Mathematical Geosciences
https://doi.org/10.1007/s11004-023-10127-0

Importance Weighting in Hybrid Iterative Ensemble Smoothers for Data Assimilation

Yuming Ba[1] · Dean S. Oliver[2]

Received: 18 September 2023 / Accepted: 26 November 2023

Abstract

Because it is generally impossible to completely characterize the uncertainty in complex model variables after assimilation of data, it is common to approximate the uncertainty by sampling from approximations of the posterior distribution for model variables. When minimization methods are used for the sampling, the weights on each of the samples depend on the magnitude of the data mismatch at the critical points and on the Jacobian of the transformation from the prior density to the sample proposal density. For standard iterative ensemble smoothers, the Jacobian is identical for all samples, and the weights depend only on the data mismatch. In this paper, a hybrid data assimilation method is proposed which makes it possible for each ensemble member to have a distinct Jacobian and for the approximation to the posterior density to be multimodal. For the proposed hybrid iterative ensemble smoother, it is necessary that a part of the mapping from the prior Gaussian random variable to the data be analytic. Examples might include analytic transformation from a latent Gaussian random variable to permeability followed by a black-box transformation from permeability to state variables in porous media flow, or a Gaussian hierarchical model for variables followed by a similar black-box transformation from permeability to state variables. In this paper, the application of weighting to both hybrid and standard iterative ensemble smoothers is investigated using a two-dimensional, two-phase flow problem in porous media with various degrees of nonlinearity. As expected, the weights in a standard iterative ensemble smoother become degenerate for problems with large amounts of

Yuming Ba and Dean S. Oliver have contributed equally to this work.

✉ Dean S. Oliver
dean.oliver@norceresearch.no

Yuming Ba
yumingba@gpnu.edu.cn

[1] School of Mathematics and Systems Science, Guangdong Polytechnic Normal University, No. 293, Zhongshan Avenue West, Tianhe District, Guangzhou 510065, Guangdong, China

[2] Technology Department, NORCE Norwegian Research Centre, Nygårdsgaten 112, 5008 Bergen, Norway

Published online: 08 January 2024

data. In the examples, however, the weights for the hybrid iterative ensemble smoother were useful for improving forecast reliability.

Keywords Weighted randomized maximum likelihood · Hybrid iterative ensemble smoother · Denoising · Data assimilation

1 Introduction

In many geoscience applications, parameters of high-dimensional models must be estimated from a limited number of noisy data. The data are often only indirectly and non-linearly related to the parameters of the model. Consequently, the parameters of the model are usually underdetermined, and estimation of a single set of model parameters that satisfy the data is not sufficient to characterize the solution of the inverse problem.

Although powerful methods for quantifying uncertainty in high-dimensional model spaces with Gaussian uncertainty are available (Martin et al. 2012), and powerful Monte Carlo methods are available for non-Gaussian low-dimensional model spaces, it is still a challenge to quantify uncertainty for situations where the model dimension is large and the posterior distribution is non-Gaussian. In that case, approximate sampling methods must typically be used. A relatively standard approach to approximate sampling is through the minimization of a stochastic cost function. The method is known by various names including geostatistical inversing (Kitanidis 1995), randomized maximum likelihood (Oliver et al. 1996), randomized maximum a posteriori (Wang et al. 2018), and randomize-then-optimize (Bardsley et al. 2014). In these methods, a realization from an approximation to the posterior distribution is generated by minimizing the weighted squared distance of the posterior sample to a realization from the prior and the squared distance between the actual data and the perturbed simulated data. The method provides exact sampling when the prior distribution is Gaussian and the relationship between data and model parameters is linear. When the posterior distribution is non-Gaussian but unimodal, it is possible to weight the realizations from minimization such that the sampling is exact (Oliver et al. 1996; Oliver 2017; Bardsley et al. 2014, 2020; Wang et al. 2018), although computation of weights in high dimensions may be difficult.

The actual posterior landscape for distributed-parameter geoscience inverse problems is difficult to ascertain, although there are known to be features of subsurface flow models that result in multimodal posterior distributions: uncertain fault displacement in a layered reservoir (Tavassoli et al. 2005), uncertain rock type location in a channelized reservoir (Zhang et al. 2003), layered non-communicating reservoir flow with independent uncertain properties (Oliver et al. 2011), and non-Gaussian prior distributions for log-permeability (Oliver and Chen 2018). For relatively simple transient single-phase flow problems, in which the prior distribution of log-permeability is multivariate normal, the posterior distribution appears to be multimodal in low-dimensional subspaces, but appears to be characterized by curved ridges in higher-dimensional subspaces (Oliver and Chen 2011).

Sampling the posterior distribution correctly for subsurface flow models is difficult when there are hundreds of thousands to millions of uncertain parameters whose magnitudes must be inferred from the data. The Markov chain Monte Carlo (MCMC) methods are usually considered the gold standard for sampling from Bayesian posteriors. Because of the high computational cost of evaluating the likelihood function for subsurface flow models, however, MCMC is seldom used for subsurface models. For a two-dimensional single-phase porous flow problem in which only the permeability field is uncertain, an MCMC method with transition proposals required hundreds of thousands of iterations to generate a useful number of independent samples (Oliver et al. 1997). For a porous flow model with a more complex posterior but only three uncertain parameters, a population MCMC approach demonstrated good mixing properties and provided good results, but at substantial cost (Mohamed et al. 2012). In order to reduce the cost of the likelihood evaluation, Maschio and Schiozer (2014) replaced the flow simulator by proxy models generated by an artificial neural network. An iterative procedure combining MCMC sampling and artificial neural network (ANN) training was applied to a reservoir model with 16 uncertain attributes.

Iterative ensemble smoothers, on the other hand, have been remarkably successful at history matching large amounts of data into reservoir models with hundreds of thousands of parameters. Based loosely on the ensemble Kalman filter (Evensen 1994), which is routinely used for numerical weather prediction, iterative ensemble smoothers use stochastic gradient approximations for minimization, so that solutions of the adjoint system are not necessary. The downside of this is that the methodology must approximate the cost function as a quadratic surface. Consequently, the method is not well suited to arbitrary posterior landscapes.

When the posteriori probability density function (pdf) has multiple modes of any type, minimization-based simulation methods will almost certainly sample occasionally from local minima of the cost function that contribute very little to the probability mass in the posterior. These samples are usually, but not always, characterized by large data mismatch after minimization. In the case of an exceptionally large data mismatch, it is common practice to omit poorly calibrated model realizations when computing mean forecasts and uncertainty quantification. It is much more difficult to decide how to treat realizations with intermediate data mismatches, or realizations in general when the unweighted distribution is known to be only approximate. Importance weighting of realizations to correct for the approximate sampling is the principled approach to uncertainty quantification in these cases.

It has been shown that a standard class of importance-weighted particle filters are degenerate in high dimensions even when the so-called optimal proposal is used (Snyder et al. 2015; van Leeuwen et al. 2019). The optimal proposal is defined as the one that minimizes the variance of weights for particle filters that are based only on the new observations and the particles generated at a previous step (Doucet et al. 2000). For updating schemes that are not limited to the use of the prior ensemble of particles, the weights need not be degenerate. Ba et al. (2022) demonstrated that the weights in a properly weighted randomized maximum likelihood (RML) scheme are not necessarily degenerate, even when the problem is nonlinear. And it is known that linear inference problems can be sampled exactly using the minimization methods, in which case all weights are identical. When the problem has many modes, however, the local minimizer is attracted to modes with small probability mass, and the weights may become degenerate simply because the weights on small local minima should be small. Weights can also become degenerate when they are inaccurately computed because of approximations in the gradient.

van Leeuwen et al. (2019) identify four approaches to reducing the variance in weights for particle filters. The minimization approach (RML) used in this paper can be considered to belong to the category in which particles are pushed from the prior into regions of high posterior probability density. For inverse problems with Gaussian priors on model parameters and linear observation operators, after minimization, the particles are distributed as the target distribution, so weighting is not required. When the posterior is multimodal, however, the problem of weighting can be relatively complex, as each particle in the prior potentially maps to multiple particles in the proposal density, each with a different weight. Ba et al. (2022) computed low-rank approximations to the particle weights for a single-phase porous media flow problem using the adjoint system for the flow simulator, but adjoint systems are not always available, and the cost can be prohibitive. In this paper, we show how approximate weights can be easily computed when an ensemble Kalman-based approach is used to solve an inverse problem.

In this paper, we develop an importance weighting approach to the problem in which the relationship between observations and model parameters is sufficiently nonlinear that the posterior distribution for model parameters is multimodal. Although importance weighting in particle filters is a standard approach for dealing with nonlinearity in small data assimilation problems, we apply it to problems with relatively large numbers of model parameters and data. We show that this can be done in a hybrid iterative ensemble smoother approach for which gradients required for minimization are computed using a combination of analytical and stochastic gradients. The method is applied to a two-phase porous media flow problem with multimodal posterior pdf. In order to be useful for large problems, we improve an earlier approach through the use of circulant embedding of the covariance matrix to allow matrix multiplication in large models. Finally, we demonstrate that the weights computed using a hybrid or ensemble Kalman-like approach are noisy approximations of the true weights and that denoising the weights improves model predictability.

2 Methodology

Consider the following generic forward model for prediction of u given $\boldsymbol{\theta}$,

$$\mathcal{B}(\boldsymbol{\theta}, u) = 0, \quad \text{in} \quad \Omega,$$

which for example could be a system of partial differential equations (PDEs) characterizing a physical problem. In the parameter estimation problem, the task is to quantify the unknown parameter $\boldsymbol{\theta}$ given some limited observations of u on parts of the domain Ω. The relationship between model parameters and observations is given by the widely used model

$$\boldsymbol{d}^o = g(\boldsymbol{m}(\boldsymbol{\theta})) + \boldsymbol{\epsilon},$$

where $g(\cdot)$ is the generic observation operator and $\boldsymbol{m}(\cdot)$ is the model operator mapping the unknown $\boldsymbol{\theta}$ to the space of an intermediate variable, such as the hierarchical model, transformation of permeability, and composite observation, including the three cases of Sect. 2. In a finite-dimensional parameter space, $\boldsymbol{\theta} \in \mathbb{R}^{N_\theta}$, $\boldsymbol{m}(\boldsymbol{\theta}) \in \mathbb{R}^{N_m}$, and $\boldsymbol{d}^o \in \mathbb{R}^{N_d}$. Assume that $\boldsymbol{\epsilon} \in \mathbb{R}^{N_d}$ is independent of $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{C}_d)$. Given a prior Gaussian distribution $\mathrm{N}(\boldsymbol{\theta}^{\mathrm{pr}}, \boldsymbol{C}_\theta)$, we expect to generate samples $\boldsymbol{\theta}^i$, $i = 1, \ldots, N_e$, from the posterior distribution

$$\pi_\Theta(\boldsymbol{\theta}|\boldsymbol{d}^o) = \frac{\pi_{\Theta D}(\boldsymbol{d}^o, \boldsymbol{\theta})}{\pi_D(\boldsymbol{d}^o)} \propto \exp(-Q(\boldsymbol{\theta})),$$

with the negative log posterior function

$$Q(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{pr}})^T \boldsymbol{C}_\theta^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{pr}}) + \frac{1}{2}(g(\boldsymbol{m}(\boldsymbol{\theta})) - \boldsymbol{d}^o)^T \boldsymbol{C}_d^{-1}(g(\boldsymbol{m}(\boldsymbol{\theta})) - \boldsymbol{d}^o).$$

In general, the normalization constant $\pi_D(\boldsymbol{d}^o)$ is unknown, but independent of $\boldsymbol{\theta}$. For simplicity, $\boldsymbol{m}(\boldsymbol{\theta})$ is denoted as $\boldsymbol{m}$.

In this paper, we apply ensemble Kalman-like approximations to the randomized maximum likelihood (RML) method (Kitanidis 1995; Oliver et al. 1996; Chen and Oliver 2012) for data assimilation. The RML method draws samples $(\boldsymbol{\theta}^{i'}, \boldsymbol{\delta}^{i'})$ from the Gaussian distribution

$$\begin{aligned} q_{\Theta'\Delta'}(\boldsymbol{\theta}', \boldsymbol{\delta}') = q_{\Theta'}(\boldsymbol{\theta}')\, q_{\Delta'}(\boldsymbol{\delta}') = &\frac{1}{(2\pi)^{\frac{N_\theta+N_d}{2}} |\boldsymbol{C}_\theta|^{1/2} |\boldsymbol{C}_d|^{1/2}} \\ &\times \exp\Big(-\frac{1}{2}(\boldsymbol{\theta}' - \boldsymbol{\theta}^{\mathrm{pr}})^T \boldsymbol{C}_\theta^{-1}(\boldsymbol{\theta}' - \boldsymbol{\theta}^{\mathrm{pr}}) - \frac{1}{2}(\boldsymbol{\delta}' - \boldsymbol{d}^o)^T \boldsymbol{C}_d^{-1}(\boldsymbol{\delta}'(\boldsymbol{\delta}' - \boldsymbol{d}^o)\Big), \end{aligned} \tag{1}$$

for given $\boldsymbol{\theta}^{\mathrm{pr}}$ and $\boldsymbol{d}^o$. The ith approximate posterior sample is then generated by computing the critical points of the cost functional

$$Q_i(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^{i'})^T \boldsymbol{C}_\theta^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^{i'}) + \frac{1}{2}(g(\boldsymbol{m}) - \boldsymbol{\delta}^{i'})^T \boldsymbol{C}_d^{-1}(g(\boldsymbol{m}) - \boldsymbol{\delta}^{i'}). \tag{2}$$

The critical points are obtained by solving $\nabla_\theta Q_i(\boldsymbol{\theta}) = 0$ for $\boldsymbol{\theta}$. In general, the maxima and stationary points contribute little, and the Levenberg–Marquardt method with a Gauss–Newton approximation of the Hessian is used for the minimization.The ith increment in the iteration is written as

$$\delta\boldsymbol{\theta}_l = \frac{\boldsymbol{\theta}^{i'} - \boldsymbol{\theta}_l}{1+\lambda_l} - \boldsymbol{C}_\theta \boldsymbol{G}_l^T \Big[(1+\lambda_l)\boldsymbol{C}_d + \boldsymbol{G}_l \boldsymbol{C}_\theta \boldsymbol{G}_l^T\Big]^{-1} \Big[\Big(g(\boldsymbol{m}_l) - \boldsymbol{\delta}^{i'}\Big) - \frac{\boldsymbol{G}_l(\boldsymbol{\theta}_l - \boldsymbol{\theta}^{i'})}{1+\lambda_l}\Big], \tag{3}$$

where $\boldsymbol{G}_l = (\nabla_{\theta_l}(g^T))^T$, and λ_l is the Levenberg–Marquardt regularization parameter for the ℓth iteration.

The ensemble Kalman approximation of the RML is asymptotically exact for Gauss-linear data assimilation problems and adopts an average sensitivity computed from the ensemble samples to approximate the downhill direction (Chen and Oliver 2012), which results in inaccurate sensitivity when the problem is highly nonlinear. To improve the accuracy of the sensitivity matrix for individual realizations, the hybrid ensemble method is introduced. (The hybrid method can refer to many different approaches. Here, we refer to approaches that use gradients that are computed partially from the ensemble and partially by direct differentiation.) Through proper forms such as Eq. (6) and Eq. (7) in the hybrid ensemble method, some derivatives are computed analytically and others are approximated from the ensemble. Consequently, instead of a single common gain matrix applied to all realizations, the gain matrix of each sample in hybrid ensemble methods is different. In a naïve implementation, the computational cost will be very high for large models. We take advantage of the block-circulant structure of the prior model covariance matrix or its square root to reduce the cost substantially, applying circulant embedding for fast multiplication of Toeplitz matrices.

For posterior distributions with multiple modes, the approximate samples from minimization-based simulation methods will almost certainly converge to local minima, some of which contribute very little to the probability mass in the posterior. These samples may result in large data mismatch after minimization. Importance weights of approximate samples from the proposal distribution are used to correct the sampling. To compute the importance weights, it is necessary to compute the proposal distribution for RML samples. Solving $\nabla_\theta Q(\boldsymbol{\theta}) = 0$ leads to a map from $(\boldsymbol{\theta}, \boldsymbol{\delta})$ to $(\boldsymbol{\theta}', \boldsymbol{\delta}')$ in Ba et al. (2022),

$$\begin{cases} \boldsymbol{\theta}' = \boldsymbol{\theta} + \boldsymbol{C}_\theta \boldsymbol{G}^T \boldsymbol{C}_d^{-1}\big(g(\boldsymbol{m}) - \boldsymbol{\delta}\big) \\ \boldsymbol{\delta}' = \boldsymbol{\delta}. \end{cases} \tag{4}$$

Based on the map of Eq. (4) and the original notation, the distribution of the transformed variables is given by

$$\begin{aligned} p_{\Theta\Delta}(\boldsymbol{\theta}, \boldsymbol{\delta}) &:= n(\boldsymbol{\theta}')^{-1} q_{\Theta'\Delta'}(\boldsymbol{\theta}', \boldsymbol{\delta}') J(\boldsymbol{\theta}, \boldsymbol{\delta}) \\ &= n(\boldsymbol{\theta}')^{-1} q_{\Theta'}\Big(\boldsymbol{\theta} + \boldsymbol{C}_\theta \boldsymbol{G}^T \boldsymbol{C}_d^{-1}\big(g(\boldsymbol{m}) - \boldsymbol{\delta}\big)\Big) q_{\Delta'}(\boldsymbol{\delta}) J(\boldsymbol{\theta}, \boldsymbol{\delta}), \end{aligned} \tag{5}$$

where $n(\boldsymbol{\theta}')$ is the total number of critical points of Eq. (2), and $J(\boldsymbol{\theta}, \boldsymbol{\delta})$ denotes the Jacobian determinant associated with the map $(\boldsymbol{\theta}, \boldsymbol{\delta}) \to (\boldsymbol{\theta}', \boldsymbol{\delta}')$. In the following, we assume that the map is locally invertible, that is, $J \neq 0$ everywhere. The form of $J(\boldsymbol{\theta}, \boldsymbol{\delta})$ is provided by

$$J(\boldsymbol{\theta}, \boldsymbol{\delta}) = \left| I + \mathcal{D}\Big(\boldsymbol{C}_\theta \boldsymbol{G}^T \boldsymbol{C}_d^{-1}\big(g(\boldsymbol{m}) - \boldsymbol{\delta}\big)\Big) \right|,$$

where $\mathcal{D}(\cdot)$ is the gradient operator for $\boldsymbol{C}_\theta \boldsymbol{G}^T \boldsymbol{C}_d^{-1}\big(g(\boldsymbol{m}) - \boldsymbol{\delta}\big)$ with respect to $\boldsymbol{\theta}$.

When importance sampling is implemented for highly nonlinear problems, the variance in the log-weights will generally not be small. This is because the RML proposal density is not identical to the target density, and the ensemble samples are only approximations of the samples that would be obtained from exact computation of minima. Because of the approximations, the actual spread in computed importance weights will be larger than it should be. Denoising of importance weights has been shown to be effective at improving the weights (Akyildiz et al. 2017). For the ensemble methods based on RML, denoising will be performed when the variance of weights is large.

2.1 Example Applications

The weighting of critical-point samples in the RML method depends on the magnitude of the data mismatch at the critical points and on the Jacobian of the transformation from the prior density to the proposal density. When standard iterative ensemble smoothers are applied for data assimilation, the Jacobian is identical for all samples. If a hybrid data assimilation method is applied, however, there is the possibility for each ensemble member to have a distinct Jacobian and for the posterior distribution of particles to be multimodal. In order to apply a hybrid method iterative ensemble smoother, it is necessary that a part of the transformation from the prior Gaussian random variable to the data be analytic. Examples might include transformation from a latent Gaussian random variable to permeability followed by a system of partial differential equations mapping permeability to state variables in porous media flow, or a Gaussian hierarchical model for variables followed by a similar transformation from permeability to state variables

$$\boldsymbol{G} = (\nabla_\theta (g^T))^T = \boldsymbol{G}_m (\nabla_\theta (\boldsymbol{m}^T))^T = \boldsymbol{G}_m \boldsymbol{M}_\theta, \tag{6}$$

where $\boldsymbol{G}$ is the sensitivity matrix of the forward operator g with respect to the latent Gaussian random variable $\boldsymbol{\theta}$, $\boldsymbol{G}_m$ is the gradient of the forward operator with respect to the log-permeability $\boldsymbol{m}$, and $\boldsymbol{M}_\theta$ is the gradient of the log-permeability m with respect to $\boldsymbol{\theta}$.

2.1.1 Hierarchical Gaussian

For a hierarchical Gaussian model in which hyper-parameters, $\boldsymbol{\beta}$, of the prior model covariance such as the principal ranges and the orientation of the anisotropy are uncertain, we might use the non-centered parameterization (Papaspiliopoulos et al. 2007) to express the relationship between the observable Gaussian variable $\boldsymbol{m}$ (e.g., log-permeability) and the model parameters $\boldsymbol{z}$ (latent Gaussian variables) and $\boldsymbol{\beta}$ as

$$\boldsymbol{m} = \boldsymbol{m}_{\text{pr}} + \boldsymbol{L}(\boldsymbol{\beta})\boldsymbol{z},$$

where $\boldsymbol{L}$ is a square root of the model covariance matrix $\boldsymbol{C}_m = \boldsymbol{L}\boldsymbol{L}^T$. In this application, the sensitivity of the observable variable $\boldsymbol{m}$ to the latent variables $\boldsymbol{z}$ and $\boldsymbol{\beta}$ is nonlocal,

$$\boldsymbol{M}_\theta = \left[\boldsymbol{L}\ (\nabla_\beta \boldsymbol{L})\boldsymbol{z}\right].$$

In a hybrid iterative ensemble smoother (IES), the sensitivity of production data to permeability and porosity, $\boldsymbol{G}_m$, would be estimated using the ensemble of predicted data and the ensemble of model perturbations.

2.1.2 Transformation of Permeability

In some applications of data assimilation to subsurface characterization, it is desirable to generate prior realizations of the permeability field with non-Gaussian structure, that is, continuous channel-like features of high permeability embedded in a low-permeability background. A property field with these characteristics can be obtained by applying a nonlinear transformation to a correlated Gaussian field, that is, $\boldsymbol{m} = f(\boldsymbol{\theta})$, where $\theta \sim \mathrm{N}(\boldsymbol{\theta}^{\mathrm{pr}}, \boldsymbol{C}_\theta)$. Unlike the hierarchical example, in which the sensitivity matrix $\boldsymbol{M}_\theta$ had dimensions $N_m \times N_m$ and was potentially full, the sensitivity of $\boldsymbol{m}$ to $\boldsymbol{\theta}$ for simple variable transformation will generally be diagonal.

2.1.3 Composite Observation Operators

For some types of subsurface data assimilation problems, the observation operator might be separable into two parts, one of which can be treated analytically, while the other might be a complex function of the parameters, determined by the solution of a partial differential equation. An example is the observation of acoustic impedance in a seismic survey. The state of the reservoir u (i.e., the pressure and saturation) is a function of permeability and porosity, which are denoted as $u(\boldsymbol{\beta}_1)$. The acoustic impedance, Z, is related to the state of the reservoir and other reservoir properties through a rock physics model, which may have several additional uncertain parameters, $\boldsymbol{\beta}_2$. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$. The composite relationship is written loosely as

$$Z(\boldsymbol{\theta}) = Z(u(\boldsymbol{\beta}_1), \boldsymbol{\beta}_2).$$

The sensitivity of acoustic impedance to the permeability and porosity can be decomposed using the chain rule as

$$\boldsymbol{G} = (\nabla_\theta(Z^T))^T = (\nabla_{u,\beta_2}(Z^T))^T(\nabla_{\beta_1}(u^T))^T = \boldsymbol{G}_{u,\beta_2}\boldsymbol{U}_{\beta_1}, \tag{7}$$

in which case the sensitivity of impedance to the state variables and parameters of the rock physics model, $\boldsymbol{G}_{u,\beta_2}$, can be computed analytically, and the sensitivity of the state variables to permeability and porosity, $\boldsymbol{U}_{\beta_1}$, can be estimated stochastically as in an iterative ensemble smoother.

2.2 Data Assimilation Based on Ensemble Methods

In practice, iterative ensemble smoothers are often an effective approach for solving large-scale geoscience inverse problems. These methods are based on the Kalman filter (Evensen 1994), which uses a low-rank approximation of the covariance matrix to replace the full covariance and avoids the need to compute adjoints of the objective functions as might be required in an extended Kalman filter. To improve the efficiency of updating the unknown parameters, a so-called smoother method using all the data simultaneously is generally adopted for parameter estimation problems. However, most parameter estimation problems are nonlinear, and a single update in which all data are simultaneously assimilated is not sufficient. For history matching, iteration is required of a smoother application. Iterative ensemble smoothers (IES) and their variants include two general approaches: multiple data assimilation (MDA) (Reich 2011; Emerick and Reynolds 2013) and IES based on randomized maximum likelihood (RML) (Chen and Oliver 2012). The iterative ensemble smoother form of the RML uses an average sensitivity to approximate the Hessian matrix. For the strongly nonlinear problems, the ensemble average sensitivity will provide a poor approximation of the local sensitivity. To partially rectify this problem, a hybrid RML-IES method has been proposed to improve the estimate of the local sensitivity; some gradients are computed analytically and others are approximated from the ensemble (Oliver 2022).

2.2.1 Iterative Ensemble Smoother

For the RML method, the computation of the gradient of the objective function with respect to the parameters is necessary. In many high-dimensional problems, the computation of derivatives is difficult. The iterative ensemble smoothers utilize ensemble realizations to approximate the first- and second-order moments, which avoids the need to compute derivatives directly. Using an iterative ensemble smoother method (Chen and Oliver 2013), the update step (Eq. (3)) for the ith ensemble member at the lth iteration can be approximated as

$$\begin{aligned}\boldsymbol{\theta}_{l+1}^{i} = \boldsymbol{\theta}_{l}^{i} - \frac{1}{1+\lambda_l}\Xi_{\theta_l}(\Xi_{\theta_l})^T\boldsymbol{C}_{\theta}^{-1}(\boldsymbol{\theta}_{l}^{i}-\boldsymbol{\theta}^{i'}) - \Xi_{\theta_l}(\Xi_{d_l})^T\Big((1+\lambda_l)\boldsymbol{C}_d + \Xi_{d_l}(\Xi_{d_l})^T\Big)^{-1} \\ \times\Big(g(\boldsymbol{m}_l) - \boldsymbol{\delta}^{i'} - \frac{1}{1+\lambda_l}\Xi_{d_l}(\Xi_{\theta_l})^T\boldsymbol{C}_{\theta}^{-1}(\boldsymbol{\theta}_{l}^{i}-\boldsymbol{\theta}^{i'})\Big),\end{aligned} \tag{8}$$

where

$$\begin{cases}\Xi_{\theta_l} = \dfrac{1}{\sqrt{N_e-1}}(\boldsymbol{\theta}_l^1,\ldots,\boldsymbol{\theta}_l^{N_e})\Big(\boldsymbol{I}_{N_e} - \dfrac{1}{N_e}\mathbf{1}_{N_e}\mathbf{1}_{N_e}^T\Big),\\ \Xi_{d_l} = \dfrac{1}{\sqrt{N_e-1}}\Big(g(\boldsymbol{m}_l^1),\ldots,g(\boldsymbol{m}_l^{N_e})\Big)\Big(\boldsymbol{I}_{N_e} - \dfrac{1}{N_e}\mathbf{1}_{N_e}\mathbf{1}_{N_e}^T\Big),\\ \boldsymbol{\delta}^{i'} \sim \mathrm{N}(\boldsymbol{d}^o, \boldsymbol{C}_d), \quad \boldsymbol{\theta}^{i'} \sim \mathrm{N}(\boldsymbol{\theta}^{\mathrm{pr}}, \boldsymbol{C}_\theta).\end{cases}$$

Here, N_e is the number of model realizations in the initial ensemble. The ensemble realizations are used to approximate the gradient of the forward operator g with respect to $\boldsymbol{\theta}$. The update in Eq. (8) is restricted to the space spanned by the initial ensemble, and the number of degrees of freedom available for calibration of the model to data is $N_e - 1$. To avoid the tendency for ensemble collapse with large amounts of data, localization is almost always used in high-dimensional problems. Additionally, for highly nonlinear problems, the sensitivity of data to model parameters estimated from the ensemble in Eq. (8) is a poor approximation to the local sensitivity, which often results in failure to converge to local minima.

2.2.2 Hybrid Iterative Ensemble Smoother

For the hybrid IES, instead of using an ensemble stochastic approximation of the sensitivity matrix $\boldsymbol{G}$, the derivative of $\boldsymbol{m}$ with respect to $\boldsymbol{\theta}$ is computed analytically, and the chain rule is used to compute $\boldsymbol{G}$ as

$$\boldsymbol{G} = \boldsymbol{G}_m \cdot \left(\nabla_\theta \left(\boldsymbol{m}^T\right)\right)^T = \boldsymbol{G}_m \boldsymbol{M}_\theta. \tag{9}$$

Then, the update equation (Eq. (3)) can be written as

$$\begin{aligned}\boldsymbol{\theta}_{l+1} = \boldsymbol{\theta}_l &- \frac{1}{1+\lambda_l}(\boldsymbol{\theta}_l - \boldsymbol{\theta}') - \boldsymbol{C}_\theta \boldsymbol{M}_\theta^T (\Xi_{m_l})^{-T} (\Xi_{d_l})^{\mathrm{T}} \\ &\times \left((1+\lambda_l)\boldsymbol{C}_d + (\Xi_{d_l})(\Xi_{m_l})^{-1} \boldsymbol{M}_\theta \boldsymbol{C}_\theta \boldsymbol{M}_\theta^T (\Xi_{m_l})^{-T} (\Xi_{d_l})^T\right)^{-1} \\ &\times \left(g(\boldsymbol{m}_l) - \boldsymbol{\delta}' - \frac{1}{1+\lambda_l}(\Xi_{d_l})(\Xi_{m_l})^{-1} \boldsymbol{M}_\theta (\boldsymbol{\theta}_l - \boldsymbol{\theta}')\right),\end{aligned} \tag{10}$$

where Ξ_{m_l} is defined similarly to Ξ_{θ_l}, and $\boldsymbol{M}_\theta = (\nabla_\theta(\boldsymbol{m}^{\mathrm{T}}))^{\mathrm{T}}$. The gain matrix for each sample is different because the sensitivity matrix $\boldsymbol{M}_\theta$ is evaluated at the model realization—not estimated from the ensemble of realizations. The main challenge with straightforward application of a hybrid IES methodology is the cost of forming and multiplying by the matrix $\boldsymbol{M}_\theta$ for all realizations (Oliver 2022).

Unlike the ensemble Kalman-based methods that take advantage of low-rank approximations of the covariance matrices, in the hybrid IES, the Toeplitz or block Toeplitz structure of the covariance matrix for stationary Gaussian random fields is utilized. It is then possible to efficiently compute the matrix–vector products using the fast Fourier transform after embedding the Toeplitz matrix in a circulant matrix (Appendix A).

2.3 Weighting of Model Realization

The RML method of sampling the posterior is only exact if the relationship between the data and the model parameters is linear. For many nonlinear problems, however, it is necessary to weight the samples to approximate the posterior distribution, in which case the computation of the gradient of the objective function is necessary. To avoid the need to compute $\boldsymbol{G}$ directly, ensemble-based methods offer an alternative. However, exact sampling using RML requires computation of additional critical points and weighting of solutions (Ba et al. 2022). The importance weight for the kth RML sample is

$$\omega_k \propto \frac{\pi_\Theta(\boldsymbol{\theta}^k)\pi_\Delta(\boldsymbol{\delta}^k|\boldsymbol{\theta}^k)}{p_{\Theta\Delta}(\boldsymbol{\theta}^k, \boldsymbol{\delta}^k)}, \tag{11}$$

where $\pi_\Theta(\boldsymbol{\theta})$ is the prior, and the likelihood $\pi_\Delta(\boldsymbol{\delta}|\boldsymbol{\theta})$ and the proposal density $p_{\Theta\Delta}(\boldsymbol{\theta}, \boldsymbol{\delta})$ are provided by Eq. (5). Ba et al. (2022) showed that in high-dimensional nonlinear cases where it is not feasible to sample all critical points, it is possible to randomly sample a single critical point for each prior realization. If the critical point is sampled uniformly from the set of all critical points, the factor $n(\boldsymbol{\theta}')$ in Eq. (5) should be set to 1.

Introducing the quantities,

$$\begin{aligned} J(\boldsymbol{\theta}, \boldsymbol{\delta}) &\approx |\boldsymbol{I}_{N_\theta} + \boldsymbol{C}_\theta \boldsymbol{G}^T \boldsymbol{C}_d^{-1} \boldsymbol{G}| \\ \boldsymbol{V}(\boldsymbol{\theta}) &= \boldsymbol{C}_d + \boldsymbol{G}\boldsymbol{C}_\theta \boldsymbol{G}^{\mathrm{T}} \\ \boldsymbol{\eta}(\boldsymbol{\theta}) &= g(\boldsymbol{m}) - \boldsymbol{d}^o - \boldsymbol{G}(\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{pr}}). \end{aligned}$$

To simplify notation, the proposal density Eq. (5) which appears in the denominator of Eq. (11) can be written as

$$\begin{aligned} & p_{\Theta\Delta}(\boldsymbol{\theta}, \boldsymbol{\delta}) \\ &= \overbrace{A_0 \exp\left[-\frac{1}{2}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{pr}}\right)^{\mathrm{T}} \boldsymbol{C}_\theta^{-1} \left(\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{pr}}\right) - \frac{1}{2}\left(g(\boldsymbol{m}) - \boldsymbol{d}^o\right)^{\mathrm{T}} \boldsymbol{C}_d^{-1} \left(g(\boldsymbol{m}) - \boldsymbol{d}^o\right)\right]}^{\pi_\Theta(\boldsymbol{\theta})} \\ & \quad \times \overbrace{A_1 \, |\boldsymbol{V}|^{1/2} \exp\left[-\frac{1}{2}\left(\boldsymbol{\delta} - g(\boldsymbol{m}) - \boldsymbol{V}^{-1}\boldsymbol{\eta}(\boldsymbol{m})\right)^{\mathrm{T}} \boldsymbol{V} \left(\boldsymbol{\delta} - g(\boldsymbol{m}) - \boldsymbol{V}^{-1}\boldsymbol{\eta}(\boldsymbol{\theta})\right)\right]}^{\pi_\Delta(\delta|\theta)} \\ & \quad \times A_2 \, |\boldsymbol{V}|^{-1/2} \exp\left[\frac{1}{2}\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathrm{T}} \boldsymbol{V}^{-1} \boldsymbol{\eta}(\boldsymbol{\theta})\right] J(\boldsymbol{\theta}, \boldsymbol{\delta}), \end{aligned} \tag{12}$$

where A_0, A_1, and A_2 are all normalization constants, independent of $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$.

Because the first two lines in Eq. (12) cancel terms in the numerator of Eq. (11), the importance weight for sample k is

$$\omega \propto |\boldsymbol{V}|^{1/2} \exp\left[-\frac{1}{2}\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathrm{T}} \boldsymbol{V}^{-1} \boldsymbol{\eta}(\boldsymbol{\theta})\right] J^{-1}(\boldsymbol{\theta}, \boldsymbol{\delta}), \tag{13}$$

where second derivatives of $\boldsymbol{G}$ at the critical point have been neglected.

2.3.1 Importance Weights for the IES

Although the IES method is based on RML, the application to multimodal distributions is limited, as all samples share a common estimate of $\boldsymbol{G}$ estimated from the ensemble of realizations

$$\begin{aligned} \boldsymbol{G}\boldsymbol{C}_\theta\boldsymbol{G}^T &\approx \Xi_d\,\Xi_d^T \\ \boldsymbol{G} &\approx \Xi_d\,\Xi_\theta^T\boldsymbol{C}_\theta^{-1} \\ \boldsymbol{C}_\theta\boldsymbol{G}^T &\approx \Xi_\theta\,\Xi_d^T. \end{aligned} \tag{14}$$

Because $\boldsymbol{V}(\boldsymbol{\theta})$ and $J(\boldsymbol{\theta}, \boldsymbol{\delta})$ are the same for all samples, the computation of weights can be simplified. Neglecting the common multiplying constant, the IES approximation to the importance weight is

$$\omega \propto \exp\left[-\frac{1}{2}\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{\eta}(\boldsymbol{\theta})\right].$$

The only difference in weights is a result of differences in the term $\boldsymbol{\eta}(\boldsymbol{\theta})$, which requires computation of $\boldsymbol{G}$ from Eq. (14). For most practical problems, the ensemble size is smaller than the dimension of $\boldsymbol{\theta}$, so the pseudo-inverse is used to approximate the inverse of prior covariance matrix $\boldsymbol{C}_\theta^{-1}$.

2.3.2 Importance Weights for the Hybrid IES

The hybrid IES is also based on the RML method of sampling, but uses a different gain matrix for each sample while still avoiding the need for solving the adjoint system. For exact sampling from the posterior in strongly nonlinear problems, the computation of weights is unavoidable. To compute the weights $\{\omega_i\}_{i=1}^{N_e}$ of samples generated by the hybrid IES, the analytic sensitivity matrix $\boldsymbol{M}_\theta$ is used, which is $N_m \times N_\theta$. The derivative $\boldsymbol{G}_m$ of the objective function with respect to the intermediate variable $\boldsymbol{m}$ can be approximated by the ensemble samples. Finally, the computation of weights in Eq. (13) can be performed by the following forms

$$\begin{aligned} \boldsymbol{G}\boldsymbol{C}_\theta\boldsymbol{G}^T &\approx \Xi_d\,\Xi_m^{-1}\boldsymbol{M}_\theta\boldsymbol{C}_\theta\boldsymbol{M}_\theta^T\,\Xi_m^{-T}\,\Xi_d^T \\ \boldsymbol{G} &\approx \Xi_d\,\Xi_m^{-1}\boldsymbol{M}_\theta \\ \boldsymbol{C}_\theta\boldsymbol{G}^T &\approx \boldsymbol{C}_\theta\boldsymbol{M}_\theta^T\,\Xi_m^{-T}\,\Xi_d^T. \end{aligned}$$

The dimensions of Ξ_m are $N_m \times N_e$. Thus, the pseudo-inverse of Ξ_m is used in the computation of $\boldsymbol{G}$. For each sample, the terms of Eq. (13) are different. When $\boldsymbol{M}_\theta$ or $\boldsymbol{C}_\theta$ has the Toeplitz properties, circulant embedding described in Appendix A is used to reduce the computational cost of the matrix multiplication.

2.4 Excess Variance of Importance Weights

For highly nonlinear sampling problems, the variance in the log-weights should be expected to be large, since the RML proposal density is not identical to the target density. On the other hand, the actual spread in computed importance weights is larger than it should be for a number of reasons, including the fact that the minimization method used for computation of samples is approximate. The iterations are generally stopped before actual convergence, and the gradient is approximated from a low-rank ensemble. A different initial ensemble would result in a different final estimate of $\boldsymbol{G}$, $\boldsymbol{V}$, and $\det \boldsymbol{V}$. All of these will result in variability in the computation of weights, and the non-normalized log-weights will consequently have a large spread. If the so-called noisy log-weights are used directly to compute weighted forecasts, almost all the weight will fall on a single model realization.

For large Bayesian inverse problems of the type encountered in the geosciences, the likelihood is often difficult to evaluate, and noisy approximations to the likelihood must instead be used (Dunbar et al. 2022). When the likelihood is noisy, however, the transition kernel in MCMC, or equivalently the weighting of particles in importance sampling, will be affected by the noise (Alquier et al. 2016; Acerbi 2020). This noise must either be removed or be otherwise accounted for if the sampling is to be efficient. The problem of sampling with noisy importance weights has been reviewed by Akyildiz et al. (2017), who showed that denoising can be an effective approach. In the application to weighting of RML samples, the errors appear primarily in the evaluation of the proposal density, not in the evaluation of the likelihood as in most previous studies. Because the importance weights are ratios of likelihood to proposal density, however, the effect of noise in either term on the weight is similar.

2.4.1 A Model for Noise in the Log-Weights

For simplicity, we denote the logarithm of the weights on the particles as ω, so that

$$\omega = -\frac{1}{2}\log\det \boldsymbol{V} - \frac{1}{2}\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{\eta},$$

where

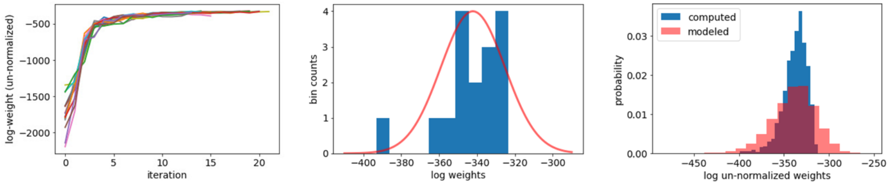
$$\boldsymbol{V} = \boldsymbol{C}_d + \boldsymbol{G}\boldsymbol{C}_\theta\boldsymbol{G}^{\mathrm{T}},$$

and

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) - \boldsymbol{d}^o - \boldsymbol{G}(\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{pr}}).$$

For a nonlinear problem, the sensitivities $\boldsymbol{G}$ at the minimizer will be variable, and since $\boldsymbol{G}$ and the data mismatch enter quadratically in ω, we expect the so-called true distribution of log-weights to be approximately chi-square. We additionally assumed a Gaussian model for the distribution of errors in the computed log-weights, which we also refer to as noise in the weights.

We can obtain an empirical characterization of the computational error by generating a number of realizations of the computed value of ω for the same prior sample $\boldsymbol{\theta}'$ but with different ensembles of realizations used for computation of $\boldsymbol{G}$. We did this for realizations of the monotonic transform of log-permeability by generating 16 independent ensembles of 199 realizations and augmenting each ensemble with another realization that was then common to each ensemble. Figure 1a shows the evolution of



(a) Evolution of log-weights of a single particle in 16 independent ensembles.

(b) Distribution of final log-weights for a unique sample from prior and Gaussian fit.

(c) Computed distribution of all log-weights from 8 ensembles of 200 realizations.

Fig. 1 The distributions of log-weights for the monotonic log-permeability transform

the log-weight on the single realization that was common to all 16 ensembles. The minimization was stopped when the iterations reached the terminated condition. (The particle that was common to all ensemble members stopped updating by iteration 10 in all ensembles, where the iteration number just contains the iteration that the mean of data mismatch is smaller than the last iteration.) Figure 1b shows the distribution of final values of ω (blue) and the Gaussian fit to the distribution of final values (red curve).

Estimating reasonable parameters in the chi-square model of the true distribution for large weights is more difficult than estimating the errors in the computation, partly because we do not have an empirical distribution of log-weights without computational noise. We instead used a trial-and-error approach in which the observed distribution of log-weights was compared with a Monte Carlo distribution of noisy samples from a chi-square distribution whose parameters were tuned to match the observed distribution. Figure 1c compares the distribution of RML-computed realizations with the realizations from the modeled distribution of noisy large weights.

The posterior distribution for noisy log-weights (Eq. (15)) is modeled as the product of a Gaussian likelihood model and a chi-square prior distribution,

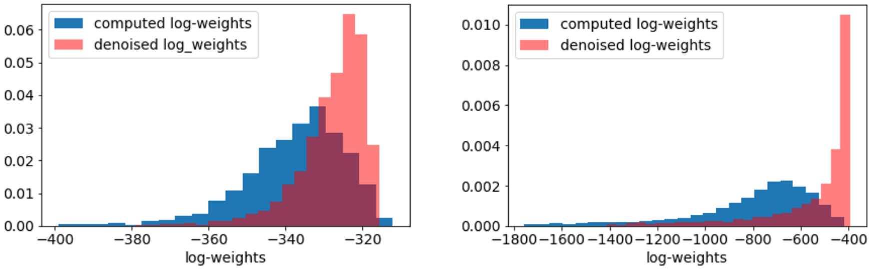
$$P(\omega|\omega^o) \propto P(\omega^o|\omega)P(\omega)$$

or

$$P(\omega|\omega^o) \propto \begin{cases} \exp\left(-\frac{(\omega-\omega^o)^2}{2\sigma_o^2}\right)\left(\frac{\omega-\omega^{\text{pr}}}{\sigma_{\text{pr}}}\right)^{\nu/2-1} \exp\left(-\frac{\omega-\omega^{\text{pr}}}{2\sigma_{\text{pr}}}\right) & \text{for} \quad \omega > \omega_{\text{pr}} \\ 0 & \text{for} \quad \omega \leq \omega_{\text{pr}}. \end{cases} \tag{15}$$

Once the parameters of the distribution have been estimated, the denoised weights are estimated by computing the maximum a posteriori values of the individual weights; that is, for each "observed" weight, we compute the maximizer of Eq. (15) to obtain the denoised weight.

For the monotonic log-permeability transform, with $\sigma_o = 16.9$, $\sigma_{\text{pr}} = 6$, and $\nu = 4$, we obtain the denoised log-weights shown in red in Fig. 2a. The effective sampling



(a) Monotonic log-permeability transformation.

(b) Non-monotonic log-permeability transformation.

Fig. 2 The distributions of log-weights after denoising (red colors)

efficiency $N_{\text{eff}}/N_e = 97.8/1600 \approx 0.109$, based on Kong's estimator Eq. (16),

$$N_{\text{Eff}} = \frac{1}{\sum_{k=1}^{N_e} \omega_k^2}, \tag{16}$$

where $\sum_{k=1}^{N_e} \omega_k = 1$.

The spread in the weights for the case with non-monotonic transform of log-permeability is much larger than in the case with monotonic permeability transform. First, the computation of weights appears to be less repeatable: Fig. 2a shows the evolution of non-normalized log-weights for the same sample when included in 16 otherwise independent ensembles. The spread of the final values for the common particle (Fig. 2b) is approximately five times as large for the non-monotonic case ($\sigma_o = 95.3$) as for the monotonic case ($\sigma_o = 16.9$). Presumably, the additional variability is a result of greater variability in $\boldsymbol{G}$ and the presence of more local minima. Additionally, the prior spread of log-weights appears to be larger, again because of multiple minima and the fact that the proposal distribution is farther from the target distribution in this case. For the non-monotonic log-permeability transform, with $\sigma_o = 95.3$, $\sigma_{\text{pr}} = 13$, and $\nu = 3$, we obtain the denoised log-weights shown in red in Fig. 2b. The effective sampling efficiency $N_{\text{eff}}/N_e = 27.2/1600 \approx 0.017$, based on Kong's estimator Eq. (16).

3 Applications of Weighting to Assimilation of Flow Data

In this section, two data assimilation methods (hybrid IES and IES) are applied to a two-dimensional, two-fluid-phase flow, incompressible problem with permeability transforms, $\boldsymbol{m} = f(\boldsymbol{\theta})$, of varying degrees of nonlinearity: a monotonic log-permeability transform and a non-monotonic transform. Here, $\boldsymbol{m}$ and $\boldsymbol{\theta}$ are the same size, which leads to a diagonal $\boldsymbol{M}_\theta$. For the two applications, the uncertain permeability field in the porous medium is estimated by assimilation of a time series of water rate observations at nine producing wells.

The state, u, of an incompressible and immiscible two-phase (aqueous (w) phase and oleic (o) phase) flow system is determined by the pressure $p(x,t)$ and saturation $s(x,t)$, which in this example are governed by

$$\begin{cases} -\nabla \cdot (K\lambda_*(s)\nabla p) = q, \\ \phi \dfrac{\partial s}{\partial t} + \nabla \cdot (f_*(s)v) = \dfrac{q_w}{\rho_w} \quad \text{in} \quad \Omega \times [0,T] \end{cases}, \tag{17}$$

with the boundary condition

$$v \cdot \mathbf{n} = 0 \quad \text{on} \quad \partial\Omega \times [0,T], \quad s(x,0) = 0 \quad \text{in} \quad \Omega = [0,2] \times [0,2],$$

where ϕ denotes the rock porosity, the source term q models sources and sinks, the fractional-flow function $f_*(s)$ measures the fraction of the total flow, $\lambda_*(s)$ is the mobility of the phase, K denotes the absolute permeability (assumed to be isotropic), q_w denotes the w phase source, and ρ_w denotes the density of the w phase. Since we only inject water and produce whatever reaches our producers, the source term for the saturation equation becomes

$$\frac{q_w}{\rho_w} = \max(q,0) + f(s)\min(q,0).$$

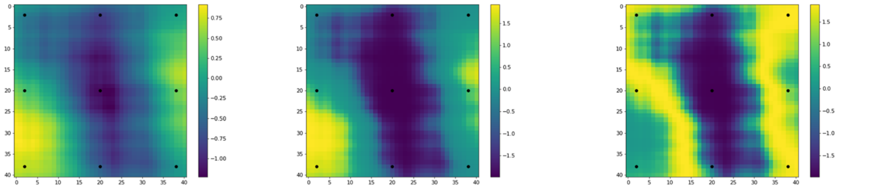
To close the model, we must supply expressions for the w phase and o phase

$$\begin{aligned} &v_i = -K\lambda_{*i}\nabla p, \quad q = \frac{q_w}{\rho_w} + \frac{q_o}{\rho_o}, \quad s_w + s_o = 1, \\ &p_w = p_o, \quad v = v_w + v_o, \quad \lambda_{*i}(s) = \frac{k_{ri}}{\mu_i}, \quad i = w, o. \end{aligned} \tag{18}$$

In Sect. 3.1, the hybrid IES methodology is compared with the IES methodology for the flow problem with a monotonic log-permeability transform. In Sect. 3.2, a similar comparison is made, but for the flow problem with a non-monotonic log-permeability transform, which has a multimodal posterior distribution of the model parameters. In both cases, the latent variable $\boldsymbol{\theta}$ is assumed to be multivariate Gaussian with covariance

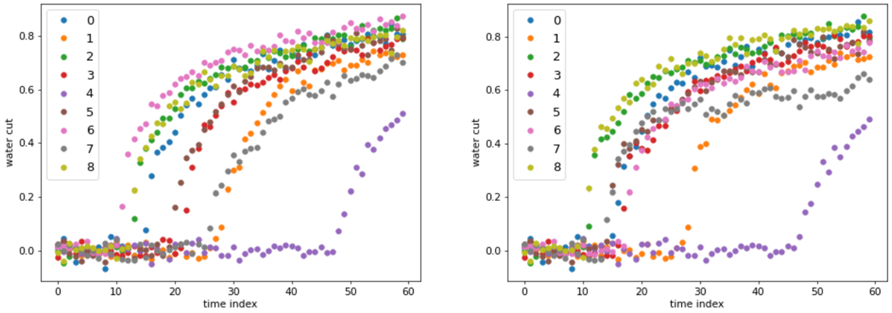
$$C_\theta(x,y) = \sigma_\theta^2 \left(1 - \frac{x^2+y^2}{\rho^2}\right) \exp\left(-\frac{x^2+y^2}{\rho^2}\right), \tag{19}$$

where x and y are the lags in the two spatial dimensions, and ρ is a measure of the correlation range. The permeability field for the data-generating model is a draw from a prior model with the range parameter for the correlation length $\rho = 1.1$ and standard deviation 0.8 for the monotonic and non-monotonic transforms. The true data-generating permeability values are as shown in Fig. 3. Figure 4 displays the corresponding water rate observations from the nine producing wells for both permeability transforms. To compare the results from the standard IES and hybrid IES, the ensemble size $N_e = 200$ for both methods.



(a) identity transform (b) monotonic transform (c) non-monotonic transform

Fig. 3 The “true” log-permeability fields used to generate production data and forecasts. Black dots show locations of producing wells



(a) monotonic transform (b) non-monotonic transform

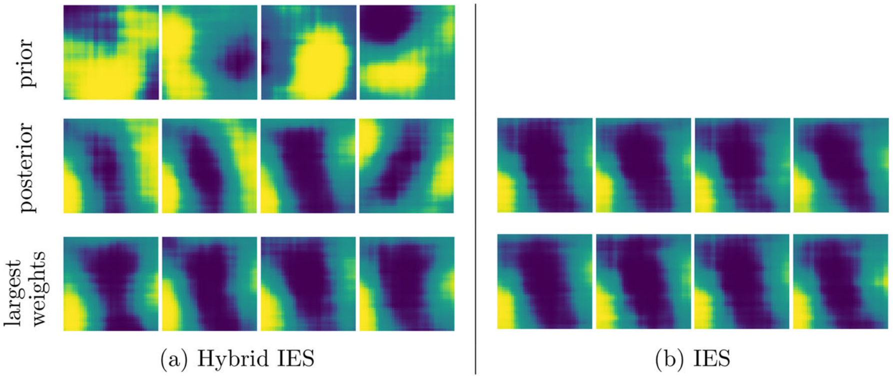
Fig. 4 Noisy observations of water cut in nine producing wells

The observation locations are distributed on a uniform 3×3 grid of the domain $[0.1, 1.9] \times [0.1, 1.9]$ as shown in Fig. 3 (black dots). The noise in the observations is assumed to be Gaussian and independent with standard deviation 0.02. The forward model (Eq. (18)) is solved by the two-point flux approximation (TPFA) scheme, which is a cell-centered finite-volume method (Aarnes et al. 2007). For the two test cases, the forward model is defined on a uniform 41×41 grid with time step $\Delta t = 0.1$. The dimension of the discrete parameter space is 1,681.

3.1 History Matching with a Monotonic Permeability Transform

To create a reservoir data assimilation test problem that is nonlinear but not obviously multimodal, a permeability transform was used that has characteristics similar to rock facies distributions, that is, regions with relatively uniform permeability and fairly sharp transitions between those regions. In some cases, the region occupied by the high-permeability facies is isolated and can be modeled well by a monotonic transformation of a Gaussian variable to log-permeability. To illustrate the effect of this type of nonlinearity on weighting in data assimilation, the transformation

$$m = \tanh\left(4\theta + 2\right) + \tanh\left(4\theta - 2\right) \tag{20}$$



(a) Hybrid IES

(b) IES

Fig. 5 Model realizations for monotonic transformation of log-permeability using the IES and hybrid IES. (The same prior ensemble is used for both methods)

is applied, where m and θ (scalars) denote the values of log-permeability and the latent Gaussian random variable in a cell, respectively. For the hybrid IES method, the gradient $\boldsymbol{M}_\theta$ of log-permeability $\boldsymbol{m}$ with respect to the Gaussian parameter $\boldsymbol{\theta}$ is necessary. The analytic derivative is given by

$$\frac{\mathrm{d}m}{\mathrm{d}\theta} = 8 - 4\tanh^2\left(4\theta + 2\right) - 4\tanh^2\left(4\theta - 2\right).$$

With this transformation, values of $\theta < 1$ are assigned $m \approx -2$, and values of $\theta > 1$ are assigned $m \approx 2$. The discretized form of the sensitivity $\boldsymbol{M}_\theta$ is diagonal with

$$\boldsymbol{M}_\theta = \begin{bmatrix} \frac{\mathrm{d}m_1}{\mathrm{d}\theta_1} & 0 & \dots & 0 \\ 0 & \frac{\mathrm{d}m_2}{\mathrm{d}\theta_2} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \frac{\mathrm{d}m_{N_\theta}}{\mathrm{d}\theta_{N_\theta}} \end{bmatrix},$$

while the covariance operator $\boldsymbol{C}_\theta$ of Eq. (19) is dense but block-Toeplitz (Zimmerman 1989; Dietrich and Newsam 1997). Multiplication by $\boldsymbol{M}_\theta$ is trivial, but the product $\boldsymbol{C}_\theta\big(\boldsymbol{M}_\theta^T(\Xi_{m_i})^{-T}\big)$ is computed using the Toeplitz property of $\boldsymbol{C}_\theta$ as described in Appendix A.

When the hybrid IES is applied to this problem, the gain matrices are potentially different for each realization, so it should be expected that some realizations will converge to local minima with small probability mass if the posterior has multiple modes. The samples with the largest weights are likely to be similar, however. Figure 5 shows the log-permeability fields for the first four prior realizations (top row) and corresponding posterior realizations of log-permeability values for both methods (middle row). The variability in the posterior realizations is smaller than the variability in the prior realizations, but still fairly large. The log-permeability values of the four posterior realizations from the hybrid IES with the largest weights (Fig. 5a (bottom row)) are much more similar, indicating that for this problem, importance weighting for the

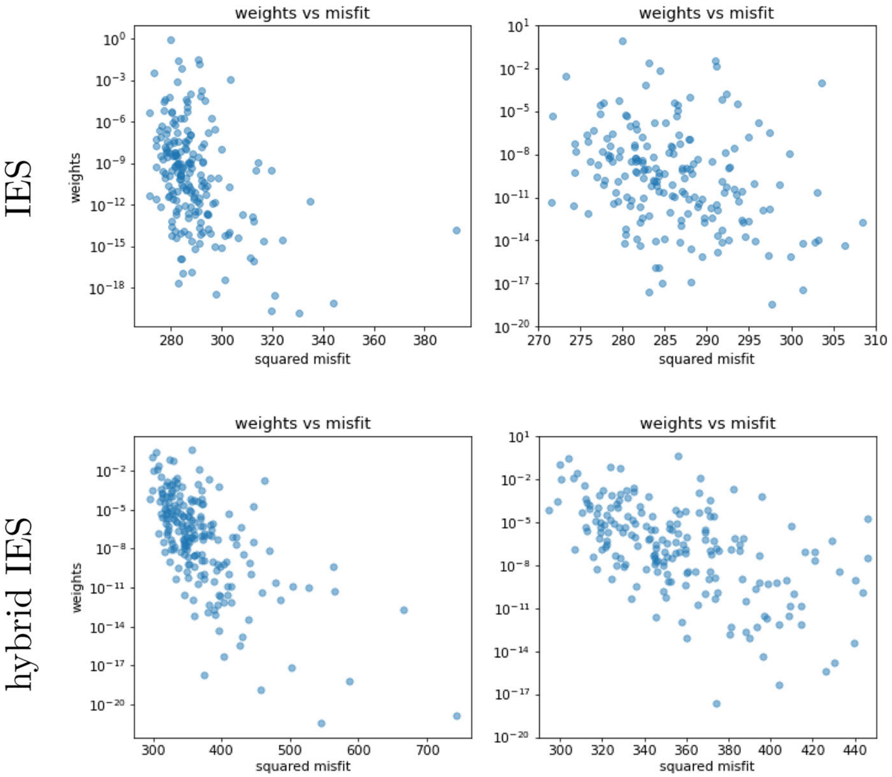


Fig. 6 The weights versus misfits for the monotonic log-permeability transform using the IES (top row) and hybrid IES (bottom row). Blue points show computed weights

hybrid IES is beneficial in selecting realizations from the posterior that are similar to the true model.

In contrast, when the IES is used for data assimilation with the same prior ensemble of realizations, the first four posterior realizations obtained from the IES (Fig. 5b (top row)) are similar to the four realizations with the largest weights (Fig. 5b (bottom row)). As the same gain matrix is used for all samples generated from the standard IES, the variability among posterior approximate realizations is smaller for the IES than for the hybrid IES, and unlike the situation with the hybrid IES, the unweighted and weighted posterior means obtained using the standard IES are almost identical (not shown).

For nonlinear problems such as this, it would be reasonable to expect the approximate posterior realizations with largest weights to have small data mismatch with observations. To investigate this hypothesis, a cross-plot of the weights versus squared data misfits generated using the IES (top row) and hybrid IES (bottom row) is shown in Fig. 6. Although the weights for the hybrid IES method are clearly correlated with squared data misfit, and the models with largest data misfit always have very small weights, there is considerable variability in weights even for small data misfit. The more important observation is that the nonlinearity in $g(\cdot)$ increases the variability in weights, and the mean of the squared misfit (369) using the hybrid IES is substantially larger than the expected value for samples from the posterior (270), while the weighted mean is 330.

Since the IES method generated less variability in the posterior realizations, the spread of weights in the IES method is expected to be smaller than the spread of weights in application of the hybrid IES. In fact, however, the spread of weights is quite large and almost independent of the data mismatch (Fig. 6 (top row)). This appears to be a result of errors in computation of $\boldsymbol{G}$, an underestimate of the magnitude of $\boldsymbol{V}$, and the degeneracy inherent in weighting of optimal proposals based purely on an ensemble of particles.

Because the range of the covariance of the permeability field is relatively large compared to the domain of interest, the observation locations are spatially distributed, and the production data from all wells are matched fairly well by the weighted and unweighted samples (Fig. 7). The posteriori means of the log-permeability fields (not shown) look similar to the truth, except that the truth is somewhat “rougher.”

The justification for data assimilation or history matching of subsurface models is generally to provide accurate assessments of future reservoir behavior. Figure 7 show the quality of the match to observed data and the predictability of future performance of the unweighted and weighted posterior ensembles at three representative wells. For this case, the differences in predictability between the weighted and unweighted realizations are small, although the prediction interval is narrower for the weighted hybrid IES, due in part to the small effective sample size.

For a Gauss-linear inverse problem, there should be no correlation between the weight on a sample from the posterior and the data mismatch—in fact, for this case, the weights should be uniform when a minimization-based sampling approach is used. For the nonlinear two-dimensional porous flow example with a monotonic log-permeability transform, the log-weights did correlate with data mismatch when the standard IES method was used for data assimilation ($r = -0.416$) and when the hybrid IES method was used ($r = -0.647$). In both cases, the quality of the data mismatch provided some information on the weighting that should be applied to a particle.

3.2 History Matching with Non-monotonic Permeability Transform

In this section, the problem of history matching and uncertainty quantification for permeability fields with a low-permeability “background” and connected high-permeability “channels” are considered. Again, soft thresholding of the Gaussian random variables is used to generate regions with relatively sharp transition to a different facies. The non-monotonic permeability transform is given by

$$m = 2\tanh\big(4\theta + 2\big) + \tanh\big(2 - 4\theta\big) - 1, \tag{21}$$

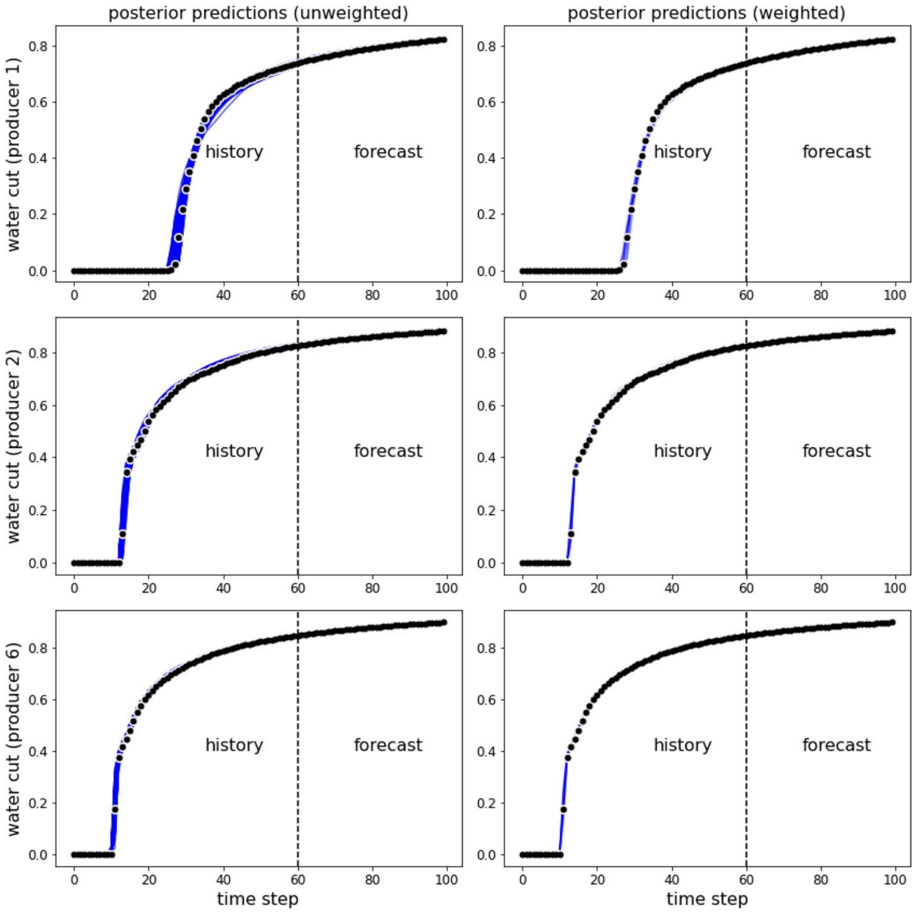


Fig. 7 The posterior predictions of wells 1, 2, and 6 using the unweighted and weighted hybrid IES for the monotonic transform. Black points show true observations

where θ is again the prior Gaussian random variable. The corresponding derivative of log-permeability with respect to the Gaussian latent variable, required for the hybrid IES, is then

$$\frac{\mathrm{d}m}{\mathrm{d}\theta} = 4 - 8\tanh^2\left(4\theta + 2\right) + 4\tanh^2\left(2 - 4\theta\right).$$

The true data-generating log-permeability field for this test problem is shown in Fig. 3c, and water cut observations for the nine producing wells are plotted in Fig. 4b. Although the data are not noticeably different from the data in the monotonic case (Fig. 4a), the presence of the channel facies makes the problem slower to converge to a mode and more likely to converge to a mode with small probability mass.

For the non-monotonic transform case, the variability of posterior realizations from the hybrid IES is larger than in the monotonic transform case, as illustrated by the first

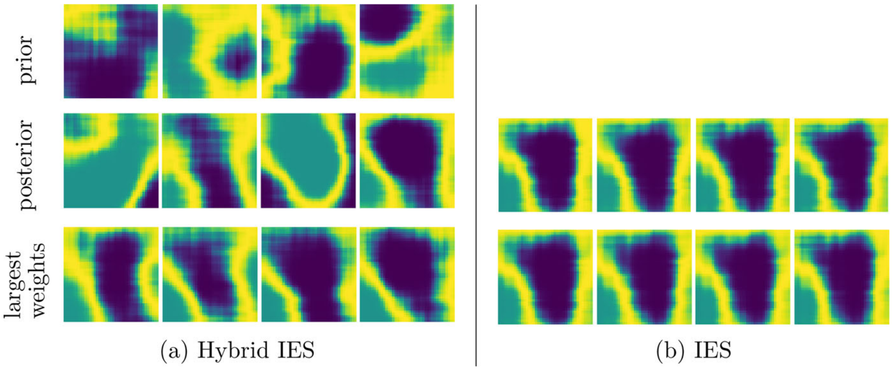


Fig. 8 Model realizations for a non-monotonic transform of log-permeability using the IES and hybrid IES

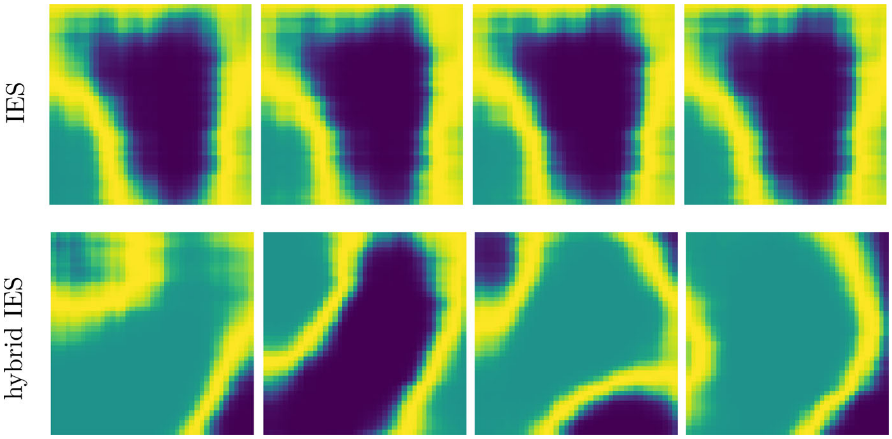


Fig. 9 Model realizations with the smallest weights for non-monotonic transformations using the IES (top row) and hybrid IES (bottom row)

four prior realizations and corresponding posterior realizations (Fig. 8a). In this case, the differences are due to the use of the analytic sensitivity $\boldsymbol{M}_\theta$ in the hybrid IES, which allows realizations to converge to different local minima. The posterior realizations with the largest importance weights (Fig. 8a (bottom row)) show reasonable similarity to the true field. When the IES method was used for data assimilation, the first four posterior realizations (Fig. 8b (top row)) and the four realizations with the largest weights (Fig. 8b (bottom row)) were almost identical. The lack of diversity results in the unweighted and weighted posterior means being very similar when a standard IES is used. Importance weighting has very little effect for the IES method on this problem. The effective sample size of the IES for the two cases is low, however, because the posterior spread has been underestimated (Chen and Oliver 2017; Ba and Jiang 2021).

In addition to the greater diversity in the realizations compared to the monotonic case, the importance weights and the data mismatch are also much more diverse for the non-monotonic transform (Fig. 10) when the hybrid IES is used. In the non-monotonic

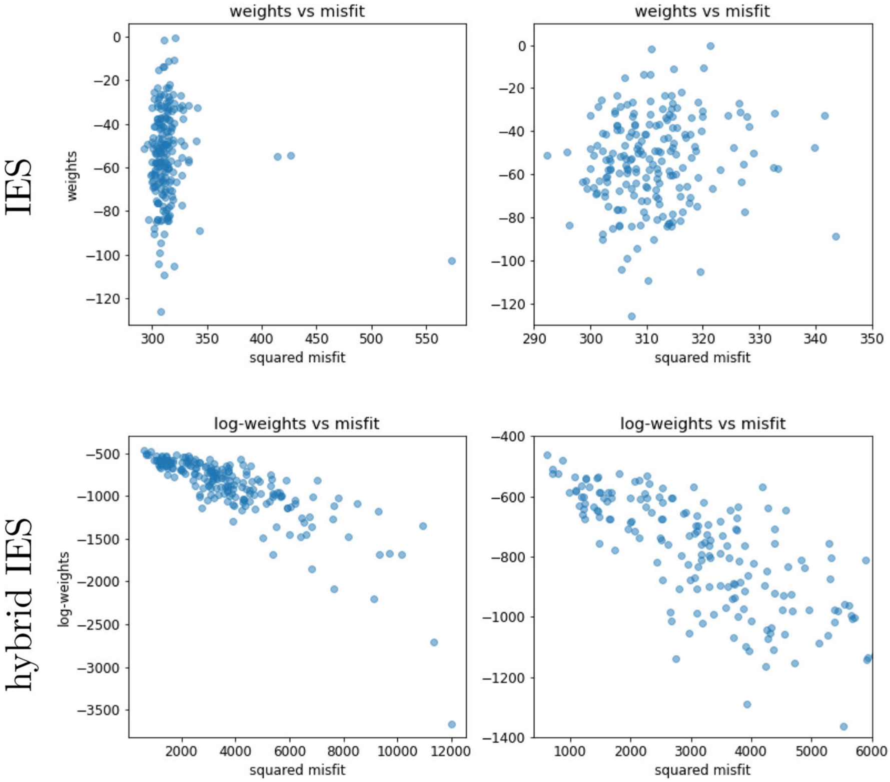


Fig. 10 The non-normalized log-weights versus misfits for the non-monotonic transform using the IES and hybrid IES. Blue points show computed log-weights

case, the expected mean of the data misfit part of the log-likelihood is still 270 (half the number of observations). Figure 10 shows, however, that the data misfits of most posterior samples are concentrated in the interval $[1,000, 6,000]$. The posterior mean of unweighted data misfits is 3,794—approximately 14 times the expected value. The posterior mean of data misfits for weighted realizations, on the other hand, is 617, which is still larger than expected but much smaller than the mean for the unweighted realizations.

Weighted and unweighted mean log-permeability fields for the non-monotonic permeability transform are shown in Fig. 11. The middle and bottom rows of Fig. 11 show mean values of log-permeability at location $(x, 1.75)$ for two different sets of results using different convergence trajectories (i.e., different values for the multiplier of λ in Levenberg–Marquardt minimization). The unweighted mean for the hybrid IES bears limited similarity to the true field and shows little connectivity of the high-permeability facies. This is a result of averaging with many dissimilar realizations that are not all well calibrated. The weighted mean looks much more like the truth, as it puts more weight on samples with higher probability mass. The effect of importance weighting is perhaps more obvious in the posterior distribution of predictions of water cut. The

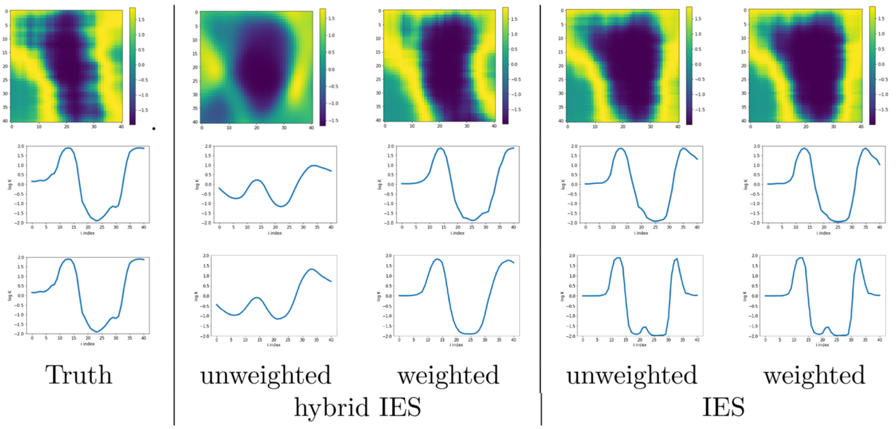


Fig. 11 The true log-permeability (left) and the unweighted and weighted posterior means using the hybrid IES (middle) and IES (right) for the non-monotonic transform

spread in the unweighted predictions is large, even during the history-matched period (Fig. 12 (left column)) and much larger than expected given the observation error. On the other hand, the quality of the weighted posterior realizations (right column) is excellent, except for well 1. The main problem with the weighted ensemble appears to be that the spread is too small, resulting from the small effective sample size.

For the more highly nonlinear two-dimensional porous flow example with non-monotonic log-permeability transform, the correlation between the importance log-weight and data mismatch was very high ($r = -0.813$) when the hybrid IES was used for data assimilation, and the data mismatch after calibration could serve as a useful tool for eliminating samples with small weights. For the standard IES, however, the approximated weights were clearly not accurate, and the correlation between log-weight and data mismatch was correspondingly small ($r = -0.087$). In this case, the data mismatch would not have provided a useful proxy for weighting.

3.3 History Matching Using the Hybrid IES

The hybrid IES algorithm is somewhat more complex than the standard IES. A naïve implementation would be very costly for realistic geoscience problems because of the additional matrix–vector operations required to compute the update step. Additionally, the weights computed from the hybrid IES are only an approximation of the weights computed in the randomized maximum likelihood method, and the weights will be noisy as a result. The following subsections address solutions to these issues.

3.3.1 Efficiency of the Update Step in the Hybrid IES

Timing experiments showed that the cost of the proposed method for computing the update step using fast Fourier transform (FFT) (see Appendix A) was substantially decreased compared with the straightforward approach used by (Oliver 2022). The

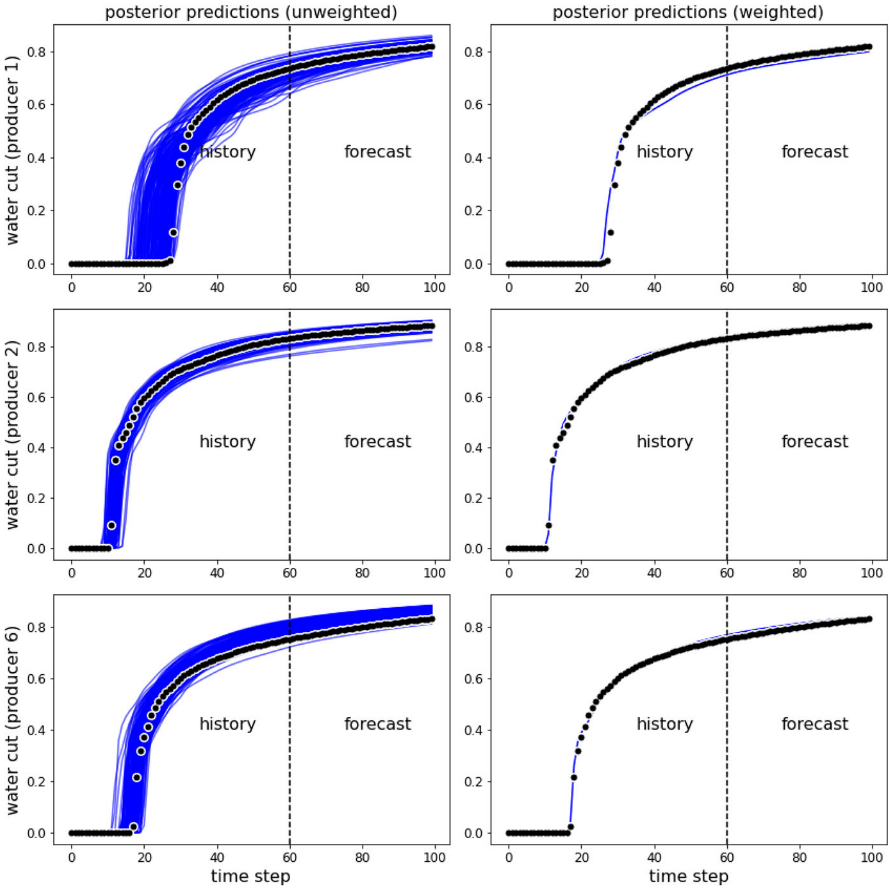


Fig. 12 The posterior predictions for wells 1, 2, and 6 using the unweighted and weighted hybrid IES for the non-monotonic transform. Black points show true observations

computational complexity of the hybrid method stems from the need to perform N_e minimizations using individual gradient estimates and the cost to generate N_e prior ensemble members from a high-dimensional parameter space. The covariance matrix is generally dense and large in the discretized space. While the sampling step from the prior distribution contributes to the cost of the hybrid IES, the cost is dominated by minimization of the objective function. For the two-phase flow example, the cost to generate 200 prior samples is about 0.054 s (timing should be considered illustrative, but for reference, all results were obtained on a computer with an i7-5500U@2.40GHz × 4 processor with 7.5 GiB memory and a 64-bit operating system). The run time is approximately 31,000 times faster when using the FFT with nonnegative definite minimal embeddings compared with a method that uses the Cholesky decomposition and matrix multiplication, as shown in Table 1. Note that the increase in cost for the non-monotonic case is a result of the varying number of iterations required for the minimizer to converge for the different settings. In the case

Table 1 Computational cost of the hybrid IES (units = second)

	Whole algotithm	Prior ensemble	$\boldsymbol{C}_\theta \boldsymbol{M}_\theta^T (\Xi_m)^{-T}$
FFT for monotonic	1427.82	0.054	0.0894
Non-FFT for monotonic	3206.29	1693.65	0.1824
FFT for non-monotonic	1691.62	0.054	0.0961
Non-FFT non-monotonic	4354.54	1697.35	0.2451

of the hybrid IES method, the cost to generate the N_e hybrid gradient is dominated by the cost to perform N_e matrix–vector multiplications with different cost functions. Hence, the computational complexity for the hybrid IES can be expected to be greater than the cost for the standard IES method. For the weighted hybrid IES, there is an additional cost incurred in the computation of the weights. Although several of the terms in the weights can be obtained at low cost through the ensemble approximations that were used for the Hessian, the N_e times matrix multiplications are necessary to compute the weights. When FFT was applied to compute hybrid gradients, the cost of computing $\boldsymbol{C}_\theta \boldsymbol{M}_\theta^T (\Xi_m)^{-T}$ was reduced from 0.18 s to 0.089 s for each sample relative to the case in which the matrix–vector multiplication did not use FFT (i.e., the FFT method was twice as fast).

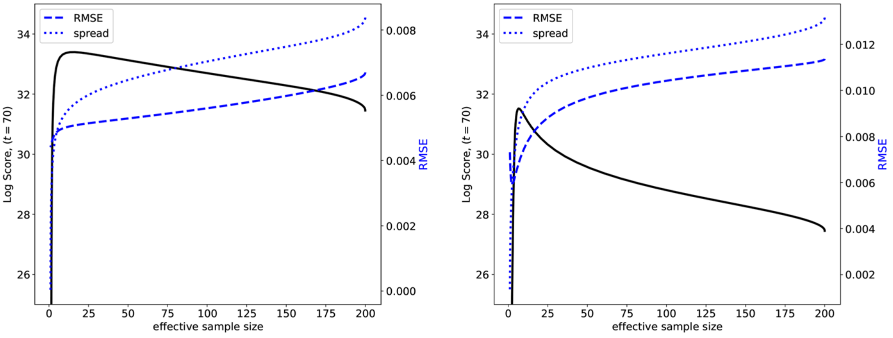
All computational costs, including the cost of minimization, could be reduced through careful modification of the algorithms. In particular, the efficiency of the weighted hybrid IES could be improved by tempering the objective function at early iterations to avoid convergence to local minima with small weights.

3.3.2 Effect of Denoising Weights on Predictability

Unweighted posterior realizations generated by minimization of a stochastic cost function are often described as well history-matched, but differences in the quality of the match to data between some realizations and observations are too large in practice to be explained by observation error. To investigate the potential benefit of weighting the samples and of different degrees of denoising, we compute the accuracy of probabilistic predictions beyond the history-matching period for data assimilation using the hybrid IES. (Optimal weighting was not investigated for the standard IES, as weighting was not useful for the non-monotonic log-permeability case.) For this investigation, observations used in history matching end at $t = 60$, and predictions are evaluated at $t = 70$ for all nine producers using the "log score" (Good 1952; Gneiting and Raftery 2007). The logarithmic score evaluates the probability of the outcome given a probability density function (pdf) empirically defined by the ensemble of predictions. The log score rewards both accuracy and sharpness of the forecasts. A higher log score signifies better probabilistic prediction,

$$\mathrm{Log}S(P, u) = -\log(\mathrm{p(u)}),$$

where a Gaussian approximation of p has been used.



(a) monotonic

(b) non-monotonic

Fig. 13 Evaluation of optimal regularization of weights for forecast predictability using the hybrid IES

The effective sample size and the log score of the forecasts were computed at $t = 70$ for all nine wells for a range of degrees of regularization of weights. Regularization of log weights was accomplished by applying a power transformation with exponents between 0 and 1 to the computed weights. The effective sample size (ESS) is affected by the degree of regularization—the ESS is 200 if equal weighting is used (i.e., if a power transformation with a very small exponent is applied). When the exponent is close to 1, one ends up using the weights as computed without denoising or regularization. Figure 13a shows predictability scores for a range of degrees of regularization using the hybrid iterative smoother with the monotonic permeability transform. Figure 13b shows corresponding results for the non-monotonic permeability transform. The solid black curves in both cases show the log score, which is somewhat small for both cases when the effective sample size is small, even though only the so-called best realizations are used for the forecast. The poor predictability for small effective sample size is a result of the small spread in the ensemble, so that even small inaccuracy of the prediction is highly improbable. As the effective sample size increases, the predictability initially increases rapidly because of the increase in the spread, but when the exponent of the power transform is decreased sufficiently, the predictability gradually decreases as more so-called bad samples are added. The impact of bad samples is smaller in the monotonic case than in the non-monotonic case because the root-mean-square error (RMSE) in the worst samples is smaller in the monotonic case.

Figure 14 compares unweighted predictions with weighted predictions and denoised weighted predictions for one of the wells (producer 4) in both two-dimensional porous flow examples. For producer 4, the agreement between the forecast from the data-generating model and weighted forecasts is nearly perfect, although the quality of the agreement at some other wells is lower. Better forecast predictability as measured by the log score is obtained using denoised importance weights, as described in Sect. 2.4.1, although in both examples (monotonic and non-monotonic log permeability transforms) the correct level of denoising was difficult to determine.

Although the effect of denoising is quantitatively different for the monotonic and non-monotonic permeability transforms, in both cases the best predictability is obtained when the weights are regularized such that the effective sample size is

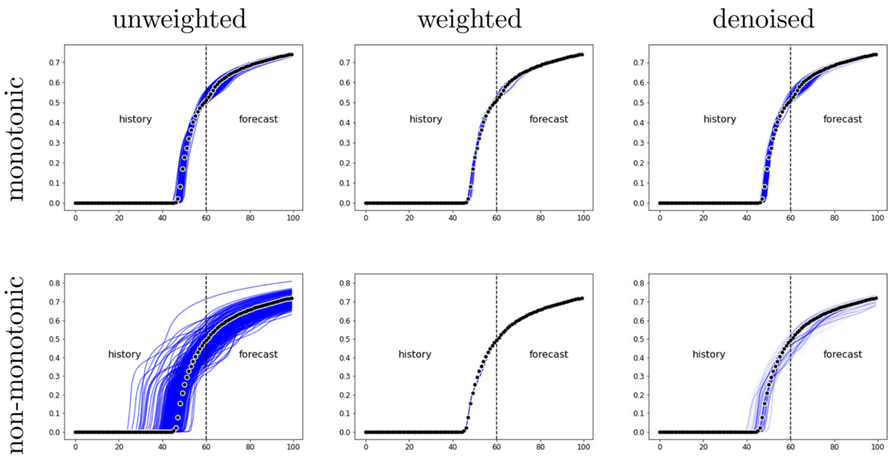


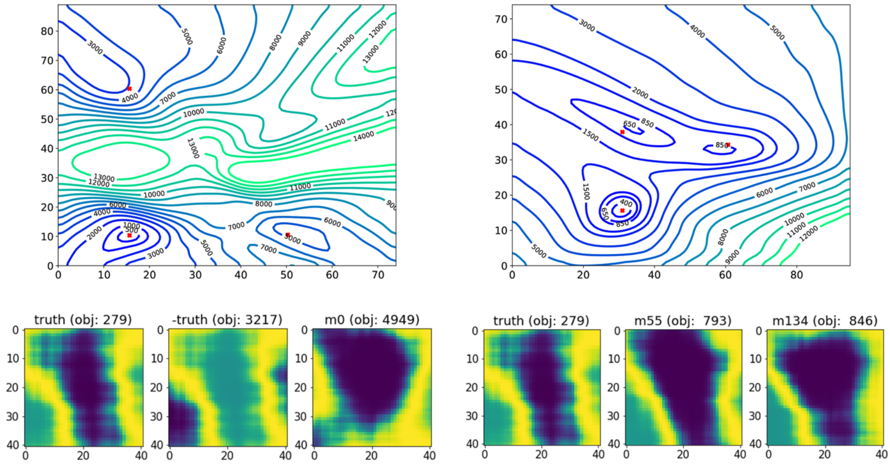
Fig. 14 The posterior distribution of forecasts conditioned to data to $t = 60$ using the hybrid IES method. The black dots show the observed data

intermediate between the ESS for unweighted samples and the ESS for the weights computed using Eq. (13).

4 Landscape of the Posterior

The efficiency of the hybrid iterative smoother for sampling the posterior was relatively low in the flow example with the non-monotonic permeability transformation. For an ensemble size of 200, the effective sample size after denoising was approximately 5.5. The small size makes probabilistic inference difficult—even though the mean weighted forecasts were generally accurate, the estimates of the uncertainty often were not. In order to obtain an effective ensemble size of approximately 40 after weighting, it would be necessary to use an initial ensemble size of approximately 1,600. As the efficiency of the randomized maximum likelihood sampler using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm for minimization with the gradient computed from the adjoint system was similar to the efficiency of the hybrid IES for a similar problem (Ba et al. 2022), it seems likely that the low efficiency is a result of the roughness of the posterior landscape rather than a problem with minimization.

The efficiency of sampling algorithms depends strongly on the landscape of the pdf to be sampled and on the goal of the sampling. If the objective is simply to sample in the neighborhood of the maximum a posteriori point, then using exact gradients is not always beneficial, especially if the log posterior is characterized by multiple scales—a smooth, long-range feature that is approximately quadratic and shorter-range fluctuations to the surface (Plecháč and Simpson 2020). If the posterior pdf is characterized, however, by a small number of nearly equivalent modes, then ensemble methods may fail to converge (Oliver and Chen 2018; Dunbar et al. 2022). In the numerical example with non-monotonic transformation of the log-permeability, the IES converged to the maximum a posteriori (MAP), but failed to sample other local



(a) Subspace generated by three very different realizations.

(b)Subspace generated by three similar realizations.

Fig. 15 The fitness landscape for the non-monotonic porous flow problem

minima. In order to clarify the behavior, the fitness landscape was evaluated in the neighborhood of the true data-generating model and over a larger region.

The dimension of the model space is too large to visualize the landscape of the posterior directly. Instead, two illustrations of the fitness landscape for the non-monotonic log-permeability transform case were created by selecting three realizations of the model parameters to define a subspace of the model space. From three realizations, an orthonormal basis was constructed, and the log-likelihood function on a grid containing the three realizations was evaluated. In the first plot of the fitness landscape (Fig. 15a), the subspace contains both the true model, θ_{true}, and the negative of the true model, $-\theta_{\text{true}}$. (Both are equally probably before conditioning to the data.) A third realization with relatively low weight was included to provide an independent basis vector needed for the two-dimensional subspace. In this case, there is a fairly large energy barrier separating the modes containing θ_{true} and $-\theta_{\text{true}}$ and a smaller barrier separating θ_{true} from θ_0. In Fig. 15b, the subspace contains the truth and two realizations with large posterior weights. Again, each realization appears to lie in a separate mode of the likelihood, although that cannot be verified without examining the surface in higher dimensions. In any case, the posterior landscape is complex, and accurate gradients may be of limited usefulness if the goal is to locate the global minimum.

5 Summary and Conclusions

Iterative, ensemble-based data assimilation methods for sampling the posterior distribution are based on minimization of stochastic objective functions. These methods of sampling are approximate when the mapping from parameters to observations is nonlinear. To correct the sampling, importance weighting can be used. It is, however, generally difficult or costly to compute the importance weights if derivatives are computed from the adjoint system. On the other hand, the cost is relatively low for ensemble-based methods which avoid the need for adjoints. It was shown that standard products from hybrid iterative ensemble smoothers could be used to compute approximations to the importance weights. The weights computed in this way are noisy—largely because of low-rank stochastic approximations of derivatives. Denoising of the importance weights increased the effective sample size, decreased the RMSE in the estimate of the posterior model mean, and increased the predictability of future reservoir behavior.

Although the IES method converged more quickly than the hybrid IES in the numerical test problem with a multimodal posterior, the posterior realizations from the IES appear to be samples from a single mode. In some cases, the IES sampled from the mode with the highest probability, so that while the uncertainty was underestimated, the fit to data was good. In other cases, however, the IES samples were centered on a mode with lower mass and the fit was not as good. The posterior mean model for the multimodal problems using the IES was very sensitive to the choice of minimization parameters. Weighting was not effective in this case because the posteriori distributions of samples were not from the critical points of the stochastic cost function. We did not evaluate the possibility of combining multiple posterior ensembles from the IES, but it seems likely that the weighted results from a large number of ensembles would provide a better representation of the posterior.

Finally, it was noted that the posterior landscape of the inverse problem for the two-dimensional, two-phase immiscible flow appears to be multimodal when the permeability field is generated from a transformation that creates channel-like features of high permeability in a low-permeability background. The characteristics of the posterior distribution have implications for the types of data assimilation methods that can be expected to provide reasonable assessment of the uncertainty. For the flow problem with “channel-like” geology, it appeared that the standard IES may be capable of generating an ensemble of well-calibrated models, but the spread in that case was artificially small. The hybrid IES method provided an ensemble of models with much greater variability, but weighting of the calibrated realizations was necessary for posterior inference, and the effective ensemble size was much smaller than the actual ensemble size.

Author Contributions Both authors contributed to the study conception and design, the writing of the manuscript, and preparation for submission.

Funding Open access funding provided by NORCE Norwegian Research Centre AS. The work of the first author was supported by the 2023 Guangzhou Basic and Applied Basic Research Project (no. 2023A04J0035) and by the Talent Special Projects of School-level Scientific Research Programs under Guangdong Polytechnic Normal University (no. 2022SDKYA025). The work of the second author was funded by the Research Council of Norway through the Petromaks2 program (NFR project number 295002) and by industry partners of the NORCE research cooperative research project “Assimilating 4D Seismic Data: Big Data Into Big Models”: Equinor Energy AS, Lundin Energy Norway AS, Repsol Norge AS, Shell Global Solutions International B.V., TotalEnergies E&P Norge AS and Wintershall Dea Norge AS.

Data Availability Not applicable.

Code Availability Code used to produce results in this manuscript will be made available on the GitHub repository.

Declarations

Conflict of interest The authors declare no competing interests.

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Permission to publish has been granted where necessary.

Appendix A: Fast Matrix–Vector Multiplication

To accelerate the update step in the hybrid IES in large inverse problems, we apply circulant embedding for fast multiplication of Toeplitz matrices. Assuming that the forward model is defined on a uniform $(n_x + 1) \times (n_y + 1)$ grid, the dimensions of the corresponding covariance matrix are $(n_x + 1)^2 \times (n_y + 1)^2$, which is extremely large in typical geoscience data assimilation applications. If the prior model covariance function is stationary and the model variables are defined on a two-dimensional regular, equispaced grid, the covariance matrix $\boldsymbol{C}_\theta$ is symmetric level-2 block-Toeplitz. To reduce the storage and the computation cost, the embedding Toeplitz covariance is applied to circulant matrices. Circulant matrix–vector products can then be computed efficiently by fast Fourier transform (FFT) (Zimmerman 1989; Dietrich and Newsam 1997). A circulant matrix $\check{\boldsymbol{C}}$ is a Toeplitz matrix that has its first column $\check{\boldsymbol{c}}$ periodic. A Toeplitz matrix $\boldsymbol{C}_\theta$ can always be augmented to generate a circulant matrix $\check{\boldsymbol{C}}$. This process is called embedding, which can be written as

$$\boldsymbol{C}_\theta = \mathcal{M}^\dagger \check{\boldsymbol{C}} \mathcal{M}(\mathcal{H}(\boldsymbol{\xi})) = \mathcal{M}^\dagger \mathcal{F}^{[-2]}\Big(\mathcal{F}^{[2]}(\check{\boldsymbol{c}})\mathcal{F}^{[2]}(\check{\boldsymbol{\xi}})\Big),$$

where the operation $\mathcal{M}(\mathcal{H}(\cdot))$ injects and embeds $\boldsymbol{\xi}$ into $\check{\boldsymbol{\xi}}$, and $\mathcal{F}^{[2]}$ is evaluated by the FFT. Under the assumption of stationarity, the random field $Y(x, y)$ has correlation function $r(x, y)$ that depends only on the separation of variables.

Let h_x and h_y be constants denoting, respectively, the horizontal and vertical mesh size of a two-dimensional rectangular domain formed by the points (x_i, y_j), where

$$\begin{cases} x_i = ih_x, & 0 \leq i \leq n_x \\ y_j = jh_y, & 0 \leq j \leq n_y. \end{cases}$$

The ordering of the grid nodes is done from left to right, bottom to top. Thus, the correlation matrix $\boldsymbol{R}$ is block-Toeplitz (Zimmerman 1989; Dietrich and Newsam 1997), which is symmetric and uniquely characterized by the first block row

$$(\boldsymbol{R}_0, \boldsymbol{R}_1, \ldots, \boldsymbol{R}_{n_y}),$$

with $\boldsymbol{R}_j$ being the square of dimension $n_x + 1$. Based on the node ordering, $\boldsymbol{R}_j$ has first row and first column entries, respectively, given by

$$r(ih_x, jh_y)_{i=0}^{n_x} \quad \text{and} \quad r(-ih_x, jh_y)_{i=0}^{n_x}.$$

The blocks $\boldsymbol{R}_j$ are symmetric only if the correlation function has the special form $r(|x|, |y|)$. When $\boldsymbol{R}_j$ is Toeplitz for $j \geq 1$, it is uniquely characterized by its first row and first column which are written as

$$(r_{0j}, r_{1j}, \ldots, r_{n_x j}) \quad \text{and} \quad (r_{0j}, r_{-1j}, \ldots, r_{-n_x j})^T.$$

The minimal circulant embedding of $\boldsymbol{R}_j$ that ensures that the embedding matrix has an even dimension (for FFT computation) is then given by the square circulant matrix $\boldsymbol{S}_j$ of dimension $2(n_x + 1)$, for which the first row is

$$(r_{0j}, r_{1j}, \ldots, r_{n_x j}, \phi_j, r_{-n_x j}, \ldots, r_{-1j}). \tag{A1}$$

The circulant embedding of Toeplitz matrices can be constructed by Eq. (A1). Once constructed, the circulant embedding is used to quickly generate samples from the prior and to evaluate products such as $\boldsymbol{C}_\theta \boldsymbol{G}^T$, which would otherwise be infeasible.

Appendix B: List of Symbols

List of Symbols

N_θ	Number of model parameters
$N(\cdot, \cdot)$	Gaussian distribution
N_d	Dimension of observational space
C_m	Covariance matrix of Gaussian prior for m
C_d	Covariance matrix of ϵ
N_m	Number of intermediate variable
C_θ	Covariance matrix of Gaussian prior for θ

β	Hyparameter or other reservoir properties
θ_l^i	ith sample for the lth iteration
$\mathcal{B}$	Generic forward model operator
θ^{pr}	Mean of Gaussian prior
$\pi_D(d^{\mathrm{o}})$	Normalization constant
$Q(\theta)$	Negative log likelihood function
$\pi_\Theta(\theta \| d^{\mathrm{o}})$	Posterior distribution
$\pi_{\Theta\Delta}(\theta, \delta)$	Target distribution
$q_{\Theta'\Delta'}(\theta', \delta')$	Proposal distribution of (θ', δ')
$p_{\Theta\Delta}(\theta, \delta)$	Proposal distribution of (θ, δ)
κ_{ri}	Relative permeability of phase i
θ,δ	Samples of target distribution
θ',δ'	Samples of Gaussian $q_{\theta'\Delta'}(\theta', \delta')$
Z	Composite observation operator
f	Nonlinear transform from θ to m
$n(\theta')$	Total number of critical points
G	Differential operator of g wrt θ
G_m	Differential operator of g wrt m
$p(u)$	Predictive distribution of u
$\check{C}$	Circulant matrix augmented by C_θ
$P(\omega^o \| \omega)$	Likelihood function of weights
M_θ	Differential operator of m wrt θ
I_N	N-dimensional identity matrix
$P(\omega \| \omega^o)$	Posterior distribution of weights
Boldletter	Vector or matrix of corresponding letter
Unboldletter	Analytic representation or scalar
N_e	Number of samples
N_{eff}	Effective sample size
m	Intermediate variable
d^o	Observations
w	Weights
ϵ	Observation error
l	Iteration index
J	Jacobian determinant
σ_o	Standard deviation of ω^o
V, $\eta(\theta)$	Auxiliary variables
$u(x)$	Model state
$p(x, t)$	Pressure
$s(x, t)$	Saturation
Ω	Spatial domain
L	Square root of C_M
ω^o	Observation of weights
$P(\omega)$	Prior of weights
h_x,h_y	Mesh size
n_x,n_y	Number of grid nodes
g	Observation operator

μ_i	Viscosity of phase i
Ξ_θ	Ensemble deviation for θ
1_N	N-dimensional one-vector
Ξ_m	Ensemble deviation for m
Ξ_d	Ensemble deviation for d
C	Covariance operator
$\check{c}$	First column of $\check{C}$
σ	Standard deviation of C
λ_l	lth Regularization parameter
ρ	Correlation length of C
$\delta\theta_l$	Increment for the lth iteration
ν	Freedom of Chi-square prior for ω
σ_{pr}	Standard deviation of prior for ω

References

Aarnes JE, Gimse T, Lie KA (2007) An introduction to the numerics of flow in porous media using Matlab. In: Hasle G, Lie K, Quak E (eds) Geometric modelling, numerical simulation, and optimization. Springer, pp 265–306

Acerbi L (2020) Variational Bayesian Monte Carlo with noisy likelihoods. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in neural information processing systems, vol 33. Curran Associates Inc, pp 8211–8222

Akyildiz OD, Marino IP, Míguez J (2017) Adaptive noisy importance sampling for stochastic optimization. In: 2017 IEEE 7th international workshop on computational advances in multi-sensor adaptive processing (CAMSAP), pp 1–5

Alquier P, Friel N, Everitt R, Boland A (2016) Noisy Monte Carlo: convergence of Markov chains with approximate transition kernels. Stat Comput 26(1):29–47. https://doi.org/10.1007/s11222-014-9521-x

Ba Y, Jiang L (2021) A two-stage variable-separation Kalman filter for data assimilation. J Comput Phys 434:110244. https://doi.org/10.1016/j.jcp.2021.110244

Ba Y, de Wiljes J, Oliver DS, Reich S (2022) Randomized maximum likelihood based posterior sampling. Comput Geosci 26(1):217–239. https://doi.org/10.1007/s10596-021-10100-y

Bardsley J, Solonen A, Haario H, Laine M (2014) Randomize-then-optimize: a method for sampling from posterior distributions in nonlinear inverse problems. SIAM J Sci Comput 36(4):A1895–A1910. https://doi.org/10.1137/140964023

Bardsley JM, Cui T, Marzouk YM, Wang Z (2020) Scalable optimization-based sampling on function space. SIAM J Sci Comput 42(2):A1317–A1347. https://doi.org/10.1137/19m1245220

Chen Y, Oliver DS (2012) Ensemble randomized maximum likelihood method as an iterative ensemble smoother. Math Geosci 44(1):1–26. https://doi.org/10.1007/s10596-016-9599-7

Chen Y, Oliver DS (2013) Levenberg–Marquardt forms of the iterative ensemble smoother for efficient history matching and uncertainty quantification. Comput Geosci 17(4):689–703. https://doi.org/10.1007/s10596-013-9351-5

Chen Y, Oliver DS (2017) Localization and regularization for iterative ensemble smoothers. Comput Geosci 21(1):13–30. https://doi.org/10.1007/s10596-016-9599-7

Dietrich CR, Newsam GN (1997) Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. SIAM J Sci Comput 18(4):1088–1107. https://doi.org/10.1137/S1064827592240555

Doucet A, Godsill S, Andrieu C (2000) On sequential Monte Carlo sampling methods for Bayesian filtering. Stat Comput 10(3):197–208. https://doi.org/10.1023/A:1008935410038

Dunbar ORA, Duncan AB, Stuart AM, Wolfram MT (2022) Ensemble inference methods for models with noisy and expensive likelihoods. SIAM J Appl Dyn Syst 21(2):1539–1572. https://doi.org/10.1137/21M1410853

Emerick AA, Reynolds AC (2013) Ensemble smoother with multiple data assimilation. Comput Geosci UK 55:3–15. https://doi.org/10.1016/j.cageo.2012.03.011

Evensen G (1994) Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J Geophys Res Oceans 99(C5):10143–10162. https://doi.org/10.1029/94JC00572

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Am Statist Assoc 102(477):359–378. https://doi.org/10.1198/016214506000001437

Good IJ (1952) Rational decisions. J R Stat Soc Ser B Stat Methodol 14(1):107–114

Kitanidis PK (1995) Quasi-linear geostatistical theory for inversing. Water Resour Res 31(10):2411–2419. https://doi.org/10.1029/95WR01945

Martin J, Wilcox L, Burstedde C, Ghattas O (2012) A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. SIAM J Sci Comput 34(3):A1460–A1487. https://doi.org/10.1137/110845598

Maschio C, Schiozer DJ (2014) Bayesian history matching using artificial neural network and Markov Chain Monte Carlo. J Petrol Sci Eng 123:62–71. https://doi.org/10.1016/j.petrol.2014.05.016

Mohamed L, Calderhead B, Filippone M, Christie M, Girolami M (2012) Population MCMC methods for history matching and uncertainty quantification. Comput Geosci 16(2):423–436. https://doi.org/10.1007/s10596-011-9232-8

Oliver DS, He N, Reynolds AC (1996) Conditioning permeability fields to pressure data. In: Proceedings of the European conference on the mathematics of oil recovery, pp 1–11. https://doi.org/10.3997/2214-4609.201406884

Oliver DS (2017) Metropolized randomized maximum likelihood for improved sampling from multimodal distributions. SIAM/ASA J Uncertain Quantif 5(1):259–277. https://doi.org/10.1137/15M1033320

Oliver DS (2022) Hybrid iterative ensemble smoother for history matching of hierarchical models. Math Geosci 54(8):1289–1313. https://doi.org/10.1007/s11004-022-10014-0

Oliver DS, Chen Y (2011) Recent progress on reservoir history matching: a review. Comput Geosci 15(1):185–221. https://doi.org/10.1007/s10596-010-9194-2

Oliver DS, Chen Y (2018) Data assimilation in truncated plurigaussian models: impact of the truncation map. Math Geosci 50(8):867–893. https://doi.org/10.1007/s11004-018-9753-y

Oliver DS, Cunha LB, Reynolds AC (1997) Markov chain Monte Carlo methods for conditioning a permeability field to pressure data. Math Geol 29(1):61–91. https://doi.org/10.1007/BF02769620

Oliver DS, Zhang Y, Phale HA, Chen Y (2011) Distributed parameter and state estimation in petroleum reservoirs. Comput Fluids 46(1):70–77. https://doi.org/10.1016/j.compfluid.2010.10.003

Papaspiliopoulos O, Roberts GO, Sköld M (2007) A general framework for the parameterization of hierarchical models. Statist Sci 22(1):59–73. https://doi.org/10.1214/088342307000000014

Plecháč P, Simpson G (2020) Sampling from rough energy landscapes. Commun Math Sci 18(8):2271–2303. https://doi.org/10.4310/CMS.2020.v18.n8.a9

Reich S (2011) A dynamical systems framework for intermittent data assimilation. BIT Numer Math 51(1):235–249. https://doi.org/10.1007/s10543-010-0302-4

Snyder C, Bengtsson T, Morzfeld M (2015) Performance bounds for particle filters using the optimal proposal. Mon Weather Rev 143(11):4750–4761. https://doi.org/10.1175/MWR-D-15-0144.1

Tavassoli Z, Carter JN, King PR (2005) An analysis of history matching errors. Comput Geosci 9(2):99–123. https://doi.org/10.1007/s10596-005-9001-7

van Leeuwen PJ, Künsch HR, Nerger L, Potthast R, Reich S (2019) Particle filters for high-dimensional geoscience applications: a review. Quart J Roy Meteorol Soc 145(723):2335–2365. https://doi.org/10.1002/qj.3551

Wang K, Bui-Thanh T, Ghattas O (2018) A randomized maximum a posteriori method for posterior sampling of high dimensional nonlinear Bayesian inverse problems. SIAM J Sci Comput 40(1):A142–A171. https://doi.org/10.1137/16M1060625

Zhang F, Reynolds AC, Oliver DS (2003) The impact of upscaling errors on conditioning a stochastic channel to pressure data. SPE J 8(1):13–21. https://doi.org/10.2118/83679-PA

Zimmerman DL (1989) Computationally exploitable structure of covariance matrices and generalized convariance matrices in spatial models. J Stat Comput Simul 321–2:1–15. https://doi.org/10.1080/00949658908811149