When full waveform inversion meets gradient flows: an outlook on uncertainty quantification

*Philippe Marchner *, Univ. Grenoble Alpes, Romain Brossier, Univ. Grenoble Alpes, and Ludovic Métivier, Univ. Grenoble Alpes/CNRS*

Full Waveform Inversion (FWI) is a high-resolution seismic imaging method and a mature technology in exploration geophysics. The typical approach consists in recovering some subsurface mechanical properties, referred to as the model $m(x), x \in \mathbb{R}^d$, by minimizing a least-squares misfit (Virieux et al., 2014)

$$\min_{m\in\mathbb{R}^N} C(m) = \frac{1}{2}\|d_{\text{cal}}[m] - d_{\text{obs}}\|^2_{L^2}, \tag{1}$$

where N is the dimension of the model space, d_{obs} are observed seismic traces recorded at receivers (typically noisy), and $d_{\text{cal}}[m]$ are calculated traces obtained through a costly numerical resolution of an acoustic or elastodynamics wave equation. For 3D FWI, the dimension of the model space is very large, typically millions or billions. Quasi-Newton iterative methods are commonly used to solve this minimization problem due to their good trade-off between robustness, convergence speed, and computational cost. Most notably, the ℓ-BFGS method (Nocedal, 1980) has become the industry standard for FWI. The update takes the form, at the k-th iteration,

$$m_{k+1} = m_k - \tau_k Q(m_k)\nabla C(m_k), \tag{2}$$

where the preconditioner $Q(m_k)$ is constructed from the previous ℓ-gradients, and the step size τ_k is determined via a line-search strategy. It is well known that the minimization problem (1) is ill-posed and nonlinear, with a large effective null-space. As a result, the obtained subsurface property lacks interpretability. The Bayesian approach to inverse problems, promoted in the 80's by Tarantola and Valette (1982), provides a probabilistic view on the uncertainty quantification problem, which has been widely adopted in many fields of applied sciences. The solution to the inverse problem consists in finding the *probability distribution* of all possible models, once all known data and prior knowledge have been incorporated. It is obtained from the Bayes theorem (Kaipio and Somersalo, 2005)

$$\pi_{\text{post}}(m) := \pi(m|d_{obs}) = \frac{\pi(d_{\text{obs}}|m)\pi_{\text{prior}}(m)}{\pi(d_{\text{obs}})}, \quad \pi(d_{\text{obs}}) \neq 0, \tag{3}$$

given a prior model distribution $\pi_{\text{prior}}(m)$, and a likelihood distribution that is related to the least-squares misfit through $\pi(d_{\text{obs}}|m) \propto \exp(-C(m))$ when additive Gaussian noise are assumed in the observations. Despite progress in computational seismology, machine learning, and the availability of computing power, a direct application of the Bayes formula remains intractable, because it involves a high-dimensional integral over the entire model space. A more reasonable goal is to find a set of highly probable solutions drawn from the posterior π_{post}, a task known in computational statistics as *sampling*. Arguably one of the oldest sampling strategy is to build a Markov Chain (MCMC) through the celebrated Metropolis-Hastings algorithm (Hastings, 1970). In FWI, advanced MCMC strategies such as HMC (Gebraad et al., 2020; Sen and Biswas, 2017), RJMCMC (Bodin and Sambridge, 2009), Stochastic-Newton MCMC (Martin et al., 2012), or MALA (Izzatullah et al., 2021) have been employed. Although these methods can achieve fast mixing rates, they remain challenging to apply when a single likelihood evaluation involves solving a wave equation with $\mathcal{O}(10^9)$ degrees of freedom. At the other end of the spectrum, Gaussian posterior approximations are computationally affordable, and enable a local sensitivity analysis around the maximum a posteriori (MAP) estimate (Fichtner and Trampert, 2011; Bui-Thanh et al., 2013). They provide a useful baseline for uncertainty quantification but fail to capture non-gaussian posteriors, which are expected in highly nonlinear problems like FWI. In the late 2000's, machine learning brought new perspectives to Bayesian inference. A major shift in perspective was the willingness to trade exact posterior representation for computational efficiency, giving rise to the field of *variational inference* (VI). Examples include Gaussian VI (Ranganath et al., 2014) and particle-based methods such as Stein Variational Gradient Descent (SVGD, Liu and Wang (2016)), which have both been applied to FWI (Zhang et al., 2023). These approaches provide scalable uncertainty estimates at a fraction of the cost of MCMC; nevertheless, it is still not well understood how accurately they approximate the posterior distribution. The same applies to normalizing flows (Siahkoohi et al., 2021; Yin et al., 2024), which learn a mapping to transform the prior into the posterior, through the composition of simple invertible functions. In parallel, the field of data assimilation has pursued similar ideas, giving rise to a class of methods known as Sequential Monte Carlo (SMC) or particle filters. A notable success in this area is the Ensemble Kalman Filter or EnKF (Evensen, 2003; Thurin et al., 2019), which offers a computationally efficient, though approximate, alternative to particle filters. Inspired by these recent advances in uncertainty quantification for FWI, we emphasize a general methodology to better understand the modern computational Bayesian approaches to sampling, leveraging the theory of gradient flows in the space of probability measures (Santambrogio, 2017; Trillos et al., 2023). We then discuss how data assimilation techniques, such as the EnKF, can help mitigate the curse of dimensionality on a challenging FWI scenario with field data, following the work of Hoffmann et al. (2024). We conclude by outlining perspectives and directions for future research.

MODERN ADVANCES ON SAMPLING USING GRADIENT FLOWS ON PROBABILITIES

The goal of this section is to propose a probabilistic interpretation of FWI. Recent advances in the theory of optimal transport (Chewi et al., 2024; Santambrogio, 2015) have brought new insights into Bayesian inference through the framework of gradient flows in the space of probability measures. To relate this perspective to a deterministic FWI update, we note that the it-

erative scheme in Eq. (2) is an explicit Euler discretization of the ODE

$$\frac{dm}{dt} = -\nabla_Q C(m), \quad m(0) = m_0, \tag{4}$$

as the step size $\tau_k \to 0$, which we refer to as a gradient flow in the Euclidean space $\mathbb{R}^N$. It is driven by the velocity vector field $\mathbf{v} = \nabla_Q C = Q\nabla C$. We can interpret the FWI minimization as the choice of three components (Chen et al., 2023):

1. a misfit function;
2. a metric to define the gradient;
3. a suitable numerical integration scheme.

Rather than inverting for a single subsurface property, we give a probabilistic analogue, by treating the model $M = m(x)$ as a realization of a random variable $M \sim \mu$ governed by the probability distribution μ. To facilitate the exposition, we assume that all probability measures are smooth probability densities. Instead of minimizing a least-squares misfit, we define a statistical discrepancy. We choose the Kullback-Leibler (KL) divergence, which measures the relative information between μ and the posterior distribution π from Eq. (3). The optimization problem becomes

$$\min_{\mu\in\mathscr{P}} \mathscr{C}(\mu), \quad \mathscr{C}(\mu) = \mathrm{KL}(\mu||\pi_{\text{post}}) := \int \mu \log\left(\frac{\mu}{\pi_{\text{post}}}\right) dm, \tag{5}$$

which is defined in the space of probability densities $\mathscr{P}$. The theory of optimal transport allows to endow the space $\mathscr{P}$ with a Riemannian geometry (Otto, 2001), thereby inducing a metric where we can assign an inner product and further define a gradient. Denoting by $\mathscr{M}(\mu)$ the (Riemannian) metric tensor, the gradient flow of the functional $\mathscr{C}$ formally evolves a curve of densities $t \to \mu_t$ in artificial time according to

$$\frac{\partial \mu_t}{\partial t} = -\nabla\mathscr{C}(\mu_t) = -\mathscr{M}(\mu_t)^{-1} \frac{\delta\mathscr{C}}{\delta\mu}\bigg|_{\mu=\mu_t}, \tag{6}$$

where $\frac{\delta\mathscr{C}}{\delta\mu}$ denotes the first variation (Fréchet derivative) of $\mathscr{C}$. Applied to the KL divergence, we find $\frac{\delta \mathrm{KL}}{\delta\mu} = \log\mu - \log\pi + 1$. A canonical choice for the metric is the 2-Wasserstein metric W_2, the transport distance that quantifies the minimal mean-squared displacement needed to move one density into another. Suppose the time-dependent density μ_t is carried from any prior π_{prior} towards the posterior π_{post} by a velocity field $\mathbf{v}_t$. Because probability mass is conserved, the evolution of μ_t must satisfy the continuity equation in the weak sense,

$$\int \left(\frac{\partial \mu_t}{\partial t} + \nabla\cdot(\mu_t \mathbf{v}_t)\right)\varphi = 0, \tag{7}$$

for any smooth test function φ of compact support. Among all velocity fields satisfying the continuity equation, the one that realizes the 2-Wasserstein distance minimizes the kinetic energy, and therefore is a potential field $\mathbf{v}_t = \nabla\phi$. This is the Benamou and Brenier (2000) dynamic formulation of optimal transport. It yields the inverse metric operator,

$$\mathscr{M}(\mu)^{-1}\phi = -\nabla(\mu\nabla\phi),$$

acting on a scalar field ϕ. Plugging the KL first variation into the gradient flow formula yields the Wasserstein gradient flow,

$$\frac{\partial \mu_t}{\partial t} = \nabla\cdot\left(\mu_t \nabla\log\left(\frac{\mu_t}{\pi_{\text{post}}}\right)\right) \tag{8}$$

which is the evolution equation pioneered by Jordan et al. (1998). This suggests that, to transform the prior to the posterior, one may follow the direction of steepest descent of the KL divergence with respect to the W_2 metric. Writing the posterior as $\pi_{\text{post}}(m) \propto e^{-C(m)}$, where $C(m)$, previously introduced as the data misfit, may now include additional regularization terms accounting for prior information, the corresponding Wasserstein gradient flow is equivalently a Fokker-Planck equation

$$\frac{\partial \mu_t}{\partial t} = \Delta\mu_t + \nabla\cdot(\mu_t\nabla C).$$

It reveals that the flow is a combination of: i) a drift term $\nabla\cdot(\mu_t\nabla C)$ that transports the density towards small values of the misfit C (i.e., the high-probability regions of the posterior), and ii) a diffusion term $\Delta\mu_t$ that spreads the mass out, allowing the flow to explore the posterior rather than collapsing to a single point. Under a convexity assumption on the (possibly regularized) misfit C, it can be shown that the KL divergence decreases monotonically along the flow, and that the evolving distribution μ_t converges to the posterior. This highlights a key strength of the framework, namely its analogy to convergence guarantees in convex optimization.

FROM GRADIENT FLOWS TO SAMPLING METHODS

We now relate the previous development to the uncertainty quantification methods used in FWI.

Markov Chain Monte Carlo and Langevin dynamics

The Fokker-Planck PDE admits the overdamped Langevin stochastic differential equation (SDE) representation

$$dM_t = -\nabla C(M_t)dt + \sqrt{2}dB_t, \quad M_t \sim \mu_t, \tag{9}$$

where B_t is the standard Brownian motion. If particles follows this SDE, they will converge in law to π_{post} as $t \to \infty$. A straightforward time discretization is the Euler-Maruyama scheme $M_{k+1} = M_k - \tau\nabla C(M_k) + \sqrt{2\tau}\xi_k, \xi_k \sim \mathcal{N}(0, I)$, which gives the unadjusted Langevin algorithm. Adding a Metropolis-Hastings correction yields MALA (Roberts and Tweedie, 1996), and removes the bias introduced by the time step. Preconditioning, using curvature information as in stochastic-Newton MCMC, or introducing momentum (underdamped Langevin / HMC) corresponds respectively to changing the metric or augmenting the dynamics with auxiliary variables. A preconditioned dynamics such as $dM_t = -Q\nabla C(M_t)dt + \sqrt{2Q}dB_t$, with Q a positive-definite matrix, can substantially accelerate mixing rates of MCMC. More generally, there is an entire family of SDEs whose marginal laws satisfy Fokker-Planck equations and inherit a gradient-flow structure (Ma et al., 2015).

Lagrangian view and SVGD

While the Langevin dynamics adopt an Eulerian point of view, the Wasserstein gradient flow admits the Lagrangian particle

representation with instantaneous velocity

$$\frac{dM_t}{dt} = \mathbf{v}_t(M_t) = -\nabla \log\left(\frac{\mu_t}{\pi_{\text{post}}}\right) = -\nabla C - \nabla \log \mu_t, \quad (10)$$

This representation is appealing because it describes a deterministic particle motion. However, it cannot be implemented directly, becuase the density μ_t is unknown. SVGD addresses this problem by restricting the admissible velocity fields to reproducing kernel Hilbert space (RKHS), i.e. a kernelized subspace of $L^2(\mu_t)$. The velocity field $\mathbf{v}_t$ is projected onto this subspace as

$$\hat{\mathbf{v}}_t(\cdot) = -\int \kappa(m,\cdot)\nabla \log\left(\frac{\mu_t(m)}{\pi_{\text{post}}(m)}\right)\mu_t(m)dm, \quad (11)$$

and after integrating by parts simplifies to $\mathbb{E}_{\mu_t}(\nabla\kappa - \kappa\nabla C)$, which can be estimated with Monte Carlo samples. The performance of SVGD critically depends on the kernel choice κ and lacks convergence guarantees in the large particle limit. In addition, kernel interactions can become ineffective in high-dimensional settings, leading to variance collapse.

Normalizing flows

Normalizing flows (NF) aim to transform a simple reference distribution, here the prior, into the target posterior via an invertible map T_θ parameterized by neural networks, such that $\pi_{\text{prior}} = (T_\theta)_\# \pi_{\text{post}}$, where $(T_\theta)_\#$ denotes the push-forward operation. In practice, T_θ is constructed as a composition of multiple local transformations $T_\theta = T_{\theta_K} \circ ... \circ T_{\theta_1}$, each of which can be interpreted as integrating a velocity field that satisfies the continuity equation over a discrete time step. From the lens of Wasserstein gradient flow, NFs learn a finite-dimensional approximation of the transport map induced by the steepest descent direction of the KL divergence. The resulting transformations are not necessarily the optimal maps associated with the dynamic optimal transport formulation. Recent developments have introduced normalizing flows in which a neural network is trained to approximate the implicit time-stepping of the JKO scheme (Mokrov et al., 2021). This construction establishes a closer geometric link between NFs and the variational structure of Bayesian inference, though, to the best of our knowledge has not yet been applied in the context of FWI.

Gaussian variational inference and EnKF

A natural simplification of the Wasserstein gradient flow is to restrict the space of admissible densities to a parametric family. Assuming a Gaussian family $\mu_t \sim \mathcal{N}(m_t, \Sigma_t)$, we can project the infinite-dimensional gradient flow onto the finite-dimensional manifold of Gaussian measures (Lambert et al., 2022). By taking the moments of Eq. (8), one obtains a closed system of ODEs governing the evolution of the mean and covariance:

$$\frac{dm_t}{dt} = -\mathbb{E}_{\mu_t}[\nabla C] \quad (12)$$

$$\frac{d\Sigma_t}{dt} = 2I - \Sigma_t \mathbb{E}_{\mu_t}[\nabla^2 C] - \mathbb{E}_{\mu_t}[\nabla^2 C]\Sigma_t, \quad (13)$$

which differs from the black-box variational inference (BBVI) where gradients are taken in the Euclidean parameter space of (m_t, Σ_t). By selecting a suitable quadrature rule for the Gaussian expectations and a time discretization scheme, we can compute the best Gaussian approximation to the posterior with respect to the KL divergence. However, note the covariance update involves $\mathcal{O}(N^2)$, which motivates low-rank or sparse approximations of the covariance. A sparse approximation was considered in the PSVI method (Zhao and Curtis, 2024), allowing FWI uncertainty quantification in 3D.

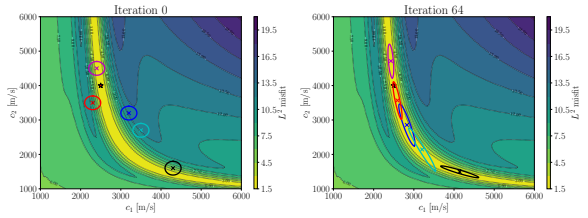

Figure 1: Evolution of a mixture of 5 Gaussians, represented in colors by their means and covariances, for a toy 2-parameters FWI problem at initialization (left) and after 64 iterations (right). The L^2-misfit is plot on the background, and shows a heavy-tailed nullspace, corresponding to models with equal travel times from the source to the receiver. The "true" model, $c^* = (2500, 4000)$ m/s, is depicted by the red star.

The Ensemble Kalman Filter (EnKF) also belongs to this class of Gaussian-approximation methods. Informally, the EnKF update can be viewed as a Gauss-Newton preconditioned gradient update for the mean, using the mean point quadrature rule (Chada et al., 2021)

$$\frac{dm_t}{dt} = -\mathbb{E}_{\mu_t}[Q\nabla C] \approx K(d_{\text{cal}}[m_t] - d_{\text{obs}}), \quad (14)$$

where K is the Kalman gain. In addition, the covariance is not updated explicitly; instead, it is estimated empirically from the ensemble, yielding a low-rank approximation of Σ_t. Because the EnKF is not designed as a sampling method, the ensemble tends to collapse over the iterations. In practice, localization and inflation techniques are used to counteract collapse and maintain diversity.

Towards new methods

In summary, many of the sampling methods used in FWI share a common ground as discretizations of the Wasserstein gradient flow related to the KL divergence. Thanks to this framework, we can also define new methods. For instance, we can enrich the Gaussian VI approximation by evolving a Gaussian mixture fully in parallel, following Lambert et al. (2022), in order to improve the posterior representation. This is particularly relevant when the effective nullspace is large, such as depicted in Fig. 1 on a toy FWI example. Although the framework is attractive, it is not straightforward to apply when the parameter space is high-dimensional. We show in the next section how the EnKF has achieved this on field data (Hoffmann et al., 2024).

LARGE-SCALE UNCERTAINTY QUANTIFICATION: AN EXAMPLE ON FIELD DATA

We propose an application of a UQ-FWI workflow on field data recorded by a 4-components OBC device. A total of 2048

receivers were deployed on the seabed along twelve cables, covering a surface of 145 km^2. 2048 reciprocal sources, located 5 m below the sea surface, are used for the inversion. The data is restricted to the frequency band 2.5-5 Hz. In this application we use only the pressure component and perform the inversion in the visco-acoustic VTI approximation. We rely on the finite-difference based full waveform modeling and inversion code `TOYxDAC_TIME` (Yang et al., 2018). We only invert for V_p and keep the other parameters as passive. The initial model has been obtained by reflection traveltime tomography. A source subsampling strategy has been applied into ensemble of shots in 16 batches.

The initial ensemble of models is designed by perturbing the initial model with Gaussian statistics, following the strategy described in Thurin et al. (2019). The initial covariance is chosen such that it matches the expected resolution of the FWI in the considered frequency band (Wu and Toksöz, 1987). The initial ensemble needs to be sufficiently rich to prevent collapse, while avoiding cycle-skipping of one of the ensemble member. This ensures that all models fall into the same local minimum valley. The FWI-EnKF scheme performs a dynamical application of the Bayes theorem, sequentially for the 16 batches of shots. Each application of the Bayes theorem consists of a forecast and an analysis step, repeated until all batches have been assimilated,

1. the forecast step applies a few BFGS iterations to each particle independently, which gives a new prior π_{prior}. Here we perform 3 BFGS iterations per batch,
2. the analysis step applies the EnKF update to the forecasted ensemble, which gives an intermediate posterior π_{post}. This requires the application of N_e forward modeling, where N_e is the ensemble size.

Reflecting on the previous section, we see that the empirical covariance is updated at each analysis step from the ensemble, but there is no diffusion term in the forecast nor the analysis steps to prevent the collapse of the ensemble. A strong advantage of this strategy is that both steps leverage quasi-Newton optimization. Once the EnKF-FWI strategy is performed, we can extract Gaussian statistics from the ensemble. Fig. 2 shows the mean and variance of the final ensemble for V_p obtained with three different ensemble sizes, for a cross-section at $x = 2.95$ km, as well as the distribution of the ensemble at a given point. We conclude that an ensemble size of $N_e = 50$ is sufficient to capture relevant information about uncertainty.

It remains to analyze the ability of the EnKF-FWI scheme to estimate the best Gaussian from the posterior with respect to the KL divergence, which is limited by i) the low-rank representation of the covariance, ii) the covariance collapse, and iii) the mean point quadrature rule. Based on our knowledge on gradient flows, we are in position to correct the covariance collapse of the EnKF by restoring a suitable diffusion effect inspired from the preconditioned Langevin SDE (Chada et al., 2021), at the cost of additional iterations. An example on a toy FWI example is shown in Fig. 3.

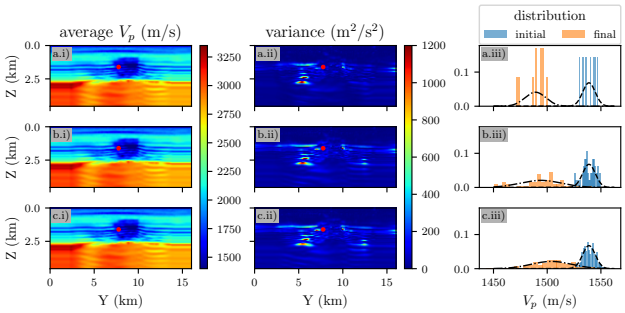

Figure 2: (a.i–c.i) Final mean V_p model obtained by EnKF-FWI, (a.ii–c.ii) final variance, (a.iii–c.iii) distribution of the models at a given point marked in red. The results are obtained with three ensemble sizes N_e: (a) $N_e = 10$, (b) $N_e = 50$ and (c) $N_e = 200$. $x = 2.95$ km.

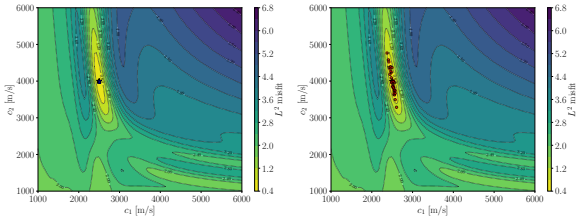

Figure 3: EnKF iterative scheme applied to a toy 2-parameters FWI problem, after 64 iterations (or analysis steps), with $N_e = 32$. While the EnKF has collapsed (left), we can correct the EnKF to ensure a good balance between drift and diffusion (right). The L^2-misfit is plot on the background. The "true" model, $c^* = (2500, 4000)$ m/s, is depicted by the red star.

OUTLOOKS AND FUTURE RESEARCH DIRECTIONS

Through dimensionality reduction, large-scale uncertainty quantification has recently become feasible for 3D FWI (Hoffmann et al., 2024; Zhang et al., 2023). Current strategies, however, still rely on a Gaussian approximation of the posterior. It is important to note that this Gaussian approximation is obtained through minimization of the KL divergence, rather than from a local estimation around the MAP. Future research directions include, but are not limited to, the following:

- developing sampling algorithms from the gradient flow perspective, which leverage second-order information and can move beyond the Gaussian representation,
- designing dimensionality reduction strategies that preserve the accuracy of the posterior representation,
- constructing accurate surrogate models to emulate wave propagation PDEs, enabling fast likelihood evaluation,
- understanding how sampling methods are affected by cycle-skipping.

These directions open exciting perspectives for uncertainty quantification in FWI in the years to come.

REFERENCES

Benamou, J.-D., and Y. Brenier, 2000, A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem: Numerische Mathematik, **84**, 375–393.

Bodin, T., and M. Sambridge, 2009, Seismic tomography with the reversible jump algorithm: Geophysical Journal International, **178**, 1411–1436.

Bui-Thanh, T., O. Ghattas, J. Martin, and G. Stadler, 2013, A computational framework for infinite-dimensional Bayesian inverse problems Part I: The linearized case, with application to global seismic inversion: SIAM Journal on Scientific Computing, **35**, A2494–A2523.

Chada, N. K., Y. Chen, and D. Sanz-Alonso, 2021, Iterative ensemble Kalman methods: A unified perspective with some new variants: Foundations of Data Science, **3**, 331–369.

Chen, Y., D. Z. Huang, J. Huang, S. Reich, and A. M. Stuart, 2023, Sampling via gradient flows in the space of probability measures: arXiv preprint arXiv:2310.03597.

Chewi, S., J. Niles-Weed, and P. Rigollet, 2024, Statistical optimal transport: arXiv preprint arXiv:2407.18163.

Evensen, G., 2003, The ensemble Kalman filter: Theoretical formulation and practical implementation: Ocean dynamics, **53**, 343–367.

Fichtner, A., and J. Trampert, 2011, Resolution analysis in full waveform inversion: Geophysical Journal International, **187**, 1604–1624.

Gebraad, L., C. Boehm, and A. Fichtner, 2020, Bayesian elastic full-waveform inversion using Hamiltonian Monte Carlo: Journal of Geophysical Research: Solid Earth, **125**, e2019JB018428.

Hastings, W. K., 1970, Monte Carlo sampling methods using Markov chains and their applications: Biometrika, **57**.

Hoffmann, A., R. Brossier, L. Métivier, and A. Tarayoun, 2024, Local uncertainty quantification for 3-D time-domain full-waveform inversion with ensemble Kalman filters: application to a North Sea OBC data set: Geophysical Journal International, **237**, 1353–1383.

Izzatullah, M., T. Van Leeuwen, and D. Peter, 2021, Bayesian seismic inversion: a fast sampling Langevin dynamics Markov chain Monte Carlo method: Geophysical Journal International, **227**, 1523–1553.

Jordan, R., D. Kinderlehrer, and F. Otto, 1998, The variational formulation of the Fokker–Planck equation: SIAM journal on mathematical analysis, **29**, 1–17.

Kaipio, J., and E. Somersalo, 2005, Statistical and computational inverse problems: Springer.

Lambert, M., S. Chewi, F. Bach, S. Bonnabel, and P. Rigollet, 2022, Variational inference via Wasserstein gradient flows: Advances in Neural Information Processing Systems, **35**, 14434–14447.

Liu, Q., and D. Wang, 2016, Stein variational gradient descent: A general purpose Bayesian inference algorithm: Advances in neural information processing systems, **29**.

Ma, Y.-A., T. Chen, and E. Fox, 2015, A complete recipe for stochastic gradient MCMC: Advances in neural information processing systems, **28**.

Martin, J., L. C. Wilcox, C. Burstedde, and O. Ghattas, 2012, A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion: SIAM Journal on Scientific Computing, **34**, A1460–A1487.

Mokrov, P., A. Korotin, L. Li, A. Genevay, J. M. Solomon, and E. Burnaev, 2021, Large-scale Wasserstein gradient flows: Advances in Neural Information Processing Systems, **34**, 15243–15256.

Nocedal, J., 1980, Updating quasi-Newton matrices with limited storage: Mathematics of computation, **35**, 773–782.

Otto, F., 2001, The geometry of dissipative evolution equations: the porous medium equation: Communications in Partial Differential Equations, **26**, 101–174.

Ranganath, R., S. Gerrish, and D. Blei, 2014, Black box variational inference: Artificial intelligence and statistics, PMLR, 814–822.

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

Santambrogio, F., 2015, Optimal transport for applied mathematicians: Birkäuser, Basel.

———, 2017, {Euclidean, metric, and Wasserstein} gradient flows: an overview: Bulletin of Mathematical Sciences, **7**, 87–154.

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

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

Tarantola, A., and B. Valette, 1982, Inverse problems= quest for information: Journal of geophysics, **50**, 159–170.

Thurin, J., R. Brossier, and L. Métivier, 2019, Ensemble-based uncertainty estimation in full waveform inversion: Geophysical Journal International, **219**, 1613–1635.

Trillos, N. G., B. Hosseini, and D. Sanz-Alonso, 2023, From optimization to sampling through gradient flows: Notices of the American Mathematical Society, **70**.

Virieux, J., A. Asnaashari, R. Brossier, L. Métivier, A. Ribodetti, and W. Zhou, 2014, 6. an introduction to full waveform inversion, *in* Encyclopedia of Exploration Geophysics: Society of Exploration Geophysicists, R1–40.

Wu, R.-S., and M. N. Toksöz, 1987, Diffraction tomography and multisource holography applied to seismic imaging: Geophysics, **52**, 11–25.

Yang, P., R. Brossier, L. Métivier, J. Virieux, and W. Zhou, 2018, A Time-Domain Preconditioned Truncated Newton Approach to Multiparameter Visco-acoustic Full Waveform Inversion: SIAM Journal on Scientific Computing, **40**, B1101–B1130.

Yin, Z., R. Orozco, M. Louboutin, and F. J. Herrmann, 2024, WISE: Full-waveform variational inference via subsurface extensions: Geophysics, **89**, A23–A28.

Zhang, X., A. Lomas, M. Zhou, Y. Zheng, and A. Curtis, 2023, 3-D Bayesian variational full waveform inversion: Geophysical Journal International, **234**, 546–561.

Zhao, X., and A. Curtis, 2024, Physically structured variational inference for Bayesian full waveform inversion: Journal of Geophysical Research: Solid Earth, **129**, e2024JB029557.

