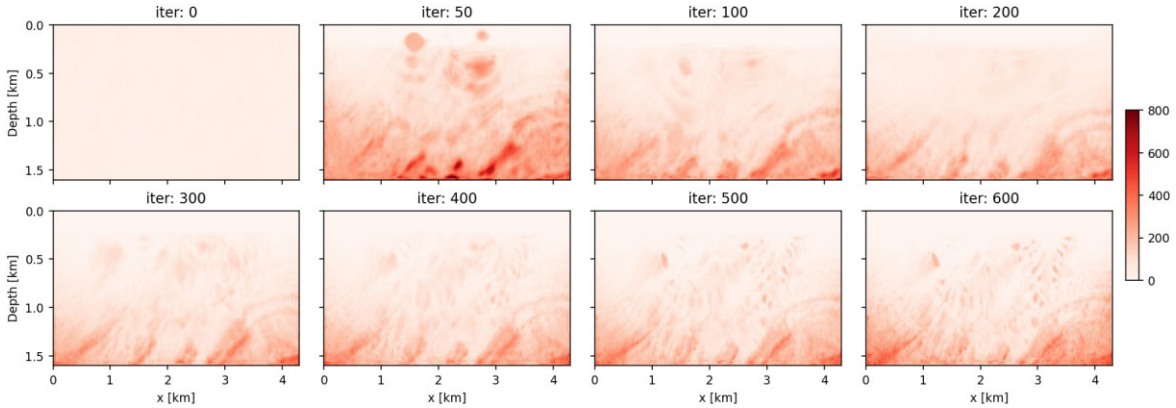

Figure B4. Multiscale scenario: Standard deviation evolution for the 200-particle experiment using annealed SVGD (tanh) and RBF kernel.

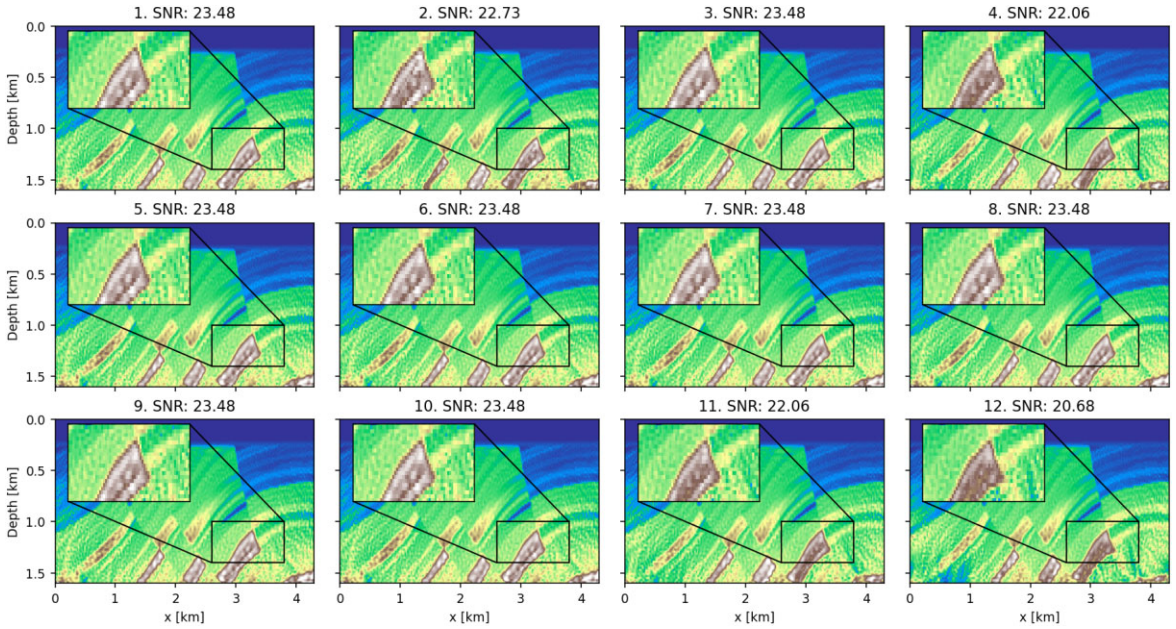

Figure B5. Multiscale scenario: visualization of 12 particles from a 200-particle experiment using vanilla SVGD with RBF Kernel after 600 iterations.

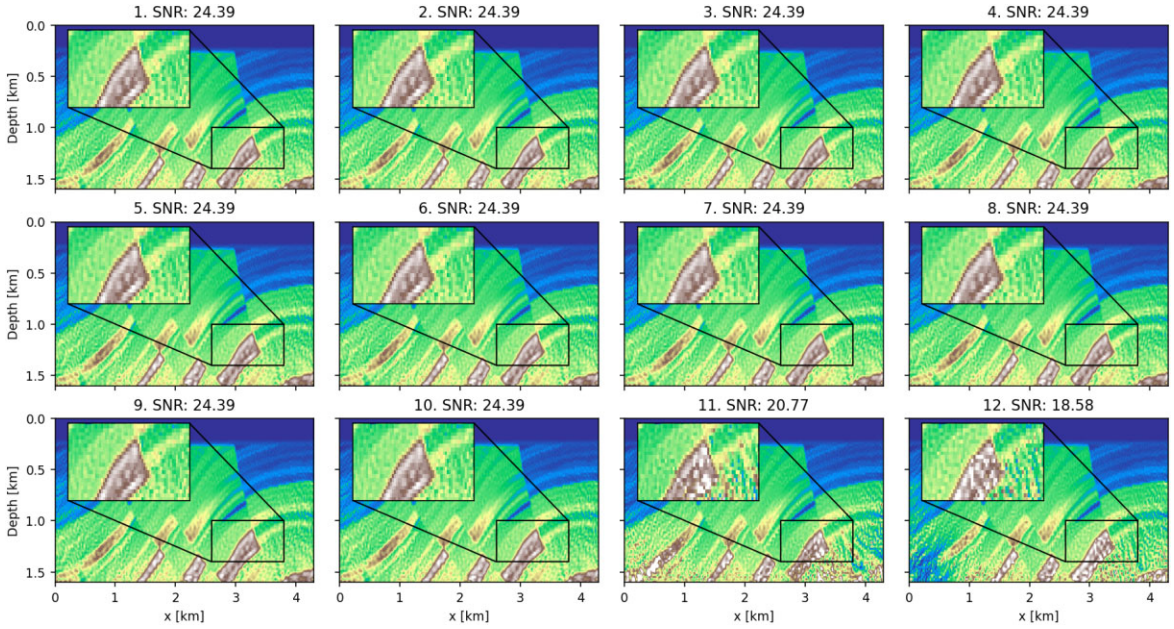

Figure B6. Multiscale scenario: visualization of 12 particles from a 200-particle experiment using annealed SVGD (tanh) with RBF Kernel after 600 iterations.

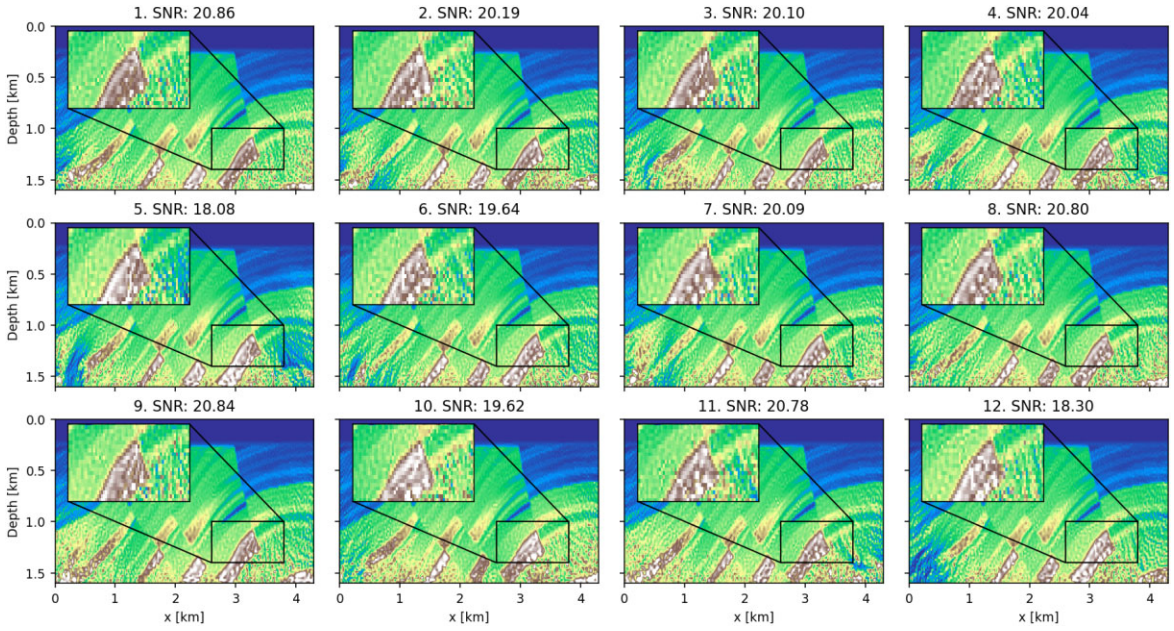

Figure B7. Multiscale scenario: visualization of 12 particles from a 200-particle experiment using annealed SVGD (tanh) with RBF kernel and constant bandwidth ($h = 2500$) after 600 iterations.

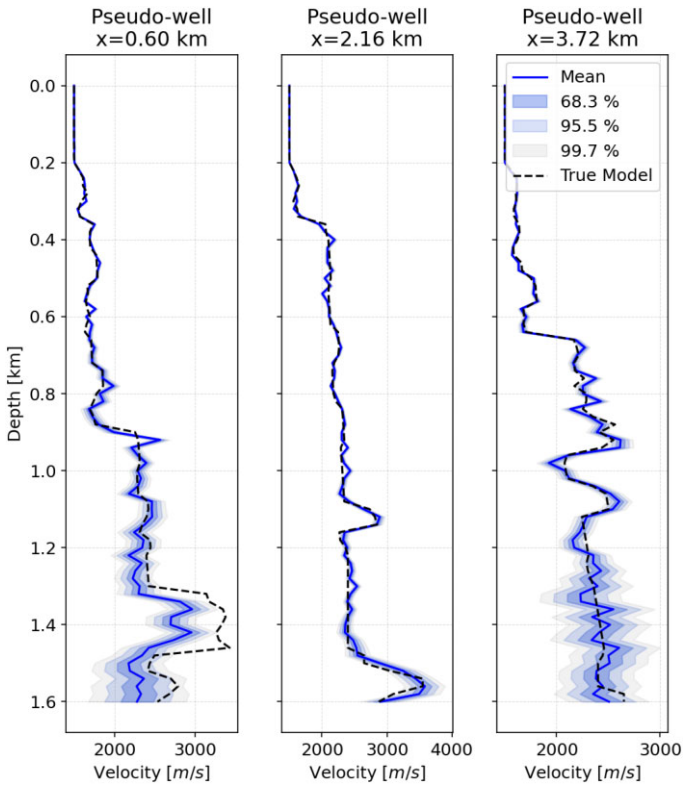

Figure B8. Multiscale scenario: pseudo-wells marginals from a 200-particle experiment using vanilla SVGD with RBF kernel.

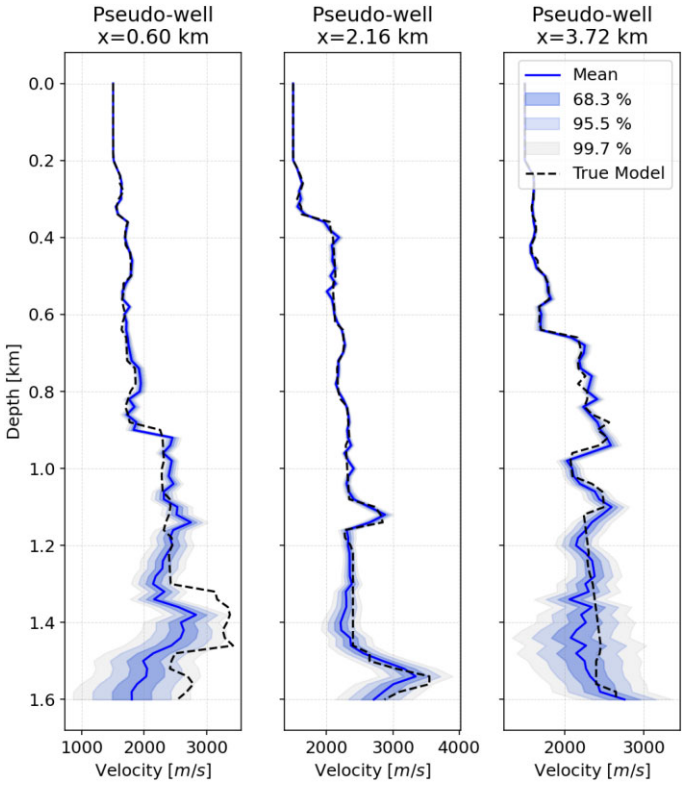

Figure B9. Multiscale scenario: pseudo-wells marginals from a 200-particle experiment using vanilla SVGD with RBF kernel and constant bandwidth ($h = 2500$).

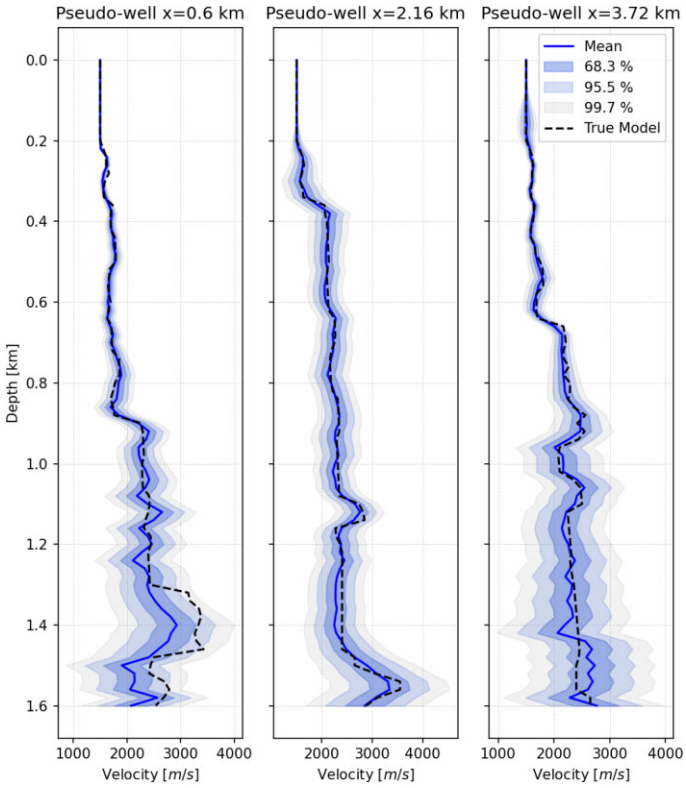

Figure B10. Multiscale scenario: pseudo-wells marginals from a 200-particle experiment using annealed SVGD (tanh) with RBF kernel.

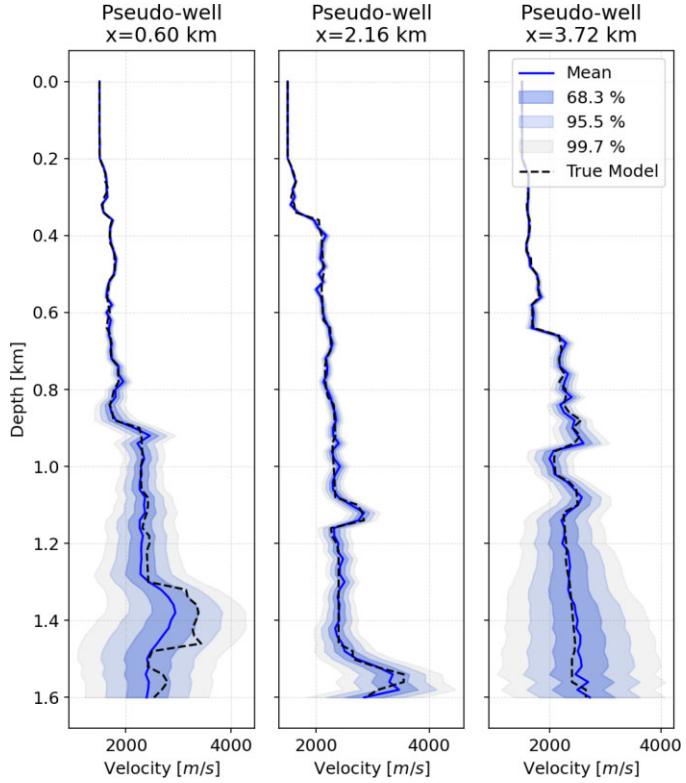

Figure B11. Multiscale scenario: pseudo-wells marginals from a 200-particle experiment using annealed SVGD (tanh) with RBF kernel and constant bandwidth ($h = 2500$).

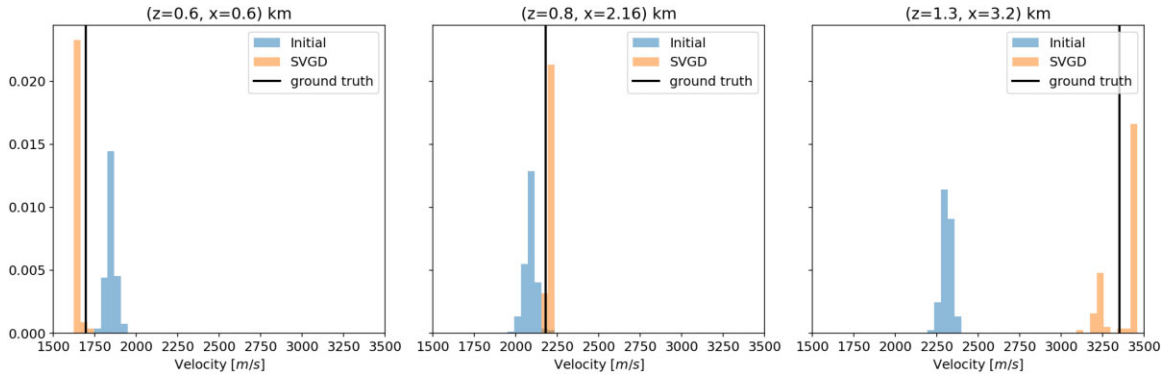

Figure B12. Multiscale scenario:, pixels marginals for vanilla SVGD with RBF kernel, 200 particles.

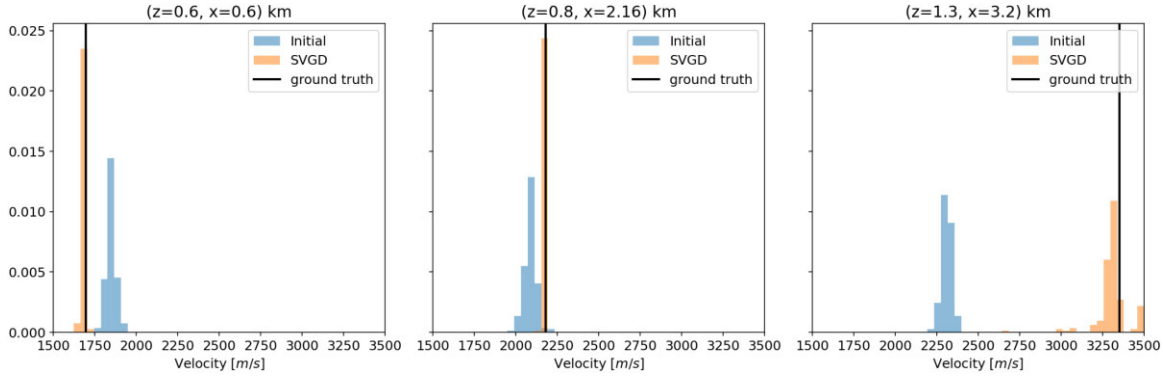

Figure B13. Multiscale scenario: pixels marginals for vanilla SVGD with RBF kernel and constant bandwidth ($h = 2500$), 200 particles.

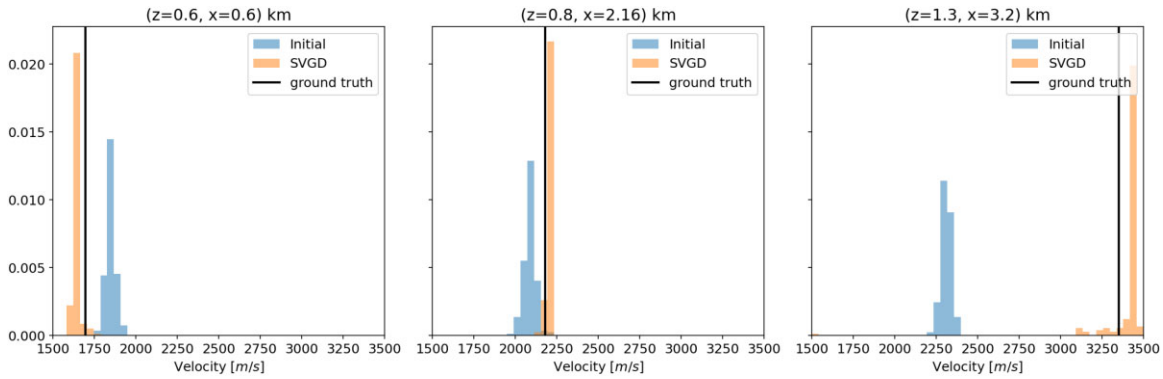

Figure B14. Multiscale scenario: pixels marginals for annealed SVGD (tanh) with RBF kernel, 200 particles.

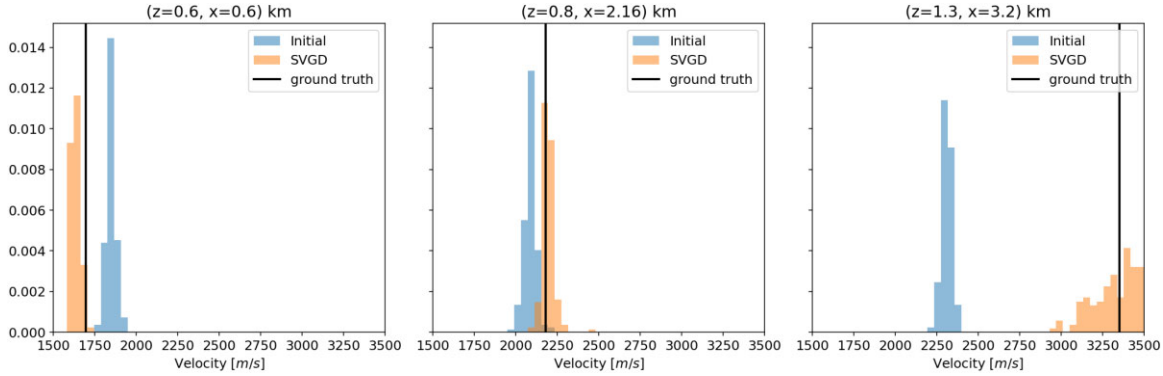

Figure B15. Multiscale scenario: pixels marginals for annealed SVGD (tanh) with RBF kernel and constant bandwidth ($h = 2500$), 200 particles.