Journal of Computational Physics 519 (2024) 113395

Contents lists available at ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

Randomized physics-informed machine learning for uncertainty quantification in high-dimensional inverse problems

Yifei Zong [a], David Barajas-Solano [b], Alexandre M. Tartakovsky [a,b,*]

[a] Department of Civil and Environmental Engineering, University of Illinois Urbana-Champaign, Urbana, IL 61801, United States of America
[b] Pacific Northwest National Laboratory, Richland, WA 99352, United States of America

ARTICLE INFO

Keywords:
Bayesian inference
Hamiltonian Monte Carlo
Inverse problems
Uncertainty quantification
Dimension reduction
Physics-informed machine learning

ABSTRACT

We propose the randomized physics-informed conditional Karhunen-Loève expansion (rPICKLE) method for uncertainty quantification in high-dimensional inverse problems. In rPICKLE, the states and parameters of the governing partial differential equation (PDE) are approximated via truncated conditional Karhunen-Loève expansions (cKLEs). Uncertainty in the inverse solution is quantified via the posterior distribution of cKLE coefficients formulated with independent standard normal priors and a likelihood containing PDE residuals evaluated over the computational domain. The maximum a posteriori (MAP) estimate of the cKLE coefficients is found by minimizing a loss function given (up to a constant) by the negative log posterior. The posterior is sampled by adding zero-mean Gaussian noises into the MAP loss function and minimizing the loss for different noise realizations. For linear and low-dimensional nonlinear problems, we show that the rPICKLE posterior converges to the true Bayesian posterior. For high-dimensional non-linear problems, we obtain rPICKLE posterior approximations with high log-predictive probability. For a low-dimensional problem, the traditional Hamiltonian Monte Carlo (HMC) and Stein Variational Gradient Descent (SVGD) methods yield similar (to rPICKLE) posteriors. However, both HMC and SVGD fail for the high-dimensional problem. These results demonstrate the advantages of rPICKLE for approximately sampling high-dimensional posterior distributions.

1. Introduction

Inverse uncertainty quantification (IUQ) problems are ubiquitous in the modeling of natural and engineering systems governed by partial differential equations (PDEs) (e.g., subsurface flow systems, geothermal systems, CO_2 sequestration, and climate modeling). IUQ has the same challenges as forward UQ, including the curse of dimensionality (COD), plus additional challenges associated with the ill-posedness of inverse PDE problems [1,2].

In this work, we are interested in IUQ for the PDE model

$$\mathcal{L}(u(\mathbf{x}), y(\mathbf{x})) = 0, \quad \mathbf{x} \in \Omega \tag{1}$$

subject to the appropriate boundary conditions, where $\mathcal{L}$ is the differential operator acting on the state variable $u(\mathbf{x})$ and the space-dependent parameter $y(\mathbf{x})$. This PDE can be discretized using a numerical method of choice, or modeled with physics-informed

* Corresponding author at: Department of Civil and Environmental Engineering, University of Illinois Urbana-Champaign, Urbana, IL 61801, United States of America.
E-mail address: amt1998@illinois.edu (A.M. Tartakovsky).

https://doi.org/10.1016/j.jcp.2024.113395
Received 8 January 2024; Received in revised form 19 July 2024; Accepted 3 September 2024

Available online 5 September 2024

machine learning methods such as the physics-informed neural networks (PINNs) [3–5]. In the deterministic inverse setting, point estimates of y and u were obtained by solving a PDE-constrained optimization problem where the objective function is the sum of the squared differences between predicted and observed y and u values. Adding regularization, which can be interpreted as prior knowledge, eliminates the ill-posedness of the inverse problem and leads to the maximum a posteriori (MAP) solution. However, the PDE-constrained MAP formulation may suffer from the COD. For example, for the diffusion equation model with an unknown space-dependent diffusion coefficient considered in this work, the MAP computational cost scales as $O(n^3)$, where n is the number of unknown parameters (discrete values of the diffusion coefficient) [6].

Models of natural and engineering systems can have thousands of (unknown) parameters. Several methods have been proposed for approximating such systems using a smaller number of parameters [7]. These methods include the pilot point method [8,9], subspace methods using singular value decompositions (SVD) [10], truncated Karhunen-Loéve expansions (KLE) [11], and truncated conditional KLEs (cKLE) [12]. In the KLE-based approach, the parameter estimation problem is reduced to estimating KLE coefficients. In [13,6], both the state variable $u(\mathbf{x})$ and space-dependent parameter $y(\mathbf{x})$ were projected onto the cKLE space and the inverse problem was formulated as a PDE residual least-square minimization problem. This method was termed the physics-informed cKLE (PICKLE) method. In [6], for a problem with 1000 intrinsic dimensions (the parameter field is represented with a 1000-dimensional truncated cKLE), it was shown that the PICKLE scales as $O(n^{1.15})$ (versus the cubic scaling of MAP). However, MAP and PICKLE do not describe uncertainty in the estimated parameters.

IUQ problems are commonly formulated in the Bayesian framework [11,14], where the distributions of the unknown parameters are sought [15,16,12]. When the Bayesian framework is combined with dimension reduction methods, the posterior parameter distribution in the reduced space is estimated, e.g., the distribution of the coefficients in the KLE expansions of parameters fields [11,12]. The posterior can be analytically estimated for linear forward models with Gaussian prior and likelihood models. In more general scenarios, the posterior computations are intractable [17].

Approximate inference methods either approximate the posterior with a parameterized distribution (variational inference (VI) methods) or generate a sequence of samples from the posterior (Markov chain Monte Carlo (MCMC) methods) [18]. VI methods optimize parameters in the assumed distribution to minimize the discrepancy between the parameterized and true posterior distributions based on a specific probabilistic distance metric [19,20]. For example, mean-field VI assumes a fully factorized structure of the posterior distribution [21]. However, such specific structures can yield biased approximations, particularly for the posterior exhibiting strongly correlated structure [22], which is common in PDE-constrained inversions. Furthermore, mean-field VI commonly underestimates the posterior variance, which is undesirable for reliability analysis and risk assessment. Stein Variational Gradient Descent (SVGD) [23] is a more general variational inference approach that uses a set of particles to approximate the posterior distribution. SVGD iteratively updates the particle positions by following the functional gradient of the Kullback-Leibler divergence given by a combination of the weighted gradient of the log-posterior and a kernel-based term. SVGD can better capture complex, multimodal posterior distributions [24] than the mean-field VI method. However, it is not immune to the COD [25].

MCMC methods aim to construct multiple chains of posterior parameter samples with stationary distributions equal to the target posterior distribution. However, MCMC, especially the random-walk MCMC, suffers from the COD. One way to accelerate MCMC is to reduce the auto-correlation of the Markov chains. Hamiltonian Monte Carlo (HMC) reduces this auto-correlation by making larger jumps in parameter space using simulated Hamiltonian dynamics and is a common choice for Bayesian statistical models [26,27]. However, there are challenges in applying HMC to high-dimensional IUQ problems; namely, HMC is also not immune to COD because the increasing number of unknown parameters generally reduces the sampling efficiency of HMC. Furthermore, the posterior covariance structure could be poorly conditioned [28] because the posterior distribution has vastly different correlations between different pairs of parameters. This can lead to biased HMC estimates [26,29]. Nonlinearity in the model generally exacerbates the complexity of the posterior distribution and increases HMC bias.

Randomized MAP [30] and similar randomization methods such as randomized maximum likelihood [31] and randomize-then-optimize [32] were proposed as alternatives to VI and MCMC methods for posterior sampling. In the randomized MAP method, a (deterministic) PDE-constrained optimization problem is replaced with a stochastic optimization problem where the randomized objective function is minimized subject to the same deterministic PDE constraint as in the standard MAP. The randomized MAP method has the same cubic dependence on the number of unknown parameters as the MAP method. Therefore, it is limited to low-dimensional problems.

Other commonly used approximate Bayesian inference methods include the iterative ensemble smoother (IES) and ensemble Kalman filter (KF) methods [33]. These methods can be combined with machine learning surrogates (e.g., Deep Operator Networks (DeepONets) [34], Fourier neural operator (FNO) [35], and Karhunen-Loéve Deep neural network (KL-DNN) [36]) to reduce the cost of multiple forward PDE simulations. In [37], a surrogate model was adaptively fine-tuned for near-optimal parameter values with online data selected during the Bayesian inference process to improve the model accuracy without a significant computational cost increase.

In this work, we propose an efficient sampling method for high-dimensional IUQ based on the residual least-square formulation of the inverse problem. The main ideas of our approach are to add independent Gaussian noise to each term of the objective function in a residual least-square method and solve the resulting problem for different realizations of the noise terms. The mean and variance of the noise distributions are chosen so that the ensemble of optimization problem solutions approximates the target posterior distribution. We test this approach for the PICKLE residual least-square formulation. We apply the “randomized PICKLE” (rPICKLE) method to a 2000-dimensional IUQ problem and demonstrate that it avoids the challenges of sampling complex posterior distributions with ill-conditioned covariances. Our analysis shows that the distribution of the rPICKLE samples converges to the exact posterior in the linear case with Gaussian priors. In nonlinear problems, the sampled posterior can deviate from the exact posterior, and these deviations

can be removed using a Metropolis rejection algorithm. In numerical experiments, we find that such deviations are minor and can be disregarded.

This paper is organized as follows. In Section 2, we review the PICKLE formulation and present the rPICKLE method for approximate Bayesian parameter estimation for Eq. (1). In Section 3, we formulate rPICKLE for an IUQ problem arising in groundwater flow modeling. In Section 4, we test rPICKLE for low-(15) and high-dimensional (2000) parameter estimation problems. A comparison with HMC and SVGD is provided for the low-dimensional case. Discussion and conclusions are given in Section 5.

2. Randomized PICKLE formulation

2.1. Conditional KLE

Here, we formulate rPICKLE for approximate Bayesian parameter estimation in the PDE model (1). The starting point of rPICKLE is to model the PDE parameter $y(\mathbf{x})$ and state $u(\mathbf{x})$ in Eq. (1) as random processes $y(\mathbf{x},\tilde{\omega})$ and $u(\mathbf{x},\tilde{\omega})$, respectively, where $\tilde{\omega}$ denotes a point in the corresponding probabilistic outcome space. Given observations $\mathbf{u}^{\text{obs}} = \{u_i^{\text{obs}}\}_{i=1}^{N_u^{\text{obs}}}$ of u and $\mathbf{y}^{\text{obs}} = \{y_i^{\text{obs}}\}_{i=1}^{N_y^{\text{obs}}}$ of y, $y(\mathbf{x},\tilde{\omega})$ and $u(\mathbf{x},\tilde{\omega})$ can be approximated with truncated cKLEs. For $y(\mathbf{x},\tilde{\omega})$, the cKLE takes the form

$$y(\mathbf{x},\tilde{\omega}) \approx \hat{y}(\mathbf{x},\boldsymbol{\xi}) = \overline{y}^c(\mathbf{x}) + \sum_{i=1}^{N_\xi} \sqrt{\lambda_i^y}\psi_i^y(\mathbf{x})\xi_i, \tag{2}$$

where $\overline{y}^c$ is the prior mean of y conditioned on y observations and $\{\lambda_i^y\}_{i=1}^{N_\xi}$ and $\{\psi_i^y(\mathbf{x})\}_{i=1}^{N_\xi}$ are the leading N_ξ eigenvalues and eigenfunctions of the prior covariance of y conditioned on y observations, $C_y^c(\mathbf{x},\mathbf{x}')$. Both $\overline{y}^c$ and $C_y^c(\mathbf{x},\mathbf{x}')$ reflect the prior knowledge about $y(x)$ and are obtained via the Gaussian process regression (GPR) equations given in Appendix A. In Eq. (2), $\boldsymbol{\xi} = [\xi_1,...,\xi_{N_\xi}]^T$ is a vector of random variables with prior independent standard normal distribution. For this prior of $\boldsymbol{\xi}$, the (prior) mean and covariance of $\hat{y}(\mathbf{x},\boldsymbol{\xi})$ are equal, up to the truncation error, to those of $y(\mathbf{x},\tilde{\omega})$. Estimating the posterior distribution of $\boldsymbol{\xi}$, i.e., the distribution constrained by the governing PDE, is the goal of rPICKLE.

Similarly, the cKLE of $u(\mathbf{x},\tilde{\omega})$ has the form

$$u(\mathbf{x},\tilde{\omega}) \approx \hat{u}(\mathbf{x},\boldsymbol{\eta}) = \overline{u}^c(\mathbf{x}) + \sum_{i=1}^{N_\eta} \sqrt{\lambda_i^u}\psi_i^u(\mathbf{x})\eta_i, \tag{3}$$

where $\overline{u}^c$ is the conditional mean of u and $\{\lambda_i^u\}_{i=1}^{N_\eta}$ and $\{\psi_i^u(\mathbf{x})\}_{i=1}^{N_\eta}$ are the leading eigenvalues and eigenfunctions of $C_u^c(\mathbf{x},\mathbf{x}')$, which is the prior covariance of u conditioned on u observations. The prior covariance of u is obtained by Monte Carlo sampling of the solution of Eq. (1) as described in Appendix A. As in the cKLE of y, $\{\eta_i\}_{i=1}^{N_\eta}$ are random variables with independent standard normal prior distribution. The posterior distribution of $\boldsymbol{\eta}$ is estimated as part of the rPICKLE inversion.

2.2. Revisiting PICKLE

We formulate rPICKLE by randomizing the PICKLE loss function. The PICKLE method was presented in [13,6] for solving high-dimensional inverse PDE problems with unknown space-dependent parameters. In PICKLE, the unknown $\hat{y}(\mathbf{x},\tilde{\boldsymbol{\xi}})$ and $\hat{u}(\mathbf{x},\tilde{\boldsymbol{\eta}})$ fields are treated as one realization of the random y and u fields, respectively, and $\tilde{\boldsymbol{\xi}}$ and $\tilde{\boldsymbol{\eta}}$ are (deterministic) parameters that are found as the solution of the minimization problem:

$$\begin{aligned}(\tilde{\boldsymbol{\xi}}^*,\tilde{\boldsymbol{\eta}}^*) &= \arg\min_{\tilde{\boldsymbol{\xi}},\tilde{\boldsymbol{\eta}}} L(\tilde{\boldsymbol{\xi}},\tilde{\boldsymbol{\eta}}) \\ &= \arg\min_{\tilde{\boldsymbol{\xi}},\tilde{\boldsymbol{\eta}}} \frac{\omega_r}{2}\|\mathcal{R}(\tilde{\boldsymbol{\xi}},\tilde{\boldsymbol{\eta}})\|_2^2 + \frac{\omega_\xi}{2}\|\tilde{\boldsymbol{\xi}}\|_2^2 + \frac{\omega_\eta}{2}\|\tilde{\boldsymbol{\eta}}\|_2^2,\end{aligned} \tag{4}$$

where $\mathcal{R}$ is the vector of PDE residuals computed on a discretization mesh with a numerical method of choice (finite volume discretization was used in [13]) and the last two terms are ℓ_2-regularization terms with respect to $\tilde{\boldsymbol{\xi}}$ and $\tilde{\boldsymbol{\eta}}$. In Section 2.3, we show that this PICKLE formulation provides the mode of the joint posterior distribution of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ given that the likelihood and prior distributions of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are Gaussian.

Weights ω_r, ω_ξ, and ω_η control the relative importance of each term in the loss function. Following [13], we set $\omega_\xi = \omega_\eta$. This choice is justified in the Bayesian context because ω_ξ and ω_η are related to the variances of the prior distributions of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$, which are the same and equal to one as stated in Section 2.1. Then, the minimization problem (4) can be re-written as

$$(\tilde{\boldsymbol{\xi}}^*,\tilde{\boldsymbol{\eta}}^*) = \arg\min_{\tilde{\boldsymbol{\xi}},\tilde{\boldsymbol{\eta}}} \frac{1}{2}\|\mathcal{R}(\tilde{\boldsymbol{\xi}},\tilde{\boldsymbol{\eta}})\|_2^2 + \frac{\gamma}{2}\|\tilde{\boldsymbol{\xi}}\|_2^2 + \frac{\gamma}{2}\|\tilde{\boldsymbol{\eta}}\|_2^2, \tag{5}$$

where $\gamma = \omega_\xi/\omega_r = \omega_\eta/\omega_r$ is the regularization parameter controlling the relative magnitude of the regularization. The value of γ is selected to minimize the ℓ_2 distance between the PICKLE solutions and reference fields. If the reference fields are unknown, γ can be selected via cross-validation.

2.3. Bayesian PICKLE

The Bayesian estimate of the posterior distribution of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ in the PICKLE model can be found from the Bayes rule:

$$P(\boldsymbol{\xi},\boldsymbol{\eta}|\mathcal{D}_{res}) = \frac{P(\mathcal{D}_{res}|\boldsymbol{\xi},\boldsymbol{\eta})P(\boldsymbol{\xi},\boldsymbol{\eta})}{\iint P(\mathcal{D}_{res}|\boldsymbol{\xi},\boldsymbol{\eta})P(\boldsymbol{\xi},\boldsymbol{\eta})d\boldsymbol{\xi}d\boldsymbol{\eta}}, \tag{6}$$

where $P(\boldsymbol{\xi},\boldsymbol{\eta})$ is the joint prior distribution of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$. We assume that the prior distributions of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are independent, i.e., $P(\boldsymbol{\xi},\boldsymbol{\eta}) = P(\boldsymbol{\xi})P(\boldsymbol{\eta})$. $P(\mathcal{D}_{res}|\boldsymbol{\xi},\boldsymbol{\eta})$ is the likelihood function, where $\mathcal{D}_{res}$ is the collection of PDE residuals. The double integral in the denominator of Eq. (6) is the normalization coefficient for the posterior distribution to integrate to one.

In PICKLE, cKLEs are conditioned on y and u observations. Therefore, the likelihood function only needs to specify the joint distribution of PDE residuals. The form of $P(\mathcal{D}_{res}|\boldsymbol{\xi},\boldsymbol{\eta})$ and $P(\boldsymbol{\xi},\boldsymbol{\eta})$ can be found from the PICKLE loss function $L(\tilde{\boldsymbol{\xi}},\tilde{\boldsymbol{\eta}})$ by requiring the PICKLE solution $\tilde{\boldsymbol{\xi}}^*$ and $\tilde{\boldsymbol{\eta}}^*$ to also maximize the posterior probability density $P(\boldsymbol{\xi},\boldsymbol{\eta}|\mathcal{D}_{res})$, i.e., by requiring the PICKLE solution to coincide with the mode of the posterior distribution.

This can be achieved by taking the negative logarithm of both sides of Eq. (6), yielding

$$-\ln[P(\boldsymbol{\xi},\boldsymbol{\eta}|\mathcal{D}_{res})] = \left[\frac{1}{\gamma}L(\boldsymbol{\xi},\boldsymbol{\eta}) + C\right] + \ln\iint P(\mathcal{D}_{res}|\boldsymbol{\xi},\boldsymbol{\eta})P(\boldsymbol{\xi},\boldsymbol{\eta})d\boldsymbol{\xi}d\boldsymbol{\eta}, \tag{7}$$

where C is a constant and $L(\boldsymbol{\xi},\boldsymbol{\eta})$ is the PICKLE loss defined in Eq. (5). The left-hand side of Eq. (7) is the negative logarithm of the posterior, and we postulated that

$$-\ln\left[P(\mathcal{D}_{res}|\boldsymbol{\xi},\boldsymbol{\eta})P(\boldsymbol{\xi},\boldsymbol{\eta})\right] = \frac{1}{\gamma}L(\boldsymbol{\xi},\boldsymbol{\eta}) + C. \tag{8}$$

Since the last term in Eq. (7) is independent of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$, PICKLE solutions $\tilde{\boldsymbol{\xi}}^*$ and $\tilde{\boldsymbol{\eta}}^*$ that minimize $L(\tilde{\boldsymbol{\xi}},\tilde{\boldsymbol{\eta}})$ also maximize $P(\boldsymbol{\xi},\boldsymbol{\eta}|\mathcal{D}_{res})$. We can break down Eq. (8) as

$$-\ln\left[P(\mathcal{D}_{res}|\boldsymbol{\xi},\boldsymbol{\eta})\right] = \frac{1}{2\gamma}\|\mathcal{R}(\mathbf{x};\boldsymbol{\xi},\boldsymbol{\eta})\|_2^2 + C_1, \tag{9}$$

and

$$-\ln\left[P(\boldsymbol{\xi},\boldsymbol{\eta})\right] = \frac{1}{2}\|\boldsymbol{\xi}\|_2^2 + \frac{1}{2}\|\boldsymbol{\eta}\|_2^2 + C_2, \tag{10}$$

where $C_1 + C_2 = C$.

In Eq. (9), we can choose C_1 such that the likelihood is

$$P(\mathcal{D}_{res}|\boldsymbol{\xi},\boldsymbol{\eta}) = \left(\frac{1}{\sqrt{2\pi}\sigma_r}\right)^N \exp\left(-\frac{1}{2}\mathcal{R}(\boldsymbol{\xi},\boldsymbol{\eta})^T\Sigma_r^{-1}\mathcal{R}(\boldsymbol{\xi},\boldsymbol{\eta})\right), \tag{11}$$

where $\Sigma_r = \sigma_r^2\mathbf{I}$ and $\sigma_r^2 = \gamma$. As desired, this likelihood states that the PDE residuals have zero mean.

Similarly, in Eq. (10), we can choose C_2 such that the prior is

$$P(\boldsymbol{\xi},\boldsymbol{\eta}) = \left(\frac{1}{\sqrt{2\pi}\sigma_\xi}\right)^{N_\xi}\left(\frac{1}{\sqrt{2\pi}\sigma_\eta}\right)^{N_\eta} \exp\left(-\frac{1}{2}\boldsymbol{\xi}^T\Sigma_\xi^{-1}\boldsymbol{\xi} - \frac{1}{2}\boldsymbol{\eta}^T\Sigma_\eta^{-1}\boldsymbol{\eta}\right), \tag{12}$$

where $\Sigma_\xi = \sigma_\xi^2\mathbf{I}$ ($\sigma_\xi^2 = 1$) and $\Sigma_\eta = \sigma_\eta^2\mathbf{I}$ ($\sigma_\eta^2 = 1$). In other words, the prior distribution of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ is independent and standard normal. Recall that this is consistent with the definition of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ in Section 2.1. With these independent and standard normal prior, the mean and the covariance of the y and u fields estimated from observations can be recovered, up to the truncation error, from their cKLEs.

2.4. rPICKLE: randomization of the PICKLE loss function

Computing the double integral in Eq. (6) is computationally intractable for high-dimensional $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$. In this section, we introduce the proposed rPICKLE method, which approximately samples the posterior distribution without computing this integral. In rPICKLE, the PICKLE loss function is randomized as

$$\begin{aligned} L^r(\boldsymbol{\xi},\boldsymbol{\eta};\boldsymbol{\omega},\boldsymbol{\alpha},\boldsymbol{\beta}) = &\frac{1}{2}\|\mathcal{R}(\boldsymbol{\xi},\boldsymbol{\eta}) - \boldsymbol{\omega}\|_{\Sigma_r}^2 + \frac{1}{2}\|\boldsymbol{\xi} - \boldsymbol{\alpha}\|_{\Sigma_\xi}^2 \\ &+ \frac{1}{2}\|\boldsymbol{\eta} - \boldsymbol{\beta}\|_{\Sigma_\eta}^2, \end{aligned} \tag{13}$$

where $\boldsymbol{\omega}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ are independent zero-mean random noise vectors with covariances $\Sigma_\omega = \Sigma_r = \sigma_r^2\mathbf{I}$, $\Sigma_\alpha = \Sigma_\xi = \mathbf{I}$, and $\Sigma_\beta = \Sigma_\eta = \mathbf{I}$, respectively. Then, the samples of the posterior distribution $P(\boldsymbol{\xi},\boldsymbol{\eta}|\mathcal{D}_{res})$ are generated by repeatedly solving the minimization problem

$$(\boldsymbol{\xi}^*,\boldsymbol{\eta}^*) = \arg\min_{\boldsymbol{\xi},\boldsymbol{\eta}} L^r(\boldsymbol{\xi},\boldsymbol{\eta};\boldsymbol{\omega},\boldsymbol{\alpha},\boldsymbol{\beta}), \tag{14}$$

for different realizations of $\boldsymbol{\omega}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$.

The rPICKLE method can be summarized as follows: We draw N_{ens} i.i.d. samples of $\boldsymbol{\omega}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ and, for each sample, minimize the loss (14) to obtain samples $(\boldsymbol{\xi}_i^*, \boldsymbol{\eta}_i^*)$ $(i = 1, ..., N_{\text{ens}})$, which approximate the posterior distribution of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$. Then, we use the cKLEs (Eq. (2) and (3)) to obtain the samples $\hat{y}(\mathbf{x}; \boldsymbol{\xi}_i^*)$ and $\hat{u}(\mathbf{x}; \boldsymbol{\eta}_i^*)$ of the predictive posterior distribution of y and u, respectively. The samples $\hat{y}(\mathbf{x}; \boldsymbol{\xi}_i^*)$ and $\hat{u}(\mathbf{x}; \boldsymbol{\eta}_i^*)$ can be used to compute the distributions of y and u or the leading moments of these distributions. For example, the first and second moments of the y distribution can be estimated as

$$\mu_{\hat{y}}(\mathbf{x}|\mathcal{D}_{res}) \approx \frac{1}{N_{\text{ens}}} \sum_{i=1}^{N_{\text{ens}}} \hat{y}(\mathbf{x}; \boldsymbol{\xi}_i^*), \tag{15}$$

$$\sigma_{\hat{y}}^2(\mathbf{x}|\mathcal{D}_{res}) \approx \frac{1}{N_{\text{ens}} - 1} \sum_{i=1}^{N_{\text{ens}}} [\hat{y}(\mathbf{x}; \boldsymbol{\xi}_i^*) - \mu_{\hat{y}}(\mathbf{x}|\mathcal{D}_{res})]^2. \tag{16}$$

In Section 2.4.1, we analytically show that rPICKLE posterior samples recover the exact posterior of $(\boldsymbol{\xi}, \boldsymbol{\eta})$ for linear problems. In Section 4.1, for a low-dimensional weakly non-linear problem, we show via comparisons with HMC and SVGD that distributions approximated with the samples $\{(\boldsymbol{\xi}_i^*, \boldsymbol{\eta}_i^*)\}_{i=1}^{N_{\text{ens}}}$ approach the *true* posterior with increasing N_{ens} regardless of the choice of σ_r^2. In Section 4.2, we numerically demonstrate for high-dimensional problems that the rPICKLE posterior approximates well the unknown parameter field. For this problem, we cannot obtain an HMC posterior approximation to validate the rPICKLE ensemble's convergence to the true posterior due to the prohibitive computational cost of HMC.

We note that the posterior distribution depends on the choice of σ_r^2. If true field y is known, one possible criterion for selecting the value of σ_r^2 is to minimize the distance between either the MAP estimate $\hat{y}_{\text{MAP}}$ or the mean estimate $\mu_{\hat{y}}(\mathbf{x}|\mathcal{D}_{res})$ and the true field. Alternatively, when the true field is unknown but some test data $\{y(\mathbf{x}_i)\}$ is available, one can select the σ_r^2 that maximizes the log-predictive probability (LPP) of the test data, defined as the sum of the pointwise log-probabilities of the data being observed given the predictive posterior [38]:

$$\text{LPP} = -\sum_{i=1}^{N} \left\{ \frac{[\mu_{\hat{y}}(\mathbf{x}_i|\mathcal{D}_{res}) - y(\mathbf{x}_i)]^2}{2\sigma_{\hat{y}}^2(\mathbf{x}_i)} + \frac{1}{2} \log[2\pi\sigma_{\hat{y}}^2(\mathbf{x}_i)] \right\}. \tag{17}$$

2.4.1. rPICKLE for a linear model

In this section, we prove that rPICKLE samples converge to the exact posterior as $N_{\text{ens}} \to \infty$ for linear (with respect to $\boldsymbol{\xi}, \boldsymbol{\eta}$) PDE residuals of the form $\mathcal{R}(\boldsymbol{\xi}, \boldsymbol{\eta}) = \mathbf{A}\boldsymbol{\xi} + \mathbf{B}\boldsymbol{\eta} - \mathbf{c} \in \mathbb{R}^N$, with $\mathbf{A} \in \mathbb{R}^{N \times N_\xi}, \mathbf{B} \in \mathbb{R}^{N \times N_\eta}$, and $\mathbf{c} \in \mathbb{R}^N$. This proof relies on the fact that the Bayesian posterior is Gaussian for Gaussian priors on $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$, and linear $\mathcal{R}$ with Gaussian likelihood [38].

First, we find the mean and covariance of the posterior given by Bayes' rule. Taking the first and second derivatives of both sides of Eq. (7) with respect to $[\boldsymbol{\xi}, \boldsymbol{\eta}]^T$ yields the relationships between the mean $\boldsymbol{\mu}_{post}$ and covariance Σ_{post} of the posterior distribution:

$$\boldsymbol{\mu}_{post} = \Sigma_{post} \begin{bmatrix} \mathbf{A}^T \Sigma_r^{-1} \mathbf{c} \\ \mathbf{B}^T \Sigma_r^{-1} \mathbf{c} \end{bmatrix}. \tag{18}$$

Then, Eq. (7) can be reformulated as:

$$\begin{aligned} &\left(\begin{bmatrix} \boldsymbol{\xi} \\ \boldsymbol{\eta} \end{bmatrix} - \boldsymbol{\mu}_{post} \right)^T \Sigma_{post}^{-1} \left(\begin{bmatrix} \boldsymbol{\xi} \\ \boldsymbol{\eta} \end{bmatrix} - \boldsymbol{\mu}_{post} \right) \\ &= \left(\begin{bmatrix} \mathbf{A}\boldsymbol{\xi} \\ \mathbf{B}\boldsymbol{\eta} \end{bmatrix} - \mathbf{c} \right)^T \Sigma_r^{-1} \left(\begin{bmatrix} \mathbf{A}\boldsymbol{\xi} \\ \mathbf{B}\boldsymbol{\eta} \end{bmatrix} - \mathbf{c} \right) + \begin{bmatrix} \boldsymbol{\xi} \\ \boldsymbol{\eta} \end{bmatrix}^T \begin{bmatrix} \Sigma_\xi^{-1} & \mathbf{0} \\ \mathbf{0} & \Sigma_\eta^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\xi} \\ \boldsymbol{\eta} \end{bmatrix}. \end{aligned} \tag{19}$$

Differentiating Eq. (19) twice with respect to $[\boldsymbol{\xi}, \boldsymbol{\eta}]^T$ yields the expression for Σ_{post}:

$$\Sigma_{post} = \begin{bmatrix} \mathbf{A}^T \Sigma_r^{-1} \mathbf{A} + \Sigma_\xi^{-1} & \mathbf{B}^T \Sigma_r^{-1} \mathbf{A} \\ \mathbf{A}^T \Sigma_r^{-1} \mathbf{B} & \mathbf{B}^T \Sigma_r^{-1} \mathbf{B} + \Sigma_\eta^{-1} \end{bmatrix}^{-1}. \tag{20}$$

Next, we derive the mean and covariance of $\boldsymbol{\xi}^*$ and $\boldsymbol{\eta}^*$ given by rPICKLE (Eq. (13)). For linear $\mathcal{R}$, the randomized minimization problem (14) is reduced to a linear least-square problem with the solution given by the following system of linear equations:

$$\frac{\partial L^r}{\partial \boldsymbol{\xi}} = \left[(\mathbf{A}\boldsymbol{\xi} + \mathbf{B}\boldsymbol{\eta} - \mathbf{c})^T \Sigma_r^{-1} \mathbf{A} - \boldsymbol{\omega}^T \Sigma_r^{-1} \mathbf{A} \right] + \left[(\boldsymbol{\xi} - \boldsymbol{\alpha})^T \Sigma_\xi^{-1} \right] = 0, \tag{21}$$

$$\frac{\partial L^r}{\partial \boldsymbol{\eta}} = \left[(\mathbf{A}\boldsymbol{\xi} + \mathbf{B}\boldsymbol{\eta} - \mathbf{c})^T \Sigma_r^{-1} \mathbf{B} - \boldsymbol{\omega}^T \Sigma_r^{-1} \mathbf{B} \right] + \left[(\boldsymbol{\eta} - \boldsymbol{\beta})^T \Sigma_\eta^{-1} \right] = 0, \tag{22}$$

or

$$\begin{bmatrix} \mathbf{A}^T \Sigma_r^{-1} \mathbf{A} + \Sigma_\xi^{-1} & \mathbf{B}^T \Sigma_r^{-1} \mathbf{A} \\ \mathbf{A}^T \Sigma_r^{-1} \mathbf{B} & \mathbf{B}^T \Sigma_r^{-1} \mathbf{B} + \Sigma_\eta^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\xi}^* \\ \boldsymbol{\eta}^* \end{bmatrix} = \begin{bmatrix} \mathbf{A}^T \Sigma_r^{-1} (\mathbf{c} + \boldsymbol{\omega}) + \Sigma_\xi^{-1} \boldsymbol{\alpha} \\ \mathbf{B}^T \Sigma_r^{-1} (\mathbf{c} + \boldsymbol{\omega}) + \Sigma_\eta^{-1} \boldsymbol{\beta} \end{bmatrix}. \tag{23}$$

The solution of these equations is:

$$\begin{bmatrix}\boldsymbol{\xi}^*\\ \boldsymbol{\eta}^*\end{bmatrix} = \Sigma\begin{bmatrix}\mathbf{A}^T\Sigma_r^{-1}(\mathbf{c}+\boldsymbol{\omega})+\Sigma_\xi^{-1}\boldsymbol{\alpha}\\ \mathbf{B}^T\Sigma_r^{-1}(\mathbf{c}+\boldsymbol{\omega})+\Sigma_\eta^{-1}\boldsymbol{\beta}\end{bmatrix}, \tag{24}$$

where

$$\Sigma = \begin{bmatrix}\mathbf{A}^T\Sigma_r^{-1}\mathbf{A}+\Sigma_\xi^{-1} & \mathbf{B}^T\Sigma_r^{-1}\mathbf{A}\\ \mathbf{A}^T\Sigma_r^{-1}\mathbf{B} & \mathbf{B}^T\Sigma_r^{-1}\mathbf{B}+\Sigma_\eta^{-1}\end{bmatrix}^{-1}. \tag{25}$$

Comparing Eqs. (25) and (20) yields $\Sigma = \Sigma_{post}$. Recall that $[\boldsymbol{\xi}^*, \boldsymbol{\eta}^*]^T$ is a function of the random noises $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\omega})$. Taking the expectation of $[\boldsymbol{\xi}^*, \boldsymbol{\eta}^*]^T$, we get

$$\begin{aligned}\mathbb{E}\begin{bmatrix}\boldsymbol{\xi}^*\\ \boldsymbol{\eta}^*\end{bmatrix} &= \mathbb{E}\left[\Sigma\begin{bmatrix}\mathbf{A}^T\Sigma_r^{-1}(\mathbf{c}+\boldsymbol{\omega})+\Sigma_\xi^{-1}\boldsymbol{\alpha}\\ \mathbf{B}^T\Sigma_r^{-1}(\mathbf{c}+\boldsymbol{\omega})+\Sigma_\eta^{-1}\boldsymbol{\beta}\end{bmatrix}\right]\\ &= \Sigma\begin{bmatrix}\mathbf{A}^T\Sigma_r^{-1}(\mathbf{c}+\mathbb{E}[\boldsymbol{\omega}])+\Sigma_\xi^{-1}\mathbb{E}[\boldsymbol{\alpha}]\\ \mathbf{B}^T\Sigma_r^{-1}(\mathbf{c}+\mathbb{E}[\boldsymbol{\omega}])+\Sigma_\eta^{-1}\mathbb{E}[\boldsymbol{\beta}]\end{bmatrix}\\ &= \Sigma\begin{bmatrix}\mathbf{A}^T\Sigma_r^{-1}\mathbf{c}\\ \mathbf{B}^T\Sigma_r^{-1}\mathbf{c}\end{bmatrix} = \boldsymbol{\mu}.\end{aligned} \tag{26}$$

Comparing Eqs. (26) and (18) yields $\boldsymbol{\mu} = \boldsymbol{\mu}_{post}$. Next, we prove that $\Sigma = \Sigma_{post}$ is the covariance of $[\boldsymbol{\xi}^*, \boldsymbol{\eta}^*]^T$. The covariance of $\boldsymbol{\zeta} = [\boldsymbol{\xi}^*, \boldsymbol{\eta}^*]^T$ can be calculated as

$$\begin{aligned}&\mathbb{E}[(\boldsymbol{\zeta}-\mathbb{E}[\boldsymbol{\zeta}])(\boldsymbol{\zeta}-\mathbb{E}[\boldsymbol{\zeta}])^T]\\ &= \Sigma\mathbb{E}\left[\begin{bmatrix}\mathbf{A}^T\Sigma_r^{-1}\boldsymbol{\omega}+\Sigma_\xi^{-1}\boldsymbol{\alpha}\\ \mathbf{B}^T\Sigma_r^{-1}\boldsymbol{\omega}+\Sigma_\eta^{-1}\boldsymbol{\beta}\end{bmatrix}\begin{bmatrix}\mathbf{A}^T\Sigma_r^{-1}\boldsymbol{\omega}+\Sigma_\xi^{-1}\boldsymbol{\alpha} & \mathbf{B}^T\Sigma_r^{-1}\boldsymbol{\omega}+\Sigma_\eta^{-1}\boldsymbol{\beta}\end{bmatrix}\right]\Sigma\\ &= \Sigma\begin{bmatrix}M_{11} & M_{12}\\ M_{21} & M_{22}\end{bmatrix}\Sigma,\end{aligned} \tag{27}$$

where

$$\begin{aligned}M_{11} &= \mathbf{A}^T\Sigma_r^{-1}\Sigma_r^{-1}\mathbf{A}\mathbb{E}[\boldsymbol{\omega}\boldsymbol{\omega}^T]+\Sigma_\xi^{-1}\Sigma_\xi^{-1}\mathbb{E}[\boldsymbol{\alpha}\boldsymbol{\alpha}^T]\\ M_{12} &= \mathbf{B}^T\Sigma_r^{-1}\Sigma_r^{-1}\mathbf{A}\mathbb{E}[\boldsymbol{\omega}\boldsymbol{\omega}^T]\\ M_{21} &= \mathbf{A}^T\Sigma_r^{-1}\Sigma_r^{-1}\mathbf{B}\mathbb{E}[\boldsymbol{\omega}\boldsymbol{\omega}^T]\\ M_{22} &= \mathbf{B}^T\Sigma_r^{-1}\Sigma_r^{-1}\mathbf{B}\mathbb{E}[\boldsymbol{\omega}\boldsymbol{\omega}^T]+\Sigma_\eta^{-1}\Sigma_\eta^{-1}\mathbb{E}[\boldsymbol{\beta}\boldsymbol{\beta}^T].\end{aligned}$$

For $\Sigma_\omega = \mathbb{E}[\boldsymbol{\omega}\boldsymbol{\omega}^T] = \Sigma_r$, $\Sigma_\alpha = \mathbb{E}[\boldsymbol{\alpha}\boldsymbol{\alpha}^T] = \Sigma_\xi$, and $\Sigma_\beta = \mathbb{E}[\boldsymbol{\beta}\boldsymbol{\beta}^T] = \Sigma_\eta$, Eq. (27) reduces to

$$\mathbb{E}[(\boldsymbol{\zeta}-\boldsymbol{\mu})(\boldsymbol{\zeta}-\boldsymbol{\mu})^T] = \Sigma\Sigma^{-1}\Sigma = \Sigma_{post}. \tag{28}$$

Finally, we note that in rPICKLE, N_{ens} samples of $\boldsymbol{\zeta}$ are obtained and used to compute the sample mean and covariance of $\boldsymbol{\zeta}$. As $N_{\text{ens}} \to \infty$, the sample mean and covariance converge to their ensemble counterparts.

2.4.2. Metropolis rejection

When the residual $\mathcal{R}(\boldsymbol{\xi}, \boldsymbol{\eta})$ is not linear with respect to $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$, rPICKLE samples may deviate from the true posterior. This deviation, however, can be corrected via Metropolis rejection sampling.

Recall that the random noise vectors $\boldsymbol{\omega}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ have an independent joint distribution:

$$p(\boldsymbol{\omega}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \exp(-\boldsymbol{\omega}^T\Sigma_\omega^{-1}\boldsymbol{\omega} - \boldsymbol{\alpha}^T\Sigma_\alpha^{-1}\boldsymbol{\alpha} - \boldsymbol{\beta}^T\Sigma_\beta^{-1}\boldsymbol{\beta}). \tag{29}$$

Next, we define a random vector $\boldsymbol{\delta} = \mathcal{R}(\boldsymbol{\xi}^*, \boldsymbol{\eta}^*) - \boldsymbol{\omega}$ and assume that there exists an invertible map $\mathcal{G} : (\boldsymbol{\omega}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \to (\boldsymbol{\delta}, \boldsymbol{\xi}^*, \boldsymbol{\eta}^*)$, where $\boldsymbol{\xi}^*$ and $\boldsymbol{\eta}^*$ minimize the rPICKLE loss function L^r. Because $p(\boldsymbol{\omega}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is known, the joint distribution of $(\boldsymbol{\delta}, \boldsymbol{\xi}^*, \boldsymbol{\eta}^*)$ can be computed as

$$f(\boldsymbol{\delta}, \boldsymbol{\xi}^*, \boldsymbol{\eta}^*) = p(\boldsymbol{\omega}, \boldsymbol{\alpha}, \boldsymbol{\beta})|\det(\mathbf{J})|, \tag{30}$$

where f is the probability density function of $(\boldsymbol{\delta}, \boldsymbol{\xi}^*, \boldsymbol{\eta}^*)$ and $\mathbf{J}$ is the Jacobian of the map $\mathcal{G}$ defined as

$$\mathbf{J} := \frac{\partial(\boldsymbol{\omega}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial(\boldsymbol{\delta}, \boldsymbol{\xi}^*, \boldsymbol{\eta}^*)}. \tag{31}$$

To find this Jacobian, we note that for a general residual function $\mathcal{R}$,

$$\frac{\partial L^r}{\partial \boldsymbol{\xi}} = \boldsymbol{\delta}^T \Sigma_r^{-1} \frac{\partial \mathcal{R}}{\partial \boldsymbol{\xi}} + (\boldsymbol{\xi} - \boldsymbol{\alpha})^T \Sigma_\xi^{-1} \tag{32}$$

$$\frac{\partial L^r}{\partial \boldsymbol{\eta}} = \boldsymbol{\delta}^T \Sigma_r^{-1} \frac{\partial \mathcal{R}}{\partial \boldsymbol{\eta}} + (\boldsymbol{\eta} - \boldsymbol{\beta})^T \Sigma_\eta^{-1} \tag{33}$$

and $(\boldsymbol{\xi}^*, \boldsymbol{\eta}^*)$ is the solution of $(\frac{\partial L^r}{\partial \boldsymbol{\xi}}, \frac{\partial L^r}{\partial \boldsymbol{\eta}})^T = \mathbf{0}$. Then, $\mathcal{G}$ is implicitly expressed as:

$$\begin{cases} \boldsymbol{\alpha} = \Sigma_\xi \frac{\partial \mathcal{R}}{\partial \boldsymbol{\xi}}^T \Sigma_r^{-1} \boldsymbol{\delta} + \boldsymbol{\xi}^* \\ \boldsymbol{\beta} = \Sigma_\eta \frac{\partial \mathcal{R}}{\partial \boldsymbol{\eta}}^T \Sigma_r^{-1} \boldsymbol{\delta} + \boldsymbol{\eta}^* \\ \boldsymbol{\omega} = \mathcal{R}(\boldsymbol{\xi}^*, \boldsymbol{\eta}^*) - \boldsymbol{\delta}. \end{cases} \tag{34}$$

The explicit form of the Jacobian can then be obtained as

$$\begin{aligned} \mathbf{J} &:= \begin{bmatrix} \frac{\partial \boldsymbol{\omega}}{\partial \boldsymbol{\delta}} & \frac{\partial \boldsymbol{\omega}}{\partial \boldsymbol{\xi}^*} & \frac{\partial \boldsymbol{\omega}}{\partial \boldsymbol{\eta}^*} \\ \frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{\delta}} & \frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{\xi}^*} & \frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{\eta}^*} \\ \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\delta}} & \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\xi}^*} & \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\eta}^*} \end{bmatrix} \\ &= \begin{bmatrix} -\mathbf{I} & \frac{\partial \mathcal{R}}{\partial \boldsymbol{\xi}} & \frac{\partial \mathcal{R}}{\partial \boldsymbol{\eta}} \\ \Sigma_\xi \frac{\partial \mathcal{R}}{\partial \boldsymbol{\xi}}^T \Sigma_r^{-1} & \Sigma_\xi (\frac{\partial^2 \mathcal{R}}{\partial \boldsymbol{\xi}^2})^T \otimes (\Sigma_r^{-1} \boldsymbol{\delta}) & \Sigma_\xi (\frac{\partial^2 \mathcal{R}}{\partial \boldsymbol{\xi} \partial \boldsymbol{\eta}})^T \otimes (\Sigma_r^{-1} \boldsymbol{\delta}) \\ \Sigma_\eta \frac{\partial \mathcal{R}}{\partial \boldsymbol{\eta}}^T \Sigma_r^{-1} & \Sigma_\eta (\frac{\partial^2 \mathcal{R}}{\partial \boldsymbol{\eta} \partial \boldsymbol{\xi}})^T \otimes (\Sigma_r^{-1} \boldsymbol{\delta}) & \Sigma_\eta (\frac{\partial^2 \mathcal{R}}{\partial \boldsymbol{\eta}^2})^T \otimes (\Sigma_r^{-1} \boldsymbol{\delta}) \end{bmatrix} \\ &= \begin{bmatrix} -\mathbf{I} & \frac{\partial \mathcal{R}}{\partial \boldsymbol{\zeta}} \\ \widehat{\Sigma} \frac{\partial \mathcal{R}}{\partial \boldsymbol{\zeta}}^T \Sigma_r^{-1} & \mathbf{I} + \widehat{\Sigma} (\frac{\partial^2 \mathcal{R}}{\partial \boldsymbol{\zeta}^2})^T \otimes (\Sigma_r^{-1} \boldsymbol{\delta}) \end{bmatrix}, \end{aligned} \tag{35}$$

where $\otimes$ denotes the tensor product and

$$\widehat{\Sigma} = \begin{bmatrix} \Sigma_\xi & \mathbf{0} \\ \mathbf{0} & \Sigma_\eta \end{bmatrix}. \tag{36}$$

The Jacobian $\mathbf{J}$ size is $(N_r + M) \times (N_r + M)$, where $M = N_\xi + N_\eta$ and N_r is the number of cells in the finite volume mesh where the PDE residuals are evaluated. However, computing $|\det(\mathbf{J})|$ only requires computing the determinant of the $M \times M$ Schur complement as

$$|\det(\mathbf{J})| = \left| \det \left[\mathbf{I} + \widehat{\Sigma} \left(\frac{\partial^2 \mathcal{R}}{\partial \boldsymbol{\zeta}^2} \right)^T \otimes (\Sigma_r^{-1} \boldsymbol{\delta}) + \widehat{\Sigma} \left(\frac{\partial \mathcal{R}}{\partial \boldsymbol{\zeta}} \right)^T \Sigma_r^{-1} \frac{\partial \mathcal{R}}{\partial \boldsymbol{\zeta}} \right] \right|. \tag{37}$$

The computational complexity of evaluating $|\det(\mathbf{J})|$ is dominated by the computation of the tensor product and the determinant itself. The tensor product requires computing N_r Hessians of size $M \times M$, and computing each Hessian has complexity $O(M^2)$, so the cost of evaluating the argument of the determinant scales as $O(M^2)$ with increasing M for fixed N_r. Evaluating the determinant itself has complexity equivalent to that of $M \times M$ matrix multiplication [39], which scales as $O(M^\epsilon)$, where $2 \leq \epsilon \leq 3$ is the exponent of the matrix multiplication algorithm employed. Therefore, for fixed N_r, the overall cost of computing $|\det(\mathbf{J})|$ scales as $O(M^\epsilon)$.

Having obtained a closed-form expression for $|\det(\mathbf{J})|$, we formulate the Metropolis rejection method for the rPICKLE as shown in Algorithm 1. First, the kth samples $\boldsymbol{\omega}_k$, $\boldsymbol{\alpha}_k$, and $\boldsymbol{\beta}_k$ $(k = 1, ..., N_{\text{ens}})$ are independently drawn. Then, the kth samples $\boldsymbol{\zeta}_k = (\boldsymbol{\xi}_k^*, \boldsymbol{\eta}_k^*)$ of the posterior distribution are found as the solution of the rPICKLE minimization problem (14) for $(\boldsymbol{\omega}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = (\boldsymbol{\omega}_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k)$. The first samples $(k = 1)$ are automatically accepted. The kth sample $(k > 1)$ is accepted or rejected according to the independent Metropolis-Hastings (IMH) method [40] with acceptance ratio α given by

$$\alpha(\boldsymbol{\zeta}_{k-1}, \boldsymbol{\zeta}_k) = \min \left\{ 1, \frac{\pi(\boldsymbol{\zeta}_k) q(\boldsymbol{\zeta}_{k-1})}{\pi(\boldsymbol{\zeta}_{k-1}) q(\boldsymbol{\zeta}_k)} \right\}, \tag{38}$$

where $\pi(\cdot)$ is a function proportional to the Bayesian posterior and $q(\cdot)$ is the proposal density function, defined as

$$\begin{aligned} q(\boldsymbol{\zeta}) &= \int f(\boldsymbol{\delta}, \boldsymbol{\xi}^*, \boldsymbol{\eta}^*) d\boldsymbol{\delta} \\ &= \int p(\boldsymbol{\omega}, \boldsymbol{\alpha}, \boldsymbol{\beta}) |\det(\mathbf{J})| d\boldsymbol{\delta}, \end{aligned} \tag{39}$$

where the mapping from $(\boldsymbol{\omega}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ to $(\boldsymbol{\xi}, \boldsymbol{\eta}, \boldsymbol{\delta})$ is given by Eq. (34). This marginalization is in general computationally prohibitive; therefore, following [30], we use the following approximate expression for the acceptance ratio:

$$\alpha(\boldsymbol{\zeta}_{k-1}, \boldsymbol{\zeta}_k) \approx \min \left\{ 1, \frac{|\det(\mathbf{J}_{\zeta_k})|^{1/2}}{|\det(\mathbf{J}_{\zeta_{k-1}})|^{1/2}} \right\}. \tag{40}$$

A number g is drawn from the continuous uniform distribution $U(0,1)$ and the kth sample is accepted if $g \leq \alpha(\boldsymbol{\zeta}_k, \boldsymbol{\zeta}_{k-1})$. Otherwise, the kth sample is rejected, i.e., $\boldsymbol{\zeta}_k$ is replaced with $\boldsymbol{\zeta}_{k-1}$.

It should be noted that in this IMH algorithm, the proposal function q is not conditioned on the previous sample. However, the acceptance ratio depends on the previous sample, and the total transition obeys the Markov property. In the case of the linear $\mathcal{R}$ model considered in Section 2.4.1, every proposal is accepted because the determinant of the Jacobian is a constant and independent of $\boldsymbol{\zeta}$.

Algorithm 1 Randomized PICKLE Algorithm.

Require: number of samples N_{ens}
for $k = 1, \ldots, N_{\text{ens}}$ **do**
  Sample random noises $\boldsymbol{\omega}_k \sim \mathcal{N}(\mathbf{0}, \Sigma_r)$, $\boldsymbol{\alpha}_k \sim \mathcal{N}(\mathbf{0}, \Sigma_\xi)$, $\boldsymbol{\beta}_k \sim \mathcal{N}(\mathbf{0}, \Sigma_\eta)$
  Propose $\boldsymbol{\zeta}' = (\boldsymbol{\xi}^*, \boldsymbol{\eta}^*)$ by optimizing the randomized loss (14)
  Calculate acceptance ratio $\alpha(\boldsymbol{\zeta}', \boldsymbol{\zeta}_{k-1}) = \min\left\{1, \frac{|\det(\mathbf{J}_{\boldsymbol{\zeta}'})|^{1/2}}{|\det(\mathbf{J}_{\boldsymbol{\zeta}_{k-1}})|^{1/2}}\right\}$
  Sample $g \sim U(0,1)$
  if $g \leq \alpha(\boldsymbol{\zeta}', \boldsymbol{\zeta}_{k-1})$ **then**
    $\boldsymbol{\zeta}_k \leftarrow \boldsymbol{\zeta}'$ (accept $\boldsymbol{\zeta}'$ and set it as the new state)
  else
    $\boldsymbol{\zeta}_k \leftarrow \boldsymbol{\zeta}_{k-1}$ (reject $\boldsymbol{\zeta}'$ and copy the old state forward)
  end if
end for

We now briefly discuss the computational complexity of Algorithm 1. The per-sample cost of rPICKLE is dominated by two operations: (1) minimizing the randomized loss (14), which depends on N_r, M, the residual function, and the iterative minimization algorithm employed, and (2) evaluating $|\det(\mathbf{J})|$. For the trust region reflective algorithm [41], the cost per minimization iteration scales as follows: For fixed M and increasing N_r, this cost is dominated by the evaluation of the Jacobian $\partial\mathcal{R}/\partial\zeta$, which scales as $O(N_r)$ if computed using reverse-mode automatic differentiation. This is consistent with [6] where, for the problem described in Section 3.1, the overall PICKLE computational cost was found to scale as $O(N_r^{1.15})$. On the other hand, for fixed N_r and increasing M, the cost per minimization iteration is dominated by the solution of a dense $M \times M$ linear problem to find the update direction, which scales as $O(M^\epsilon)$, where again ϵ is the exponent of the matrix multiplication algorithm employed.

Assuming that the number of iterations does not depend on M, it follows that the cost of the overall minimization process scales as $O(M^\epsilon)$ with increasing M. This scaling is the same as that of the evaluation of $|\det(\mathbf{J})|$. Nevertheless, when performing the tests presented in Section 4.1 we find empirically that the per-sample wall-clock cost is dominated by the evaluation of $|\det(\mathbf{J})|$, which indicates that the wall-clock cost of this operation has a higher constant factor than the minimization operation.

The per-sample computational cost of HMC is proportional to the cost of computing the gradient of the log-posterior (which scales as $O(1)$ for fixed N_r and increasing M) times the number L of leapfrog steps per HMC proposal. The scaling of L with problem dimension d is known to be $O(d^{1/4})$ for the simplified case of a posterior consisting of independent, identically distributed components [42]. However, this scaling is not known for a more general case, and the overall HMC computational complexity for the problem presented in Section 3.1 cannot be established. In Section 4.4 we empirically find that the efficiency of HMC decreases with increasing M and decreasing σ_r, so that HMC becomes prohibitively expensive for sufficiently large problems and sufficiently small σ_r.

3. Application to the Hanford Site groundwater flow model

3.1. Governing equations

We test the rPICKLE method for a two-dimensional steady-state groundwater flow model described by the boundary value problem (BVP)

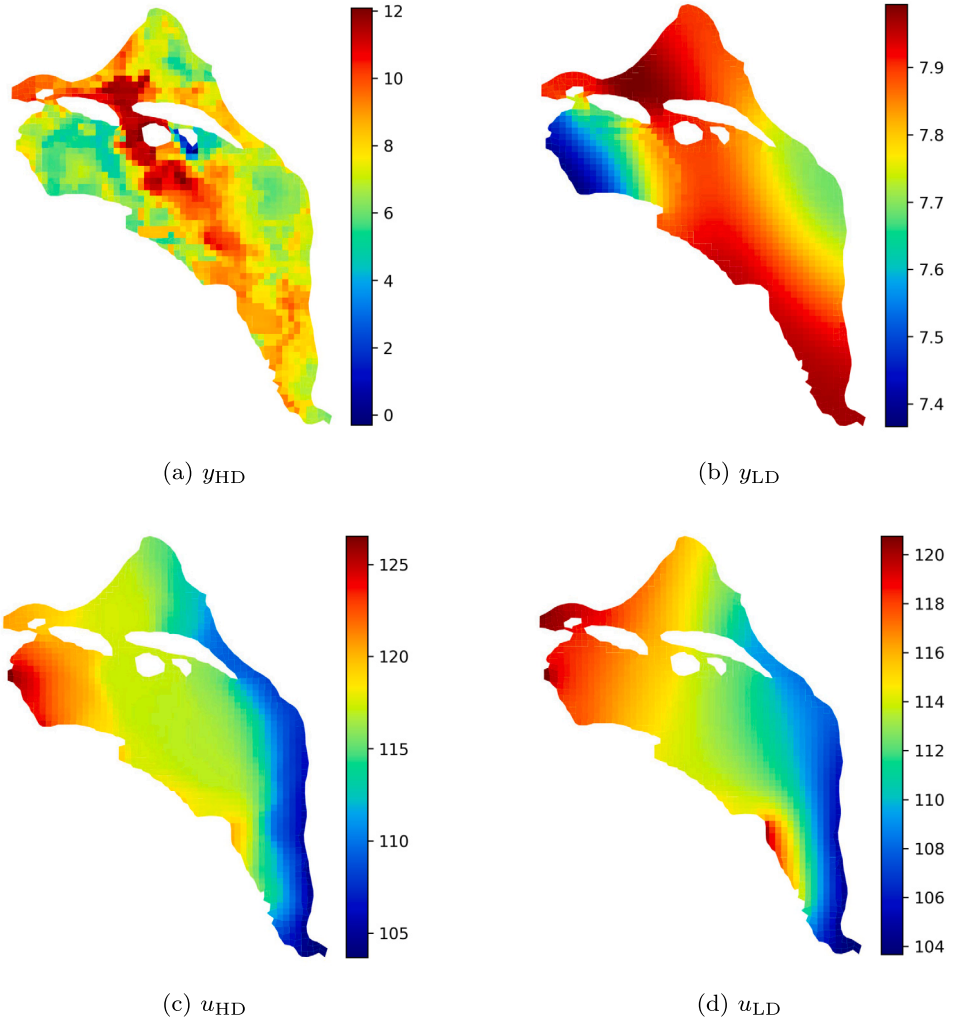
$$\nabla \cdot (e^{y(\mathbf{x})} \nabla u(\mathbf{x})) = 0 \qquad \mathbf{x} \in \Omega \tag{41}$$

$$e^{y(\mathbf{x})} \nabla u(\mathbf{x}) \cdot n(\mathbf{x}) = q_{\mathcal{N}}(\mathbf{x}) \quad \mathbf{x} \in \partial\Omega_{\mathcal{N}} \tag{42}$$

$$u(\mathbf{x}) = u_D(\mathbf{x}) \quad \mathbf{x} \in \partial\Omega_D, \tag{43}$$

where $y(\mathbf{x})$ is the logarithm of the transmissivity field, $u(\mathbf{x})$ is the hydraulic head, $\partial\Omega_{\mathcal{N}}$ is the Neumann boundary, $\partial\Omega_D$ is the Dirichlet boundary, $\partial\Omega_{\mathcal{N}} \cap \partial\Omega_D = \varnothing$, $q_{\mathcal{N}}(\mathbf{x})$ is the prescribed normal flux at the Neumann boundary, $n(\mathbf{x})$ is the unit normal vector, and $u_D(\mathbf{x})$ is the prescribed hydraulic head at the Dirichlet boundary. It is common to treat $y(\mathbf{x})$ as a realization of a correlated Gaussian field. Also, it was found that solving an inverse problem for y instead of transmissivity decreases the level of non-convexity in optimization problems [43].

We use a cell-centered finite volume (FV) discretization for Eqs. (41)–(43) and the two-point flux approximation (TPFA) to compute residuals in the rPICKLE objective function. The numerical domain Ω is discretized into N finite volume cells, and $u(\mathbf{x})$ and $y(\mathbf{x})$ are approximated via the $\mathbf{u} = \{u_1, \ldots, u_N\}$ and $\mathbf{y} = \{y_1, \ldots, y_N\}$ vectors of their respective values evaluated at the cell centers $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. For the inverse problem, we assume there are N_y^{obs} observations of y, $\mathbf{y}^{\text{obs}} = \{y_i^{\text{obs}}\}_{i=1}^{N_y^{\text{obs}}}$, and N_u^{obs} observations of u, $\mathbf{u}^{\text{obs}} = \{u_i^{\text{obs}}\}_{i=1}^{N_u^{\text{obs}}}$.



(a) y_{HD} (b) y_{LD}

(c) u_{HD} (d) u_{LD}

Fig. 1. Reference (a) high-dimensional y_{HD} and (b) low-dimensional y_{LD} log-transmissivity fields and the corresponding reference (c) high-dimensional u_{HD} and (d) low-dimensional u_{LD} hydraulic head fields. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

3.2. Hanford Site case study

The rPICKLE method is tested for estimating subsurface parameters at the Hanford Site, a U.S. Department of Energy site situated on the Columbia River in Washington state. We use Eqs. (41)–(43) to describe the two-dimensional (depth-averaged) groundwater flow at the Hanford Site. The ground truth transmissivity field is taken from a previous calibration study reported in [44]. Following [6], we employ an unstructured quadrilateral mesh with $N = 1475$ cells to compute PDE residuals.

Fig. 1a shows the ground truth $y(\mathbf{x})$ field. To retain 95% of the variance of this field, a 1000-term cKLE ($N_\xi = 1000$) is needed. We denote the ground truth y field as the high-dimensional or y_{HD} field. In the following, we will show that such high dimensionality in combination with relatively small σ_r reduces the efficiency of HMC, resulting in prohibitively large computational costs.

To enable a comparison between rPICKLE and baseline approximate Bayesian inference methods (HMC and SVGD), we generate a lower-dimensional (smoother) y field y_{LD} via iterative local averaging of y_{HD}. Such averaging reduces the variance and increases the correlation length of the field [45] and, therefore, reduces the dimensionality of the inverse problem. At the kth iteration, for the ith FV element, we calculate the geometric mean of $y_j^{(k)}$ over the adjacent j elements and assign this value to $y_i^{(k+1)}$. Here, $y_{\mathrm{HD}} = y^{(k=0)}$ and $y_{\mathrm{LD}} = y^{(k=30)}$. We find that a 10-term cKLE of y_{LD} retains 95% of the total variance of this field.

We generate the hydraulic head fields u_{HD} and u_{LD} corresponding to the y_{HD} and y_{LD} fields by solving the Darcy flow equation with the Dirichlet and Neumann boundary conditions from the calibration study [44] using the previously described FV method. Fig. 1 shows the y_{HD} and y_{LD} fields and the corresponding u fields. In the PICKLE representation, we set $N_\eta = 1000$ and $N_\eta = 5$ in cKLEs of u_{HD} and u_{LD}, respectively, to retain no less than 95% of the total variance of the hydraulic head field.

The Hanford Site calibration study [44] provides coordinates of 558 wells, with some of these wells located within the same FV cells. In this work, we only allow a single well per cell resulting in 323 wells. We assume that N_u^{obs} observations of u and N_y^{obs}

Table 1
Values of σ^2 and l for different cases.

	σ^2	l
y_{LD}, $N_y^{\mathrm{obs}} = 10$	0.369	0.419
y_{HD}, $N_y^{\mathrm{obs}} = 50$	3.185	0.038
y_{HD}, $N_y^{\mathrm{obs}} = 100$	3.236	0.035
y_{HD}, $N_y^{\mathrm{obs}} = 200$	2.821	0.029

observations of y are available. The locations of these observations are randomly selected from the well locations. The measurements of y and u at the selected locations are drawn from the y_{HD} and u_{HD} fields for the high-dimensional case and the y_{LD} and u_{LD} fields for the low-dimensional case.

The HMC, SVGD, and rPICKLE algorithms are implemented in TensorFlow 2 and TensorFlow-Probability. These implementations are publicly available at https://github.com/geekyifei/Bayesian_PICKLE_Hanford.git. All simulations are performed using an Intel® Xeon® Gold 6230R workstation.

3.3. Prior mean and covariance models of y and u

Following [6], for the log-transmissivity field $y(\mathbf{x})$, we select the 5/2-Matérn type prior covariance kernel:

$$C_y(\mathbf{x}, \mathbf{x}'; \sigma, l) = \sigma^2 \left(1 + \sqrt{5}\frac{\|\mathbf{x} - \mathbf{x}'\|}{l} + \frac{5}{3}\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{l^2}\right) \exp\left(-\sqrt{5}\frac{\|\mathbf{x} - \mathbf{x}'\|}{l}\right), \tag{44}$$

where σ and l are the standard deviation and correlation length of $y(\mathbf{x})$ found by minimizing the negative marginal log-likelihood function of the $y(\mathbf{x})$ measurements [38]. In Table 1, we summarize the values of σ^2 and l for different examples in Section 4. We find that the iterative local averaging reduces the variance and increases the correlation length of the low-dimensional field.

Then, the cKLE of y is constructed by first computing the mean and covariance of y conditioned on the y measurements using GPR (Eqs. (A.1) and (A.2)), and then evaluating the eigenvalues and eigenfunctions by solving the eigenvalue problem (A.3).

Next, we generate N_{MC} number of realizations of the stochastic y field by independently sampling $\{\xi_i\}_{i=1}^{N_\xi}$ from the normal distribution and solving Eqs. (41)–(43) for each realization of y using the FV method described above. The ensemble of u solutions is used to compute the (ensemble) mean and covariance of u using Eqs. (A.4) and (A.5). Then, the mean and covariance of u are conditioned on u observations using the GPR equations. Finally, the cKLE of u in Eq. (3) is constructed by performing the eigenvalue decomposition of the conditional covariance of u, i.e., by solving the eigenvalue problem similar to Eq. (A.3). In this work, we set $N_{\mathrm{MC}} = 5,000$.

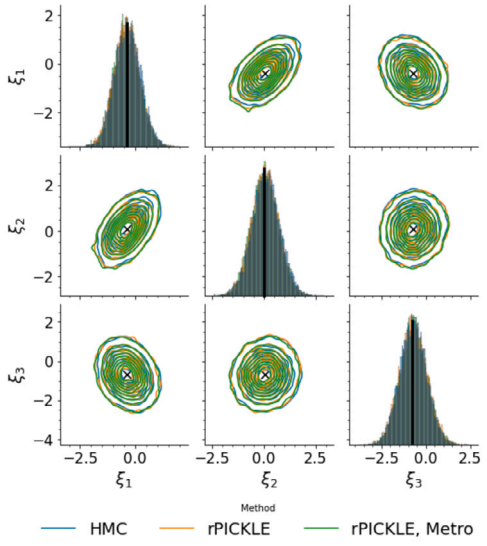
4. Numerical results

4.1. Low-dimensional problem

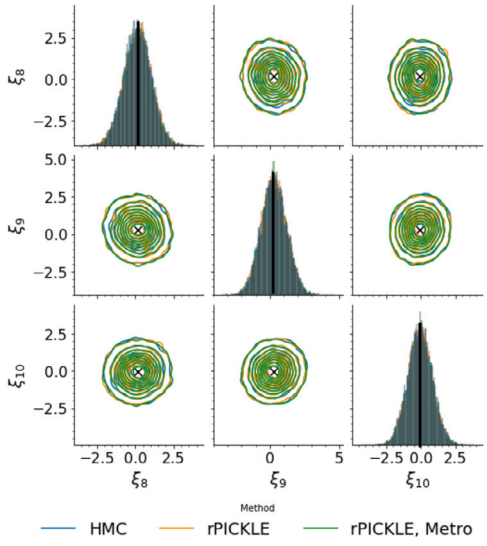
We first present results for the low-dimensional case with the ground truth y field given by y_{LD}. We assume that 10 y and u observations are available. The locations of the observations are randomly selected from the well locations. We use the posterior distributions sampled via HMC and SVGD, and the PICKLE-estimated MAP to benchmark the rPICKLE method with and without Metropolization. As stated earlier, we set $\sigma_\xi^2 = \sigma_\eta^2 = 1$. The joint problem of minimizing the rPICKLE loss over $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ and maximizing the LPP of the true field over σ_r^2 is computationally challenging. Instead, we compute the inverse solutions for several values of σ_r^2 in the range $[10^{-5}, 1]$ and select the σ_r^2 producing the largest LPP. We also study the effect of σ_r^2 on uncertainty in the inverse solution and the performance of HMC, SVGD, and rPICKLE.

For rPICKLE, we compute $N_{\mathrm{ens}} = 10^4$ samples by solving the rPICKLE minimization problem (13) for N_{ens} different realizations of $\boldsymbol{\omega}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$. In Section 4.3, we show that this number of samples is sufficient for the first two moments of the rPICKLE-sampled posterior distribution to converge. For HMC, we initialize three dispersed Markov chains over the parameter space. We set the number of HMC burn-in steps to 2×10^4 and the number of samples to $N_{\mathrm{ens}} = 10^4$ as the stopping criterion. We employ the No-U-Turn Sampler (NUTS) HMC method [46], which adaptively determines the number of integration steps taken within one HMC iteration. Furthermore, we use the dual averaging algorithms [46] to determine the optimal step size for NUTS to maintain a reasonable acceptance rate. Following [47], we set the target acceptance rate to 70%. For SVGD, we randomly initialize $N_{\mathrm{ens}} = 2 \times 10^3$ particles from an isotropic normal distribution. Here, we must use a smaller ensemble size than in rPICKLE and HMC due to memory constraints during SVGD computations, which require operations with matrices of dimension equal to the ensemble size. The Adam optimizer [48] is used to update the particle positions with the learning rate set to 1×10^{-3}. The radial basis function is used as the kernel function, with correlation length determined automatically during sampling following the so-called median trick [23].

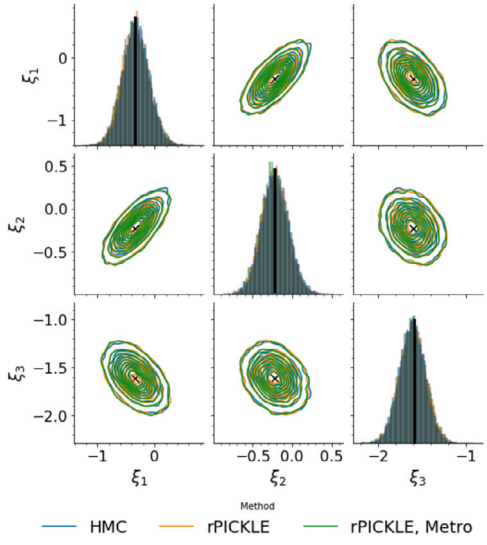
Fig. 2 depicts the marginal and bivariate distributions of the first and last three components of the $\boldsymbol{\xi}$ vector computed from HMC and rPICKLE with and without Metropolization using kernel density estimation (KDE) for $\sigma_r^2 = 10^{-4}$ and 10^{-2}. It can be seen that the distributions produced by the three different methods are similar. We also find that the marginal and bivariate distributions are approximately symmetric. For $\sigma_r^2 = 10^{-4}$, the bivariate distributions are narrower than for $\sigma_r^2 = 10^{-2}$, i.e., stronger physical constraints



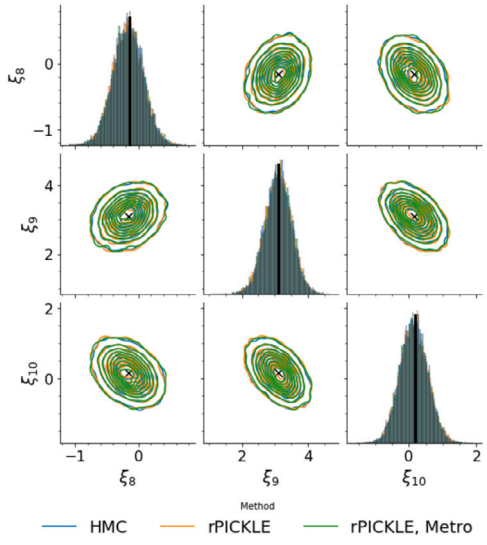
(a) $\sigma_r^2 = 10^{-2}$, ξ_1 - ξ_3



(b) $\sigma_r^2 = 10^{-2}$, ξ_8 - ξ_{10}



(c) $\sigma_r^2 = 10^{-4}$, ξ_1 - ξ_3



(d) $\sigma_r^2 = 10^{-4}$, ξ_8 - ξ_{10}

Fig. 2. Bivariate joint and marginal distributions of the first (first column) and last (second column) three components of ξ obtained from HMC, rPICKLE, and rPICKLE with Metropolization for the low-dimensional case. The black cross symbols and lines indicate the coordinates of the mode of the joint posterior distribution computed from PICKLE. The top and bottom rows show results for $\sigma_r^2 = 10^{-2}$ and 10^{-4}, respectively.

result in more certain predictions. Also, we see that for smaller σ_r^2, the bivariate distributions exhibit stronger correlations between the ξ components. These correlations are much stronger for the first three components of ξ than for the last three components.

Fig. 2 also shows the coordinates of the joint posterior mode given by the PICKLE solution. The coordinates of the modes of the marginal and bivariate distributions obtained from HMC and rPICKLE are close to the coordinates of the joint distribution mode. It should be noted that unless the posterior is Gaussian, the coordinates of the modes of marginal distributions and the corresponding coordinates of the joint distribution mode may not coincide.

We proceed to discuss the Bayesian estimates of y. For $\sigma_r^2 = 10^{-2}$, Fig. 3 shows the estimated mean $\mu_{\hat{y}}(\mathbf{x}|\mathcal{D}_{res})$ fields, the absolute differences between the mean and the reference field $|y - \mu_{\hat{y}}(\mathbf{x}|\mathcal{D}_{res})|$, the standard deviations $\sigma_{\hat{y}}(\mathbf{x}|\mathcal{D}_{res})$, and the coverage plots obtained from the four methods. The coverage plots indicate locations where the reference solution falls within the predicted 95% credibility interval, with zero and one values representing whether the reference field is outside or inside the credibility interval, respectively. We summarize in Table 2 the results obtained from the four methods in terms of the relative ℓ_2 error and ℓ_∞ error between the predicted mean and the reference y_{LD} fields, the LPP of the reference field, and the percentage of coverage (percentage

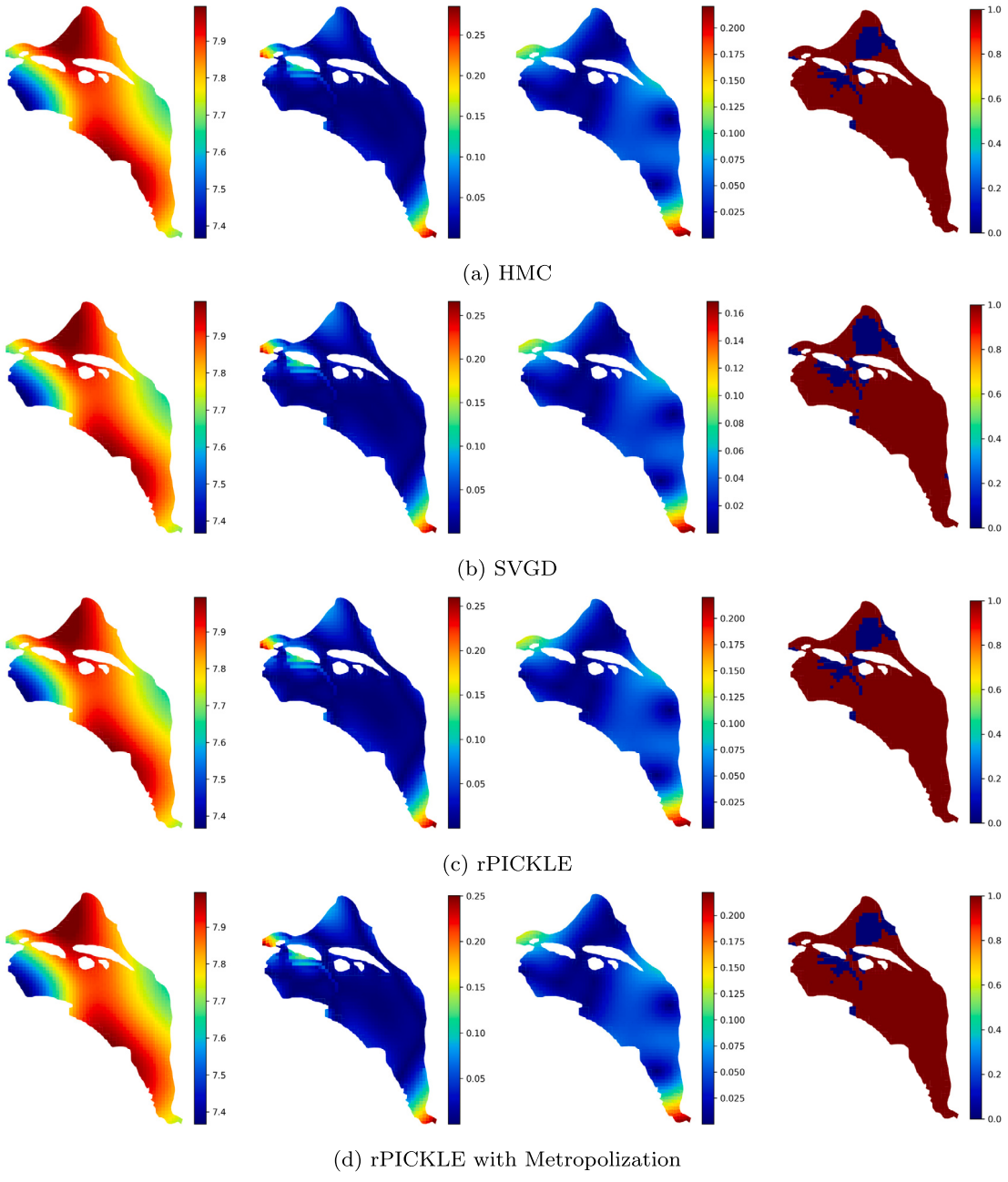


Fig. 3. The low-dimensional y field estimated from (a) HMC, (b) SVGD, (c) rPICKLE, and (d) rPICKLE with Metropolization for $\sigma_r^2 = 10^{-2}$: (first column) the posterior mean estimates; (second column) the point errors in the predicted mean with respect to the reference y_{LD} field; (third column) the posterior standard deviation of y; and (fourth column) the coverage of the reference field with the 95% credibility interval.

of nodes where the reference solution is within the 95% credibility interval) for $\sigma_r^2 = 10^{-5}$, 10^{-4}, 10^{-2}, 10^{-1}, and 10^{0}. For a given value of σ_r^2, HMC and rPICKLE produce posterior mean estimates close to the PICKLE MAP estimate, and similar LPPs and coverages. This indicates that (i) the HMC and rPICKLE sampled distributions converge to the same posterior (note that in Section 2.4.1, we only prove the convergence of rPICKLE in the linear case), and (ii) Metropolization makes the posterior more descriptive of the reference field, i.e., it yields larger LPP and smaller ℓ_2 error, but the improvements are less than 1%. We also find that, for this problem, the total acceptance rate for Metropolization is above 95%, which agrees with the findings in [49,30] that the acceptance rate in randomized algorithms is, in general, very high. It should be noted that Metropolization requires the estimation of a Jacobian determinant that becomes computationally expensive for high-dimensional problems. For this reason, in the high-dimensional case presented in Section 4.2 we do not perform Metropolization and accept all samples generated by the rPICKLE algorithm. SVGD also

Table 2
Summary of rPICKLE, Metropolized rPICKLE (Metro-rPICKLE), HMC, SVGD, and PICKLE results for the low-dimensional problem with different values of σ_r^2. Shown are the relative ℓ_2 and ℓ_∞ errors in the estimated y field with respect to y_{LD}, the LPP, and the percentage of coverage of y_{LD} by the 95% confidence interval. Note that the PICKLE solution only provides MAP and LPP.

σ_r^2	Method	ℓ_2 error	ℓ_∞ error	LPP	Coverage
10^0	PICKLE	8.05×10^{-3}	3.14×10^{-1}	–	–
	HMC	8.20×10^{-3}	3.15×10^{-1}	6972	92%
	SVGD	8.11×10^{-3}	3.14×10^{-1}	6523	85%
	rPICKLE	8.08×10^{-3}	3.14×10^{-1}	6971	92%
	Metro-rPICKLE	7.92×10^{-3}	3.15×10^{-1}	6984	92%
10^{-1}	PICKLE	7.28×10^{-3}	3.06×10^{-1}	–	–
	HMC	7.55×10^{-3}	3.07×10^{-1}	7206	90%
	SVGD	7.32×10^{-3}	3.06×10^{-1}	6935	86%
	rPICKLE	7.35×10^{-3}	3.08×10^{-1}	7226	90%
	Metro-rPICKLE	7.17×10^{-3}	3.04×10^{-1}	7240	90%
10^{-2}	PICKLE	5.90×10^{-3}	2.50×10^{-1}	–	–
	HMC	6.33×10^{-3}	2.85×10^{-1}	7561	89%
	SVGD	6.21×10^{-3}	2.67×10^{-1}	7196	86%
	rPICKLE	6.03×10^{-3}	2.60×10^{-1}	7564	89%
	Metro-rPICKLE	5.91×10^{-3}	2.50×10^{-1}	7546	89%
10^{-4}	PICKLE	7.14×10^{-3}	3.30×10^{-1}	–	–
	HMC	7.25×10^{-3}	3.40×10^{-1}	−21314	41%
	SVGD	7.26×10^{-3}	3.41×10^{-1}	−21947	40%
	rPICKLE	7.17×10^{-3}	3.33×10^{-1}	−20725	41%
	Metro-rPICKLE	7.17×10^{-3}	3.32×10^{-1}	−20232	41%
10^{-5}	PICKLE	9.36×10^{-3}	4.62×10^{-1}	–	–
	HMC	9.36×10^{-3}	4.63×10^{-1}	−275219	16%
	SVGD	9.37×10^{-3}	4.64×10^{-1}	−283043	16%
	rPICKLE	9.37×10^{-3}	4.62×10^{-1}	−277842	16%
	Metro-rPICKLE	9.37×10^{-3}	4.62×10^{-1}	−282418	16%

results in a posterior mean close to the PICKLE MAP estimate. However, it yields slightly smaller LPP and coverage than HMC and rPICKLE.

Table 2 shows that the point errors and the relative ℓ_2 errors are the smallest across all methods when $\sigma_r^2 = 10^{-2}$. We find that the LPPs are also the largest for this value of σ_r^2. This indicates that $\sigma_r^2 = 10^{-2}$ provides the posterior that best describes the true field. We also note that the PICKLE ℓ_2 and ℓ_∞ errors are also smallest for $\gamma = \sigma_r^2 = 10^{-2}$, which indicates this value also provides the optimal regularization for this problem.

Finally, we find that in this low-dimensional problem, the runtime of rPICKLE (i.e., the time to obtain a solution of the rPICKLE minimization problem) is independent of the value of σ_r^2. For all considered values of σ_r^2, the runtime per sample was approximately 0.02 seconds. On the other hand, the HMC runtime is found to increase with decreasing σ_r^2 from 1.12 seconds per sample for $\sigma_r^2 = 1$ to 2.01 for $\sigma_r^2 = 10^{-5}$. The SVGD runtime increases with decreasing σ_r^2 from 2.33 seconds per sample for $\sigma_r^2 = 1$ to 4.35 for $\sigma_r^2 = 10^{-5}$.

4.2. High-dimensional problem

Here, we present results for the high-dimensional case where the reference y field is given by y_{HD}. First, we estimate the posteriors for different values of σ_r^2 (10^{-4}, 10^{-2}, and 10^{-1}) given 100 observations of the y_{HD} field. Then, we estimate posteriors for different values of N_y^{obs} (50, 100, and 200) given $\sigma_r^2 = 10^{-4}$ to study the dependence of the posterior on the number of y measurements. In all examples in this section, we assume that $N_u^{\mathrm{obs}} = 323$, i.e., u measurements are available at all wells.

Based on the results in the previous section, we do not perform Metropolis rejection in rPICKLE. We find that for this high-dimensional problem, the HMC step size, computed from the dual averaging step size adaptation algorithm (which is designed to maintain a prescribed acceptance rate) becomes extremely small. As a result, for some values of σ_r^2, the HMC implementation fails to reach the stopping criterion after running for more than 30 days. For comparison, rPICKLE generates the same number of samples in 4 to 5 days depending on σ_r^2 (around 30–40 seconds per sample). The posterior obtained from SVGD is very close to the (Gaussian) prior distribution, and its mean deviates significantly from the PICKLE MAP estimate. Therefore, for the high-dimensional case, we only present the rPICKLE and PICKLE results. We attribute HMC's large computational time to the high condition number of the posterior covariance matrix, which we find to increase with increasing dimensionality and decreasing σ_r^2. We investigate this dependence in detail in Section 4.4.

Table 3 summarizes the relative ℓ_2 and ℓ_∞ errors, LPP, and coverage of the rPICKLE estimates of y for $\sigma_r^2 = 10^{-4}$, 10^{-2}, and 10^{-1}. The smallest rPICKLE errors and the largest LPP are achieved for $\sigma_r^2 = 10^{-2}$. However, we find that the LPP is more sensitive to σ_r^2 than the ℓ_2 error–ℓ_2 errors vary by less than 7% for the considered σ_r^2 values, while LPP values change by more than 100%. Also,

Table 3
Summary of rPICKLE and PICKLE results for the high-dimensional problem for the priors based on different numbers of y observations and σ_r^2. Shown are the relative ℓ_2 and ℓ_∞ errors in the estimated y field with respect to the reference field, the LPP, and the percentage of coverage of y_{LD} by the 95% confidence interval.

N_y^{obs}	Method	$r\ell_2$ error	ℓ_{inf} error	LPP	Coverage
	$\sigma_r^2 = 10^{-1}$				
100	PICKLE	1.08×10^{-1}	4.38×10^{0}	–	–
	rPICKLE	1.11×10^{-1}	4.41×10^{0}	1730	87%
	$\sigma_r^2 = 10^{-2}$				
100	PICKLE	1.01×10^{-1}	4.14×10^{0}	–	–
	rPICKLE	1.04×10^{-1}	4.13×10^{0}	2032	82%
	$\sigma_r^2 = 10^{-4}$				
50	PICKLE	1.91×10^{-1}	6.49×10^{0}	–	–
	rPICKLE	1.71×10^{-1}	6.50×10^{0}	−2792	65%
100	PICKLE	1.26×10^{-1}	6.14×10^{0}	–	–
	rPICKLE	1.10×10^{-1}	5.33×10^{0}	1070	67%
200	PICKLE	7.91×10^{-2}	5.34×10^{0}	–	–
	rPICKLE	7.99×10^{-2}	5.38×10^{0}	5220	75%

the ℓ_2 error is 1% smaller for $\sigma_r^2 = 10^{-4}$ than for $\sigma_r^2 = 10^{-1}$ but the LPP is 70% larger for $\sigma_r^2 = 10^{-1}$ than for $\sigma_r^2 = 10^{-4}$. Errors in the PICKLE and rPIKCLE predictions of y are very similar, and $\sigma_r^2 = 10^{-2}$ also provides the best value of the regularization coefficient for PICKLE, i.e., the PICKLE error is smallest for $\gamma = \sigma_r^2 = 10^{-2}$.

Fig. 4 depicts the marginal and bivariate distributions of the first and last three components of $\boldsymbol{\xi}$ for $\sigma_r^2 = 10^{-1}$ and 10^{-2}. Compared with the low-dimensional case, we observe that the posterior of the first three components is more non-symmetric and correlated. The posterior distributions become narrower as σ_r^2 becomes smaller. The modes of these distributions have non-zero coordinates (while the prior distributions are centered at zero). On the other hand, the last three terms have symmetric marginal distributions approximately centered at zero and circular-shaped bi-variate distributions, the latter indicating the lack of cross-correlation. Fig. 4 also shows the coordinates of the joint posterior mode obtained from PICKLE. They slightly deviate from the coordinates of the marginal and bivariate modes because of the non-Gaussianity of the posterior.

Fig. 5 shows the rPICKLE estimate of the posterior mean of y, the absolute point difference between the mean and reference $y_{\mathrm{HD}}(x)$ fields, the posterior standard deviation of $y(x)$, and the coverage for $\sigma_r^2 = 10^{-2}$ and $\sigma_r^2 = 10^{-4}$. We see significant differences in rPICKLE predictions for different σ_r^2. Errors in the predictions with $\sigma_r^2 = 10^{-4}$ are in general larger than in the predictions with $\sigma_r^2 = 10^{-2}$ with the maximum point error being 50% larger. As expected, the posterior standard deviations are generally larger in the prediction with the larger σ_r^2. However, the maximum pointwise standard deviation is larger for the smaller σ_r^2. We also see that $\sigma_r^2 = 10^{-2}$ produces a better coverage–the reference y_{HD} field is within the confidence interval in 82% of all predicted locations versus 65% for $\sigma_r^2 = 10^{-2}$. We also note that the LPP for $\sigma_r^2 = 10^{-4}$ is -2792, which is significantly smaller than that for $\sigma_r^2 = 10^{-2}$ (2032).

Next, we study the inverse rPICKLE solution as a function of N_y^{obs}. Fig. 6 shows the estimates of y obtained with rPICKLE for $N_y^{\mathrm{obs}} = 50$ and 200 and $\sigma_r^2 = 10^{-4}$. The estimates for $N_y^{\mathrm{obs}} = 100$ are given in Fig. 5. Table 3 summarizes the relative ℓ_2 and ℓ_∞ errors, LPP, and the percent of coverage of the corresponding posteriors. As N_y^{obs} increases, the posterior mean becomes closer to y_{HD}, and the posterior variance of y decreases. The LPP increases with N_y^{obs}, indicating that the posterior distribution better fits the true field. Also, we see that for all values of N_y^{obs} the coverage for the y_{HD} field is adequate, with the best coverage (75%) achieved for $N_y^{\mathrm{obs}} = 200$.

4.3. Convergence of rPICKLE and HMC with the ensemble size

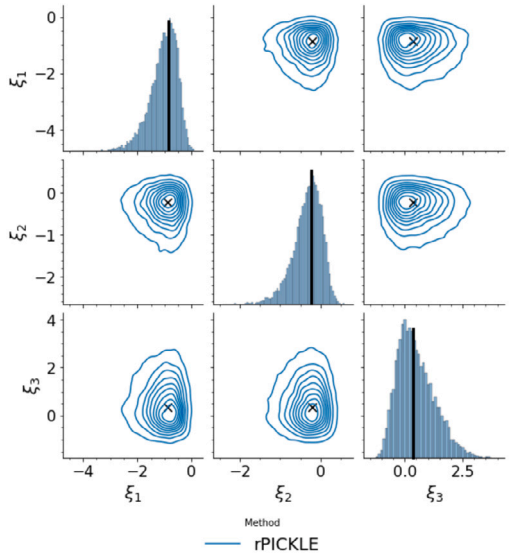
Next, we examine the convergence properties of rPICKLE and HMC estimates with increasing N_{ens} for the low- and high-dimensional cases. For the i-th component of the $\boldsymbol{\xi}$ vector, we analyze the relative mean $\overline{\xi}_i^b(m)$ ($b =$ rPICKLE or HMC):

$$R_{\overline{\xi}_i^b}(m) = \left| \frac{\overline{\xi}_i^b(m)}{\overline{\xi}_i^b(N_{\mathrm{ens}})} \right| \tag{45}$$

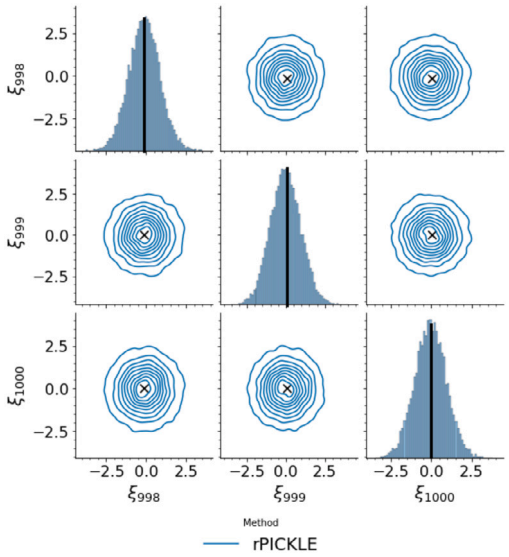
and relative standard deviation, $\sigma_{\xi_i}^b(m)$:

$$R_{\sigma_{\xi_i}^b}(m) = \left| \frac{\sigma_{\xi_i}^b(m)}{\sigma_{\xi_i}^b(N_{\mathrm{ens}})} \right|, \tag{46}$$

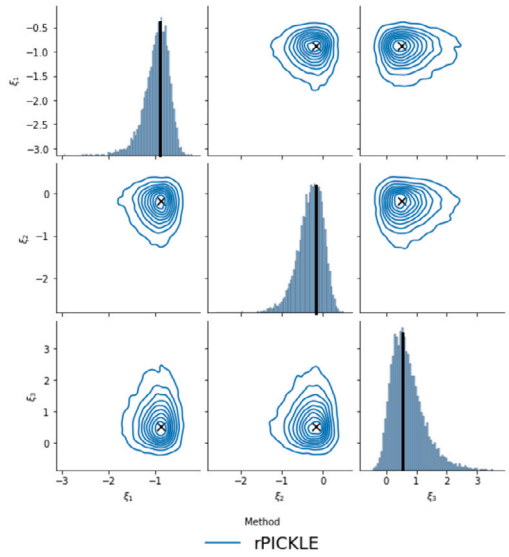
as functions of the ensemble size m ($N_{\mathrm{ens}} = 10^4$ is the maximum ensemble size). Fig. 7 shows the dependence of $R_{\overline{\xi}_i^b}(m)$ and $R_{\sigma_{\xi_i}^b}(m)$ on m for the low-dimensional case ($i = 1$ and 10) and the high-dimensional case ($i = 1$ and 100). Here, we set $\sigma_r^2 = 10^{-2}$ and 10^{-4}.



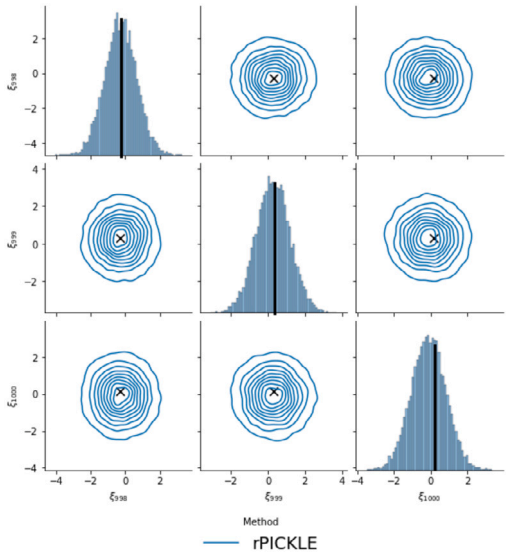
(a) $\sigma_r^2 = 10^{-1}$, ξ_1 - ξ_3



(b) $\sigma_r^2 = 10^{-1}$, ξ_{998} - ξ_{1000}



(c) $\sigma_r^2 = 10^{-2}$, ξ_1 - ξ_3



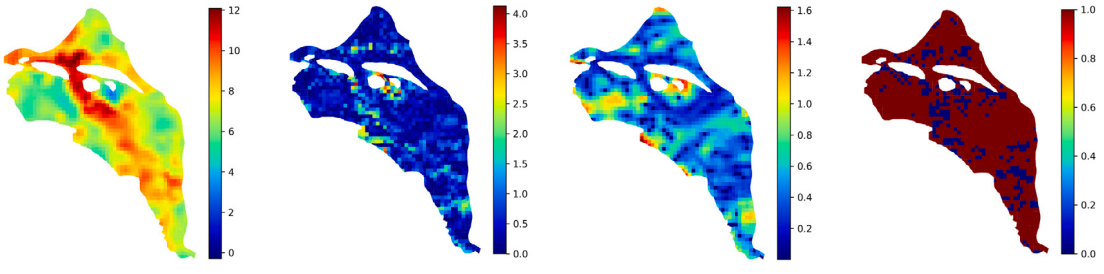
(d) $\sigma_r^2 = 10^{-2}$, ξ_{998} - ξ_{1000}

Fig. 4. Bivariate joint and marginal distributions of the first (first column) and last (second column) three components of ξ obtained from rPICKLE for the high-dimensional case. The black cross symbols and lines indicate the coordinates of the mode of the joint posterior distribution computed from PICKLE. The top and bottom rows show results for $\sigma_r^2 = 10^{-1}$ and 10^{-2}, respectively.

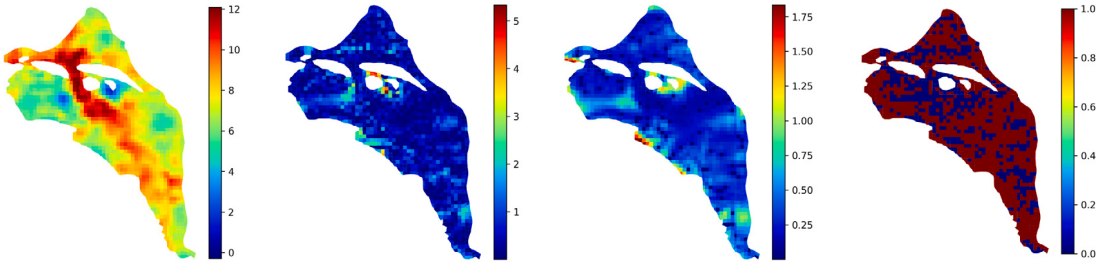
For the high-dimensional case, the number of y observations for the high-dimensional case is set to $N_y^{\text{obs}} = 100$, and we only show the convergence of rPICKLE because of the prohibitively large computational time of HMC.

For the low-dimensional case, Fig. 7 shows that the convergence properties of HMC and rPICKLE are similar. We also find that in both methods, the required number of samples for mean and variance to reach asymptotic values increases with σ_r. Furthermore, we see that in rPICKLE, the required number of samples is not significantly affected by the dimensionality, which is to be expected because the rPICKLE samples are generated independently.

It should be noted that, as shown in Section 4.4, the condition number of the posterior covariance increases with decreasing σ_r. The increasing condition number decreases the time step in the HMC algorithm. As a result, we find that the computational time of HMC to get a set number of samples increases with decreasing σ_r. On the other hand, the computational time of rPICKLE is not significantly affected by the value of σ_r.

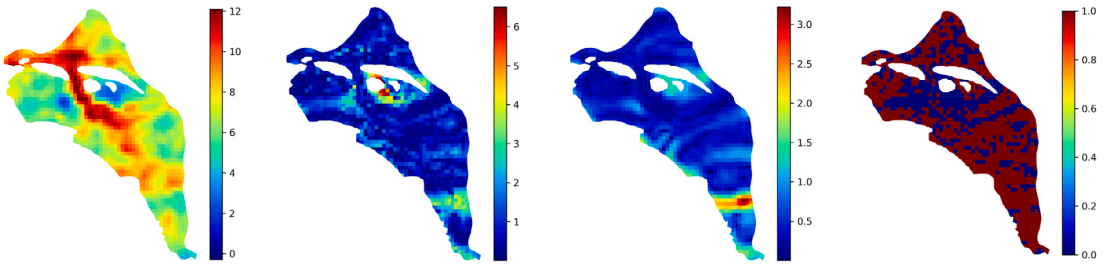


(a) $\sigma_r^2 = 10^{-2}$

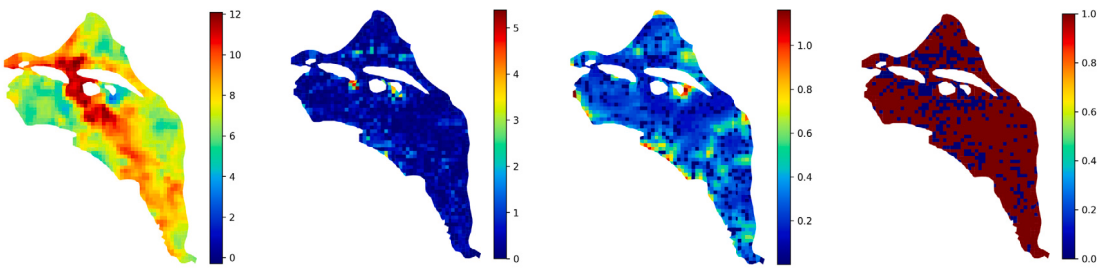


(b) $\sigma_r^2 = 10^{-4}$

Fig. 5. rPICKLE estimates of y_{HD} with (a) $\sigma_r^2 = 10^{-2}$ and (b) $\sigma_r^2 = 10^{-4}$ given $N_y^{\text{obs}} = 100$ observations of y: (first column) the posterior mean of y, (second column) point errors computed as the difference between the posterior mean of y and the reference y field, (third column) the posterior standard deviation of y, and (fourth column) the coverage of the reference y by the 95% credibility interval.

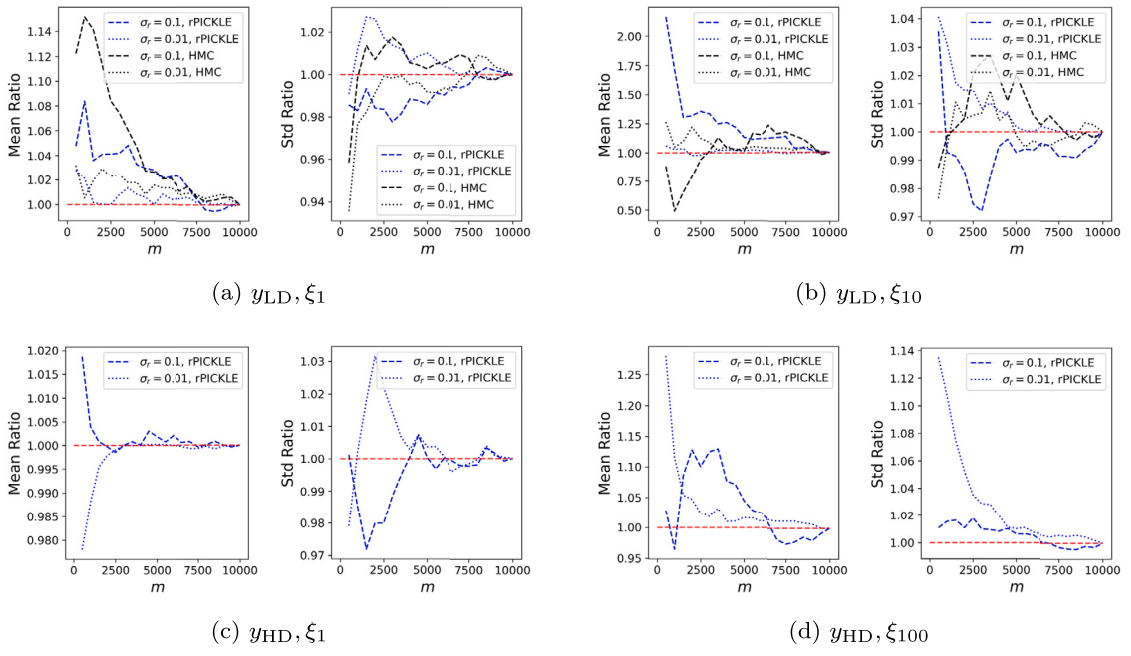


(a) $\text{N}_\text{y}^{\text{obs}} = 50$



(b) $\text{N}_\text{y}^{\text{obs}} = 200$

Fig. 6. rPICKLE estimates of the high-dimensional y field obtained with $\sigma_r^2 = 10^{-4}$ given (a) $N_y^{\text{obs}} = 50$ and (b) $N_y^{\text{obs}} = 200$ observations of y: (first column) the posterior mean of y, (second column) point errors computed as the difference between the posterior mean of y and the reference y field, (third column) the posterior standard deviation of y, and (fourth column) the coverage of the reference y by the 95% credibility interval.



(a) y_{LD}, ξ_1 (b) $y_{\mathrm{LD}}, \xi_{10}$

(c) y_{HD}, ξ_1 (d) $y_{\mathrm{HD}}, \xi_{100}$

Fig. 7. $R_{\overline{\xi}_i^k}(m)$ and $R_{\sigma^k_{\xi_i}}(m)$ as functions of the ensemble size m for the low- and high-dimensional cases and $\sigma_r^2 = 10^{-4}$ and $\sigma_r^2 = 10^{-2}$: (a) ξ_1 of y_{LD}, (b) ξ_{10} of y_{LD}, (c) ξ_1 of y_{HD}, and (d) ξ_{100} of y_{LD}. HMC results are not available for the high-dimensional case.

4.4. HMC performance for low- and high-dimensional problems

In this section, we investigate the effect of σ_r^2 and problem dimensionality on the efficiency of HMC.

As mentioned earlier, for certain values of σ_r^2, the HMC does not reach the stopping criterion (10^4 samples) after more than 30 days of running the code. For comparison, rPICKLE generates 10^4 samples in 4 to 5 days depending on the value of σ_r^2. For the low-dimensional case, rPICKLE takes approximately 4 minutes to generate the same number of samples for all considered σ_r^2 while the computational time of HMC varies from 3 to 6 hours depending on the σ_r^2 value.

The increase in HMC computational time is mainly due to smaller time steps in the Hamiltonian dynamics equation integration required to maintain a desirable acceptance rate in the dual averaging algorithm. It was shown in [50,28] that the large condition number of the posterior covariance matrix leads to decreasing HMC performance. Here, we demonstrate that the condition number increases with increasing problem dimension ($M = N_\xi + N_\eta$) and decreasing σ_r^2.

In theory, the posterior covariance can be computed directly from posterior samples obtained, for example, from rPICKLE. Here, we focus on a priori estimates of the posterior covariance that can be used as a criterion for using HMC. To obtain an a priori estimate of Σ_{post}, we employ the Laplace approximation and approximate the posterior covariance via the inverse of the Hessian of the log posterior. We start by approximating the log-posterior (6) using the Taylor expansion around the MAP (which we assume is known from PICKLE) as

$$\log P(\boldsymbol{\zeta}|\mathcal{D}_{res}) \approx \log P(\boldsymbol{\zeta}^*|\mathcal{D}_{res}) + \frac{1}{2}(\boldsymbol{\zeta} - \boldsymbol{\zeta}^*)^T (\nabla\nabla \log P(\boldsymbol{\zeta}|\mathcal{D}_{res})|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}^*})(\boldsymbol{\zeta} - \boldsymbol{\zeta}^*), \tag{47}$$

where $\boldsymbol{\zeta}^*$ is the MAP, and $\nabla\nabla \log P(\boldsymbol{\zeta}|\mathcal{D}_{res})|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}^*}$ is the Hessian, which we compute by automatic differentiation, evaluated at the MAP. The first-order term in Eq (47) vanishes because the $\log P$ gradient at the MAP point is zero. The right-hand side of Eq. (47) is equivalent, up to a constant, to the log probability density of a Gaussian distribution. Under the Laplace assumption, the posterior distribution can be approximated with a Gaussian distribution, the mean of which is given by the MAP. The covariance is found by the inverse of the Hessian evaluated at the MAP point, i.e., $(\nabla\nabla \log P(\boldsymbol{\zeta}|\mathcal{D}_{res})|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}^*})^{-1}$.

Fig. 8 shows the eigenvalues (arranged in descending order) of the approximated posterior covariance for the low- and high-dimensional cases with $\sigma_r^2 = 10^{-2}$ and 10^{-4}. The red dashed lines indicate eigenvalues of the prior covariance (all eigenvalues are equal to one because of the diagonal form of the prior covariance and unit prior variances of the parameters). In the low-dimensional problem, the condition numbers are approximately 21 for $\sigma_r^2 = 10^{-2}$ and 1760 for $\sigma_r^2 = 10^{-4}$. In the high-dimensional problem, the condition numbers are $\approx 2 \times 10^7$ for $\sigma_r^2 = 10^{-2}$ and 2×10^9 for $\sigma_r^2 = 10^{-4}$.

Larger condition numbers indicate the presence of a stronger correlation between the components of the estimated $\boldsymbol{\xi}$. Also, the eigenvalues λ_i of Σ_{post} are proportional to the variances σ_i^2 of the posterior marginal distributions of the ξ_i components of $\boldsymbol{\xi}$. Therefore, a large condition number indicates a large range of σ_i^2 values, which gives rise to geometrically pathological features of the posterior

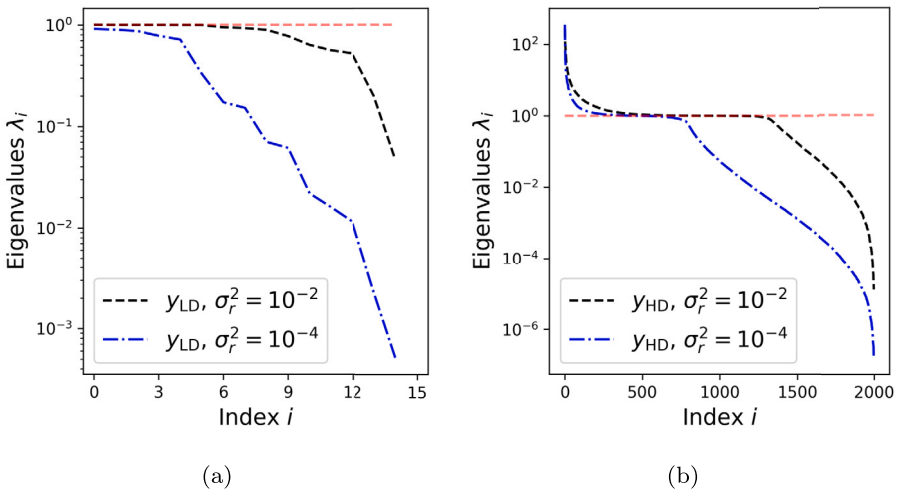


Fig. 8. Covariance spectrum corresponding to the Laplace approximated posterior for (a) y_{LD} and (b) y_{HD} for $\sigma_r^2 = 10^{-4}$ and 10^{-2}. The red dashed line represents the covariance structure informed by the normal prior.

parameter space (e.g., high curvature in the corners of equal-density posterior manifolds). The increase in correlation and the range of σ_i^2 with σ_r^2 and N_ξ can be seen in Figs. 2 and 4. These geometric complexities reduce the efficiency of HMC [29].

5. Discussions and conclusions

We presented the rPICKLE method for approximate sampling of high-dimensional Bayesian posterior distributions of unknown parameters. The rPICKLE is derived by randomizing the PICKLE objective function. In the Bayesian framework, the prior distribution expresses prior knowledge about the modeled system. In the application of rPICKLE to the diffusion (Darcy) equation with an unknown space-dependent diffusion coefficient (transmissivity of an aquifer), it is assumed that the log-diffusion coefficient $y(\mathbf{x})$ has a Gaussian prior with mean and covariance estimated from y measurements. This prior model determines the eigenfunctions in the cKLE expansion of y and the Gaussian prior distribution of the parameters $\boldsymbol{\xi}$ in this cKLE. We note that it is also possible to choose different prior distributions, such as the H^1 seminorm prior [51], for $\boldsymbol{\xi}$. The formulation of randomized methods for such priors will be the subject of future work.

The likelihood function is defined by assuming a Gaussian error model for the PDE residuals, with error variance σ_r^2. This variance then becomes a free parameter controlling the posterior distribution. We study the effect of σ_r^2 on the difference between the true field and the posterior mean and on the LPP of the true field. We demonstrated that PICKLE provides the mode of the posterior distribution if the PICKLE regularization coefficient γ is set to σ_r^2. Because the PICKLE-based mode estimation does not require posterior sampling, choosing $\sigma_r^2 = \gamma$ to minimize the distance from the posterior mode to the true field is computationally straightforward. Generally, there is no guarantee that σ_r^2 selected according to this criterion would also maximize the LPP. However, we found that for considered problems, σ_r^2 minimizing the distance between the posterior mean and mode also maximizes the LPP of the posterior distribution. We also found that LPP is more sensitive to σ_r^2 than the distance between the mode and data. Therefore, LPP should also be considered when selecting σ_r^2.

The robustness of rPICKLE was demonstrated by estimating the log-transmissivity field of a Hanford Site groundwater flow model with 2000-dimensional cKLE representations of the parameter and state variables. We found that rPICKLE produces posteriors with the mean close to the PICKLE MAP estimate for a wide range of values of σ_r^2. On the other hand, HMC did not reach the stopping criterion (10^4 samples) after running for more than 1 month. For comparison, our rPICKLE implementation generated the same number of samples in 4 to 5 days depending on the value of σ_r^2.

To compare rPICKLE and HMC, we considered a lower-dimensional problem where the parameters and states were represented with 10-dimensional cKLE parameters. For this problem, we found an excellent agreement between rPICKLE and HMC. We also found that the y predictions given by the posterior mean of rPICKLE and HMC are close to the PICKLE MAP estimate. For this low-dimensional problem, rPICKLE generated 10^4 samples in approximately 4 minutes regardless of the considered values of σ_r^2. The HMC time was found to increase from $\approx$ 3 hours for $\sigma_r^2 = 1$ to $\approx$ 6 hours for $\sigma_r^2 = 10^{-5}$. The efficiency of HMC is known to decrease with the increasing condition number of the posterior covariance matrix. We demonstrated that for the considered problem, the condition number increases with increasing dimensionality and decreasing σ_r^2, which explains the observed trend in HMC computational time. In summary, our results demonstrate the advantage of rPICKLE for high-dimensional problems with strict physics constraints (small values of σ_r^2).

CRediT authorship contribution statement

Yifei Zong: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis. **David Barajas-Solano:** Writing – original draft, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Alexandre M. Tartakovsky:** Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This research was partially supported by the U.S. Department of Energy (DOE) Advanced Scientific Computing program, the DOE project "Science-Informed Machine Learning to Accelerate Real-time (SMART) Decisions in Subsurface Applications Phase 2 – Development and Field Validation," and the United States National Science Foundation, grant 2141503. Pacific Northwest National Laboratory is operated by Battelle for the DOE under Contract DE-AC05-76RL01830.

Appendix A. Gaussian process regression and the conditional covariance

Given observations $\{y_i^{\text{obs}}\}_{i=1}^{N_y^{\text{obs}}}$, we estimate the parameters in the covariance function $C_y(\mathbf{x},\mathbf{x}')$ in Eq. (44) by minimizing the negative marginal log-likelihood function [38]. Then, we set the prior distribution of $y(x)$ to $\mathcal{GP}(\overline{y}^c(\mathbf{x}), C_y^c(\mathbf{x},\mathbf{x}'))$, where the mean $\overline{y}^c(\mathbf{x})$ and covariance $C_y^c(\mathbf{x},\mathbf{x}')$ are computed using the Gaussian process regression (GPR) equations:

$$\overline{y}^c(\mathbf{x}) = \mathbf{c}_y(\mathbf{x})\mathbf{C}_y^{-1}\mathbf{y}^{\text{obs}} \tag{A.1}$$

$$C_y^c(\mathbf{x},\mathbf{x}') = \mathbf{C}_y(\mathbf{x},\mathbf{x}') - \mathbf{c}_y(\mathbf{x})\mathbf{C}_y^{-1}\mathbf{c}_y(\mathbf{x}'). \tag{A.2}$$

Here, the superscript c denotes that Gaussian distribution is conditioned on the y measurements, $\mathbf{C}_y \in \mathbb{R}^{N_y^{\text{obs}}\times N_y^{\text{obs}}}$ is the observation covariance matrix with the elements $C_{y,(ij)} = C_y(\mathbf{x}_i^{\text{obs}}, \mathbf{x}_j^{\text{obs}})$, and $\mathbf{c}_y(\mathbf{x}) \in \mathbb{R}^{1\times N_y^{\text{obs}}}$ is the covariance vector with the elements $c_{y,i} = C_y(\mathbf{x}, \mathbf{x}_i^{\text{obs}})$.

The eigenfunctions and eigenvalues of Eq. (2) are obtained by solving the following eigenvalue problem:

$$\int C_y^c(\mathbf{x},\mathbf{x}')\psi_i^y(\mathbf{x}')d\mathbf{x}' = \lambda_i^y\psi_i^y(\mathbf{x}') \quad i = 1,...,N_\xi. \tag{A.3}$$

This eigenvalue problem is solved on the mesh used to generate the training dataset and compute residuals in the minimization problem. As such, the eigenvalue problem reduces to the eigendecomposition in the finite-dimensional vector space. In the KLE of y, we use N_ξ terms corresponding to N_ξ largest eigenvalues. The selection criterion for N_ξ is given in [6].

The mean $\overline{u}^c(\mathbf{x})$ and covariance $C_u^c(\mathbf{x},\mathbf{x}')$ in the KLE of u are computed as the solutions of the (stochastic) Eq. (1) with the stochastic y field whose mean and covariance are given by Eqs. (A.1) and (A.2). We do not use a parameterized covariance function to model u because the random field for u is generally not stationary and parameterized covariance functions do not enforce the physical constraint.

Here we use the Monte Carlo (MC) simulations to compute the mean and covariance of u. We randomly draw an ensemble of N_{MC} cKLE coefficients $\{\xi_i\}_{i=1}^{N_{\text{MC}}}$ from $\mathcal{N}(\mathbf{0},\mathbf{I})$ and use Eq. (2) to generate as many realizations of y. Then, for each realization y_i we solve Eq. (1) to get N_{MC} solutions $\mathbf{u}_i$ on the finite volume mesh. Then, the mean solution vector $\overline{\mathbf{u}}$ and the covariance matrix $\mathbf{C}_u$ of $\mathbf{u}$ are computed as

$$\overline{\mathbf{u}} = \frac{1}{N_{\text{MC}}}\sum_{i=1}^{N_{\text{MC}}}\mathbf{u}_i, \tag{A.4}$$

$$\mathbf{C}_u = \frac{1}{N_{\text{MC}}-1}\sum_{i=1}^{N_{\text{MC}}}[\mathbf{u}_i - \overline{\mathbf{u}}][\mathbf{u}_i - \overline{\mathbf{u}}]^T. \tag{A.5}$$

The GPR equations are used to condition the mean and covariance of $\mathbf{u}$ on the u measurements yielding $\overline{\mathbf{u}}^c(\mathbf{x})$ and $\mathbf{C}_u^c(\mathbf{x},\mathbf{x}')$. Finally, the eigenvalue and (discretized on the finite volume mesh) eigenfunctions are obtained via the eigenvalue decomposition of $\mathbf{C}_u^c(\mathbf{x},\mathbf{x}')$.

References

[1] H. Zhou, J.J. Gómez-Hernández, L. Li, Inverse methods in hydrogeology: evolution and recent trends, Adv. Water Resour. 63 (2014) 22–37.
[2] N. Linde, D. Ginsbourger, J. Irving, F. Nobile, A. Doucet, On uncertainty quantification in hydrogeology and hydrogeophysics, Adv. Water Resour. 110 (2017) 166–181.
[3] A.M. Tartakovsky, C.O. Marrero, P. Perdikaris, G.D. Tartakovsky, D. Barajas-Solano, Physics-informed deep neural networks for learning parameters and constitutive relationships in subsurface flow problems, Water Resour. Res. 56 (5) (2020) e2019WR026731.
[4] Q. He, A.M. Tartakovsky, Physics-informed neural network method for forward and backward advection-dispersion equations, Water Resour. Res. 57 (7) (2021) e2020WR029479.
[5] Y. Zong, Q. He, A.M. Tartakovsky, Improved training of physics-informed neural networks for parabolic differential equations with sharply perturbed initial conditions, Comput. Methods Appl. Mech. Eng. 414 (2023) 116125.
[6] Y.-H. Yeung, D.A. Barajas-Solano, A.M. Tartakovsky, Physics-informed machine learning method for large-scale data assimilation problems, Water Resour. Res. 58 (5) (2022), https://doi.org/10.1029/2021WR031023.
[7] M.P. Anderson, W.W. Woessner, R.J. Hunt, Applied Groundwater Modeling: Simulation of Flow and Advective Transport, Academic Press, 2015.
[8] B.S. RamaRao, A.M. LaVenue, G. De Marsily, M.G. Marietta, Pilot point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields: 1. Theory and computational experiments, Water Resour. Res. 31 (3) (1995) 475–493.
[9] J.E. Doherty, M.N. Fienen, R.J. Hunt, Approaches to Highly Parameterized Inversion: Pilot Point Theory, Guidelines, and Research Directions, US Department of the Interior, vol. 2010, US Geological Survey, 2010.
[10] M.J. Tonkin, J. Doherty, A hybrid regularized inversion methodology for highly parameterized environmental models, Water Resour. Res. 41 (10) (2005).
[11] Y.M. Marzouk, H.N. Najm, Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems, J. Comput. Phys. 228 (6) (2009) 1862–1902.
[12] J. Li, A.M. Tartakovsky, Gaussian process regression and conditional polynomial chaos for parameter estimation, J. Comput. Phys. (2020) 109520.
[13] A.M. Tartakovsky, D.A. Barajas-Solano, Q. He, Physics-informed machine learning with conditional Karhunen-Loève expansions, J. Comput. Phys. 426 (2021) 109904.
[14] A.M. Stuart, Inverse problems: a Bayesian perspective, Acta Numer. 19 (2010) 451–559.
[15] D. Herckenrath, C.D. Langevin, J. Doherty, Predictive uncertainty analysis of a saltwater intrusion model using null-space Monte Carlo, Water Resour. Res. 47 (5) (2011).
[16] H. Yoon, D.B. Hart, S.A. McKenna, Parameter estimation and predictive uncertainty in stochastic inverse modeling of groundwater flow: comparing null-space Monte Carlo and multiple starting point methods, Water Resour. Res. 49 (1) (2013) 536–553.
[17] S. Brooks, Markov chain Monte Carlo method and its application, J. R. Stat. Soc., Ser. D, Stat. 47 (1) (1998) 69–100.
[18] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U.R. Acharya, et al., A review of uncertainty quantification in deep learning: techniques, applications and challenges, Inf. Fusion 76 (2021) 243–297.
[19] L. Sun, J.-X. Wang, Physics-constrained Bayesian neural network for fluid flow reconstruction with sparse and noisy data, Theor. Appl. Mech. Lett. 10 (3) (2020) 161–169.
[20] R. Gou, Y. Zhang, X. Zhu, Bayesian physics-informed neural networks for seismic tomography based on the Eikonal equation, arXiv preprint, arXiv:2203.12351, 2022.
[21] D.M. Blei, A. Kucukelbir, J.D. McAuliffe, Variational inference: a review for statisticians, J. Am. Stat. Assoc. 112 (518) (2017) 859–877.
[22] J. Yao, W. Pan, S. Ghosh, F. Doshi-Velez, Quality of uncertainty quantification for Bayesian neural network inference, arXiv preprint, arXiv:1906.09686, 2019.
[23] Q. Liu, D. Wang, Stein variational gradient descent: a general purpose Bayesian inference algorithm, Adv. Neural Inf. Process. Syst. 29 (2016).
[24] Q. Liu, Stein variational gradient descent as gradient flow, Adv. Neural Inf. Process. Syst. 30 (2017).
[25] J. Ba, M.A. Erdogdu, M. Ghassemi, S. Sun, T. Suzuki, D. Wu, T. Zhang, Understanding the variance collapse of svgd in high dimensions, in: International Conference on Learning Representations, 2021.
[26] R.M. Neal, Mcmc using Hamiltonian dynamics, Handb. Markov Chain Monte Carlo 2 (11) (2011) 2.
[27] A. Fichtner, A. Zunino, L. Gebraad, Hamiltonian Monte Carlo solution of tomographic inverse problems, Geophys. J. Int. 216 (2) (2019) 1344–1363.
[28] I. Langmore, M. Dikovsky, S. Geraedts, P. Norgaard, R. Von Behren, Hamiltonian Monte Carlo in inverse problems. Ill-conditioning and multimodality, Int. J. Uncertain. Quantificat. 13 (1) (2023).
[29] M. Betancourt, A conceptual introduction to Hamiltonian Monte Carlo, arXiv preprint, arXiv:1701.02434, 2017.
[30] K. Wang, T. Bui-Thanh, O. Ghattas, A randomized maximum a posteriori method for posterior sampling of high dimensional nonlinear Bayesian inverse problems, SIAM J. Sci. Comput. 40 (1) (2018) A142–A171.
[31] Y. Chen, D.S. Oliver, Ensemble randomized maximum likelihood method as an iterative ensemble smoother, Math. Geosci. 44 (2012) 1–26.
[32] J.M. Bardsley, A. Solonen, H. Haario, M. Laine, Randomize-then-optimize: a method for sampling from posterior distributions in nonlinear inverse problems, SIAM J. Sci. Comput. 36 (4) (2014) A1895–A1910.
[33] J.T. White, A model-independent iterative ensemble smoother for efficient history-matching and uncertainty quantification in very high dimensions, Environ. Model. Softw. 109 (2018) 191–201, https://doi.org/10.1016/j.envsoft.2018.06.009.
[34] L. Lu, P. Jin, G. Pang, Z. Zhang, G.E. Karniadakis, Learning nonlinear operators via deeponet based on the universal approximation theorem of operators, Nat. Mach. Intell. 3 (3) (2021) 218–229.
[35] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, A. Anandkumar, Fourier neural operator for parametric partial differential equations, arXiv preprint, arXiv:2010.08895, 2020.
[36] Y. Wang, Y. Zong, J.L. McCreight, J.D. Hughes, A.M. Tartakovsky, Bayesian reduced-order deep learning surrogate model for dynamic systems described by partial differential equations, Comput. Methods Appl. Mech. Eng. 429 (2024) 117147.
[37] Z. Gao, L. Yan, T. Zhou, Adaptive operator learning for infinite-dimensional Bayesian inverse problems, arXiv preprint, arXiv:2310.17844, 2023.
[38] C.E. Rasmussen, C.K. Williams, et al., Gaussian Processes for Machine Learning, vol. 1, Springer, 2006.
[39] F. Chyzak, Algorithms Seminar, 2000-2001, Research Report RR-4406, INRIA, 2002, https://inria.hal.science/inria-00072182.
[40] L. Tierney, Markov chains for exploring posterior distributions, Ann. Stat. (1994) 1701–1728.
[41] M.A. Branch, T.F. Coleman, Y. Li, A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems, SIAM J. Sci. Comput. 21 (1) (1999) 1–23.
[42] A. Beskos, N. Pillai, G. Roberts, J.-M. Sanz-Serna, A. Stuart, Optimal tuning of the hybrid Monte Carlo algorithm, Bernoulli 19 (5A) (2013) 1501–1534, https://doi.org/10.3150/12-BEJ414.
[43] J. Carrera, S.P. Neuman, Estimation of aquifer parameters under transient and steady state conditions: 2. Uniqueness, stability, and solution algorithms, Water Resour. Res. 22 (2) (1986) 211–227, https://doi.org/10.1029/WR022i002p00211.
[44] C.R. Cole, M.P. Bergeron, S.K. Wurstner, P.D. Thorne, S. Orr, M.I. Mckinley, Transient inverse calibration of Hanford site-wide groundwater model to Hanford operational impacts-1943 to 1996, Tech. Rep., Pacific Northwest National Laboratory (PNNL), Richland, Washington, United States, 2001.
[45] A.M. Tartakovsky, M. Panzeri, G.D. Tartakovsky, A. Guadagnini, Uncertainty quantification in scale-dependent models of flow in porous media, Water Resour. Res. 53 (2017) 9392–9401.

[46] M.D. Hoffman, A. Gelman, et al., The no-u-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo, J. Mach. Learn. Res. 15 (1) (2014) 1593–1623.
[47] M.J. Betancourt, S. Byrne, M. Girolami, Optimizing the integrator step size for Hamiltonian Monte Carlo, arXiv:1411.6669, 2015.
[48] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint, arXiv:1412.6980, 2014.
[49] D.S. Oliver, N. He, A.C. Reynolds, Conditioning permeability fields to pressure data, in: ECMOR V-5th European Conference on the Mathematics of Oil Recovery, European Association of Geoscientists & Engineers, 1996, pp. cp–101.
[50] M.A. Girolami, B. Calderhead, Riemann manifold Langevin and Hamiltonian Monte Carlo methods, J. R. Stat. Soc., Ser. B, Stat. Methodol. 73 (2011).
[51] Y.L. Keung, J. Zou, Numerical identifications of parameters in parabolic systems, Inverse Probl. 14 (1) (1998) 83, https://doi.org/10.1088/0266-5611/14/1/009.