Computational Geosciences (2018) 22:885–908
https://doi.org/10.1007/s10596-018-9731-y

ORIGINAL PAPER

Analysis of iterative ensemble smoothers for solving inverse problems

Geir Evensen[1,2]

Received: 26 July 2017 / Accepted: 11 February 2018 / Published online: 3 March 2018

Abstract

This paper examines the properties of the Iterated Ensemble Smoother (IES) and the Multiple Data Assimilation Ensemble Smoother (ES–MDA) for solving the history matching problem. The iterative methods are compared with the standard Ensemble Smoother (ES) to improve the understanding of the similarities and differences between them. We derive the three smoothers from Bayes' theorem for a scalar case which allows us to compare the equations solved by the three methods, and we can better understand which assumptions are applied and their consequences. When working with a scalar model, it is possible to use a vast ensemble size, and we can construct the sample distributions for both priors and posteriors, as well as intermediate iterates. For a linear model, all three methods give the same result. For a nonlinear model, the iterative methods improve on the ES result, but the two iterative methods converge to different solutions, and it is not clear which should be the preferred choice. It is clear that the ensemble of cost functions used to define the IES solution does not represent an exact sampling of the posterior-Bayes' probability density function. Also, the use of an ensemble representation for the gradient in IES introduces an additional approximation compared to using an exact analytic gradient. For ES–MDA, the convergence, as a function of increasing number of uniform update steps, is studied for a huge ensemble size. We illustrate that ES–MDA converges to a solution that differs from the Bayesian posterior. The convergence is also examined using a realistic sample size to study the impact of the number of realizations relative to the number of update steps. We have run multiple ES–MDA experiments to examine the impact of using different schemes for choosing the lengths of the update steps, and we have tried to understand which properties of the inverse problem imply that a non-uniform update step length is beneficial. Finally, we have examined the smoother methods with a highly nonlinear model to examine their properties and limitations in more extreme situations.

Keywords Ensemble smoothers · IES · ES–MDA · Data assimilation · History matching

1 Introduction

Ensemble methods for data assimilation and parameter estimation [9, 11, 12] are now well established as a standard tool in the reservoir-engineering community for history matching reservoir models. Following the first application of Ensemble Kalman Filter (EnKF) with a reservoir simulation model by Nævdal et al. [19], there is now a large number of publications that address the estimation of parameters in reservoir simulation models using EnKF. We refer to the review by Aanonsen et al. [1] and references therein.

Skjervheim et al. [24] introduced the use of Ensemble Smoother (ES) as an alternative to the sequential EnKF for history matching reservoir models and showed that similar performance and results were obtained using ES and EnKF in a reservoir test case.

van Leeuwen and Evensen [26] initially proposed ES and also found that EnKF provides superior results to ES in an application with an ocean circulation model. EnKF and ES both solve the same Bayesian formulation, which in the case of EnKF is written as a recursion in time under the assumption of a Markov reservoir model and measurements that are independent in time. Thus ES differs from EnKF by computing a global update of the model parameters using all the observations simultaneously rather than using recursive updates in time. For linear dynamical models and measurement operators, EnKF and ES provide

✉ Geir Evensen
geir.evensen@iris.no

[1] International Research Institute of Stavanger, Bergen, Norway

[2] Nansen Environmental and Remote Sensing Center, Bergen, Norway

identical solutions as is shown by Evensen [10]. However, for nonlinear dynamical models, and in particular models with chaotic dynamics, EnKF is shown to be superior to ES [13, 26]. The reason is that the recursive updates keep the model on track and close to the true solution represented by the measurements. The acceptable performance of ES with a reservoir model was attributed by Skjervheim et al. [24] to the relatively "weakly nonlinear" nature of reservoir models.

In ES one integrates the whole ensemble of model realizations once to generate a prediction. Then the prior ensemble of uncertain parameters is updated using the "Kalman Filter" equations with all data assimilated simultaneously.

Finally, the model is rerun using the updated parameters to create the final history-matched ensemble of model predictions. Thus, ES solves a parameter-estimation problem that is easy to grasp by reservoir engineers since it is very similar to the concept used in most other software developed for history matching.

Following the introduction of ES for use in history matching by Skjervheim et al. [24], two iterative variants of the smoother formulation were introduced. Chen and Oliver [4, 5] published Iterative ES (IES) which was initially named Ensemble Randomized Likelihood (EnRML) [16, 21]. Emerick and Reynolds [7, 8] developed Multiple-Data-Assimilation ES (ES–MDA). The iterations of the smoother update turn out to partly resolve issues with nonlinearity and lead to better results than what is obtained by ES. There is now a range of new smoother developments and applications based on the original iterative variants, e.g., Bocquet and Sakov [2], Luo et al. [18], Iglesias [14, 15], Le et al. [17], and Rafiee and Reynolds [22].

In Evensen and Eikrem [Strategies for conditioning reservoir models on rate data using ensemble smoothers, under review] ES, ES–MDA, and IES were used with a real reservoir model. They observed that IES and ES–MDA with a different number of update steps gave slightly different results. In particular, the variance obtained from IES was lower than the one from ES–MDA with 16 MDA steps, which again was smaller than the one from ES–MDA with eight MDA steps. It was also challenging to determine which is the preferred scheme of IES and ES–MDA, the number of update steps to use in ES–MDA, and whether there was any point in using non-uniform step lengths in ES–MDA.

In this paper, we will discuss the data-assimilation methods ES, ES–MDA, and IES for a simple scalar problem and try to explain the similarities and differences between these smoothers to understand better what to expect when we use them. We start by restating the history matching problem in the next section and present a set of equations and their assumptions and illustrate how they can be used to derive ES, ES–MDA, and IES. Then, in Section 3, we present a detailed derivation of ES, ES–MDA, and IES for the scalar case while discussing the approximations and simplifications used. In Section 4, we run several experiments with the different smoothers to illustrate and discuss their properties with a weakly nonlinear and monotonic scalar model. Finally, in Section 5, we study the highly nonlinear case to establish limits of applicability of the methods and to better understand their limitations.

2 History matching problem

We start by formally restating the history matching problem as usually formulated in the petroleum industry. A first fundamental assumption is that we have a perfect forward model

$$\mathbf{y} = \mathbf{g}(\mathbf{x}). \tag{1}$$

From evaluating the model operator $\mathbf{g}(\mathbf{x})$, given a realization of the model parameters $\mathbf{x} \in \Re^n$, we uniquely determine the predicted measurements $\mathbf{y} \in \Re^m$ (corresponding to the real measurements $\mathbf{d} \in \Re^m$). Here n is the number of parameters and m the number of measurements. We have measurements $\mathbf{d}$ of $\mathbf{y}$, and we want to use the measurements to estimate the variable $\mathbf{x}$, i.e., we are solving a standard inverse problem.

In history matching, it is common to define a prior for the parameters since we usually will have more degrees of freedom in the parameters than we have independent information in the measurements. Bayes' theorem with a perfect model gives the joint posterior pdf for $\mathbf{x}$ and $\mathbf{y}$ as

$$\begin{aligned} f(\mathbf{x},\mathbf{y}|\mathbf{d}) &\propto f(\mathbf{x},\mathbf{y})f(\mathbf{d}|\mathbf{y}) \\ &= f(\mathbf{x})f(\mathbf{y}|\mathbf{x})f(\mathbf{d}|\mathbf{y}) \\ &= f(\mathbf{x})\delta(\mathbf{y}-\mathbf{g}(\mathbf{x}))f(\mathbf{d}|\mathbf{y}), \end{aligned} \tag{2}$$

where the transition density $f(\mathbf{y}|\mathbf{x})$ becomes the Dirac delta function in the case with no model errors. We are interested in the marginal pdf for $\mathbf{x}$, which we obtain by integrating Eq. 2 over $\mathbf{y}$, giving

$$\begin{aligned} f(\mathbf{x}|\mathbf{d}) &\propto \int f(\mathbf{x})\delta(\mathbf{y}-\mathbf{g}(\mathbf{x}))f(\mathbf{d}|\mathbf{y})d\mathbf{y} \\ &= f(\mathbf{x})f(\mathbf{d}|\mathbf{g}(\mathbf{x})). \end{aligned} \tag{3}$$

To be able to solve for the posterior pdf of $\mathbf{x}$, we need to impose another assumption. In the ensemble methods, the approach is to assume a Gaussian prior $f(\mathbf{x})$ and likelihood $f(\mathbf{d}|\mathbf{g}(\mathbf{x}))$. The marginal posterior pdf in Eq. 3 then becomes

$$\begin{aligned} f(\mathbf{x}|\mathbf{d}) \propto \exp - \tfrac{1}{2}\Big(&\big(\mathbf{x}-\mathbf{x}^{\mathrm{f}}\big)^{\mathrm{T}}\mathbf{C}_{xx}^{-1}\big(\mathbf{x}-\mathbf{x}^{\mathrm{f}}\big) \\ &+\big(\mathbf{g}(\mathbf{x})-\mathbf{d}\big)^{\mathrm{T}}\mathbf{C}_{dd}^{-1}\big(\mathbf{g}(\mathbf{x})-\mathbf{d}\big)\Big), \end{aligned} \tag{4}$$

where $\mathbf{x}^{\mathrm{f}}$ is the prior estimate for $\mathbf{x}$, $\mathbf{C}_{xx} \in \Re^{n\times n}$ is the error covariance of $\mathbf{x}^{\mathrm{f}}$, and $\mathbf{C}_{dd} \in \Re^{m\times m}$ is the error covariance of the measurements $\mathbf{d}$. We have dropped a superscript "f" which is often used on $\mathbf{C}_{xx}$. Maximizing $f(\mathbf{x}|\mathbf{d})$ is equivalent to minimizing the cost function

$$\begin{aligned}\mathcal{J}(\mathbf{x}) = &\big(\mathbf{x}-\mathbf{x}^{\mathrm{f}}\big)^{\mathrm{T}}\mathbf{C}_{xx}^{-1}\big(\mathbf{x}-\mathbf{x}^{\mathrm{f}}\big)\\ &+\big(\mathbf{g}(\mathbf{x})-\mathbf{d}\big)^{\mathrm{T}}\mathbf{C}_{dd}^{-1}\big(\mathbf{g}(\mathbf{x})-\mathbf{d}\big).\end{aligned} \tag{5}$$

Most methods for history matching are based on the assumptions of a perfect model and Gaussian priors. The posterior is still nonlinear and non-Gaussian due to the non-linear model $\mathbf{g}(\mathbf{x})$. In cases with, e.g., channelized reservoirs, the prior for $\mathbf{x}$ becomes non-Gaussian; however, in most cases we can represent non-Gaussian parameters by underlying Gaussian parameters, so the assumption of Gaussian priors holds.

In the case of a linear model $\mathbf{y} = \mathbf{G}\mathbf{x}$, we can derive a direct solution for the minimum of the cost function (5), and this corresponds to the Kalman-Filter update equations

$$\mathbf{x}^{\mathrm{a}} = \mathbf{x}^{\mathrm{f}} + \mathbf{K}(\mathbf{d}-\mathbf{G}\mathbf{x}), \tag{6}$$

$$\mathbf{C}_{xx}^{\mathrm{a}} = (\mathbf{I}-\mathbf{K}\mathbf{G})\mathbf{C}_{xx}, \tag{7}$$

$$\mathbf{K} = \mathbf{C}_{xx}\mathbf{G}\big(\mathbf{G}\mathbf{C}_{xx}\mathbf{G}^{\mathrm{T}}+\mathbf{C}_{dd}\big)^{-1}, \tag{8}$$

where the superscript "a" denote the analysis update. Here $\mathbf{K}$ is the standard Kalman gain matrix.

Continuing in the linear case, we can use an ensemble representation of the error covariances $\mathbf{C}_{xx}^{\mathrm{e}} = \mathbf{A}\mathbf{A}^{\mathrm{T}}/(n_{\mathrm{e}}-1)$ as in the EnKF where $\mathbf{A} \in \Re^{n\times n_{\mathrm{e}}}$ contains the ensemble anomalies (ensemble members with the ensemble average subtracted), and n_{e} is the number of ensemble members. We can then rewrite the Kalman-Filter equations as

$$\mathbf{x}_j^{\mathrm{a}} = \mathbf{x}_j^{\mathrm{f}} + \mathbf{K}^{\mathrm{e}}(\mathbf{d}_j-\mathbf{G}\mathbf{x}_j), \tag{9}$$

$$\mathbf{K}^{\mathrm{e}} = \mathbf{C}_{xx}^{\mathrm{e}}\mathbf{G}\big(\mathbf{G}\mathbf{C}_{xx}^{\mathrm{e}}\mathbf{G}^{\mathrm{T}}+\mathbf{C}_{dd}\big)^{-1}, \tag{10}$$

and it is shown by Evensen [11] that with an infinite ensemble size, the update in Eq. 9 implies Eq. 7. Here, $\mathbf{d}_j = \mathbf{d} + \boldsymbol{\epsilon}_j$ denotes perturbed observations [3]. It is also easy to show that Eqs. 9 and 10 can be derived by minimizing the following cost function for each of $\mathbf{x}_j$, i.e.,

$$\begin{aligned}\mathcal{J}(\mathbf{x}_j) = &\big(\mathbf{x}_j-\mathbf{x}_j^{\mathrm{f}}\big)^{\mathrm{T}}\mathbf{C}_{xx}^{-1}\big(\mathbf{x}_j-\mathbf{x}_j^{\mathrm{f}}\big)\\ &+\big(\mathbf{G}\mathbf{x}_j-\mathbf{d}_j\big)^{\mathrm{T}}\mathbf{C}_{dd}^{-1}\big(\mathbf{G}\mathbf{x}_j-\mathbf{d}_j\big).\end{aligned} \tag{11}$$

In the linear case, ES and any method that minimizes the cost function (11) will correctly sample the posterior Gaussian pdf $f(\mathbf{x}|\mathbf{d})$.

In the nonlinear case, we may write the cost function (11) as

$$\begin{aligned}\mathcal{J}(\mathbf{x}_j) = &\big(\mathbf{x}_j-\mathbf{x}_j^{\mathrm{f}}\big)^{\mathrm{T}}\mathbf{C}_{xx}^{-1}\big(\mathbf{x}_j-\mathbf{x}_j^{\mathrm{f}}\big)\\ &+\big(\mathbf{g}(\mathbf{x}_j)-\mathbf{d}_j\big)^{\mathrm{T}}\mathbf{C}_{dd}^{-1}\big(\mathbf{g}(\mathbf{x}_j)-\mathbf{d}_j\big),\end{aligned} \tag{12}$$

which is the cost function being approximately minimized using the IES, but as will be seen below, the minimizing solutions will no longer exactly sample the posterior non-Gaussian distribution.

The ES update equations can in the nonlinear case be derived from Eq. 5 to get the nonlinear analogs of Eqs. 6–8. Then, we introduce the ensemble approximation to obtain

$$\mathbf{x}_j^{\mathrm{a}} = \mathbf{x}_j^{\mathrm{f}} + \mathbf{K}^{\mathrm{e}}\Big(\mathbf{d}_j-\mathbf{g}\big(\mathbf{x}_j^{\mathrm{f}}\big)\Big), \tag{13}$$

$$\mathbf{K}^{\mathrm{e}} = \mathbf{C}_{xx}^{\mathrm{e}}\mathbf{g}'(\mathbf{x}^{\mathrm{f}})\left(\mathbf{g}'(\mathbf{x}^{\mathrm{f}})\mathbf{C}_{xx}^{\mathrm{e}}\mathbf{g}'(\mathbf{x}^{\mathrm{f}})^{\mathrm{T}}+\mathbf{C}_{dd}\right)^{-1}, \tag{14}$$

where the tangent-linear operator $\mathbf{g}'$ is evaluated at the mean of the prior ensemble. Note that these equations can also be derived directly from the cost function (12), which is the starting point for IES. We will show how to obtain these equations in the scalar case, and how we in the ensemble formulation can replace the tangent-linear operator $\mathbf{g}'(\mathbf{x}^{\mathrm{f}})$ with an ensemble representation.

The purpose of this discussion is to show that ES, and also ES–MDA, can be derived from the same cost function that is minimized using IES. Furthermore, we formally link ES and ES–MDA to Bayes. We have also established a link between ES, ES–MDA, and IES that can be used to explain the methods. For now, we note that they all sample the posterior distribution in the Gauss-linear case.

Direct minimization of the cost function (5) is possible using a gradient method where we usually compute the gradient from an adjoint model. Alternatively, genetic sampling algorithms can be used to sample the posterior pdf in Eq. 4. However, a problem with such methods is that they require very many model predictions to converge, and they can only be used to estimate a few, $\mathcal{O}(10)$, parameters due to the size of the parameter space. It will be shown below how the ensemble methods replace the tangent-linear operator or gradient with an ensemble representation and thereby eliminate the need for adjoint calculations.

Note also that the Iterative EnKF (IEnKF) as discussed by Sakov et al. [23] and Bocquet and Sakov [2] solves the same kind of problem as given by the marginal conditional pdf in Eq. 3 or the cost function in Eq. 5. Sakov et al. [23] derived IEnKF for state estimation where the model state at the time t_i is updated using measurements of the state at time t_{i+1}. The purpose was to handle nonlinear dynamical models and observation operators better. However, the approach used in IEnKF is similar to IES and ES–MDA where we estimate parameters $\mathbf{x}$ using measurements of the nonlinear model prediction in Eq. 1. IEnKF solves for the update of the mean in the ensemble subspace spanned by the ensemble anomalies as

$$\mathbf{x}^{\mathrm{a}} = \mathbf{x}^{\mathrm{f}} + \mathbf{A}\mathbf{w}. \tag{15}$$

Thus, the problem is reduced to compute a vector of coefficients $\mathbf{w} \in \Re^{n_{\mathrm{e}}}$ by minimizing the cost function (5) where Eq. 15 is used to write it as a cost function for $\mathbf{w}$. Finally, one computes the updated ensemble anomalies by sampling perturbations utilizing an estimate of the error covariance of the analysis obtained from approximating the inverse of the Hessian of the cost function. The IEnKF is an exciting method for solving the history matching problem, but it is not practical in its standard form with the vast ensemble size used in this paper.

3 Derivation of the smoothers

We will from now on, for clarity, consider a scalar case with a single measurement. The posterior marginal pdf in Eq. 3 is written as

$$\begin{aligned} f(x|d) &\propto f(x)f(d|g(x)) \\ &\propto \exp -\tfrac{1}{2}\Big(\big(x-x^{\mathrm{f}}\big)C_{xx}^{-1}\big(x-x^{\mathrm{f}}\big)\Big) \\ &\quad\times \exp -\tfrac{1}{2}\Big(\big(g(x)-d\big)C_{dd}^{-1}\big(g(x)-d\big)\Big). \end{aligned} \tag{16}$$

The prior pdf and likelihood are assumed to be Gaussian, and we represent the Gaussian priors by ensembles of realizations

$$\begin{aligned} f(x) &= \mathcal{N}\big(x^{\mathrm{f}}, C_{xx}\big) \rightarrow \Big\{x_j^{\mathrm{f}}\Big\}, \\ f(d|g(x)) &= \mathcal{N}\big(d, C_{dd}\big) \rightarrow \Big\{d_j\Big\}. \end{aligned} \tag{17}$$

Then, corresponding to each pair of realizations x_j^{f} and d_j we can compute a posterior realization x_j by minimizing the cost function

$$\begin{aligned} \mathcal{J}(x_j) = &\big(x_j - x_j^{\mathrm{f}}\big)C_{xx}^{-1}\big(x_j - x_j^{\mathrm{f}}\big) \\ &+\big(g(x_j)-d_j\big)C_{dd}^{-1}\big(g(x_j)-d_j\big), \end{aligned} \tag{18}$$

which measures the distance between x_j and a prior value x_j^{f} and the distance between the prediction $y_j = g(x_j)$ and a measurement d_j. The two terms are weighted by the variances of the prior and the measurement, respectively, and Eq. 18 defines a least-squares solution for x_j.

In the nonlinear case, the posterior ensemble obtained by minimizing the cost function (18) for x_j, $j = 1, n_{\mathrm{e}}$, will not precisely sample the posterior, but it will provide an approximation of it. ES, ES–MDA, and IES will then lead to three different answers: IES attempts to give the distribution of minima of the cost functions as long as the method converges and there are no local minima (although we will see below that the use of an ensemble representation for the gradient introduces an approximation). ES uses a single linear update step and only finds an estimate of the minima of the cost function, but ES also solves for an approximate variance minimizing solution of the marginal pdf. ES–MDA has a similar interpretation as ES, but we will show how ES–MDA uses a sequence of linear steps, that leads to a more accurate solution than ES.

To minimize $\mathcal{J}(x_j)$, we need the gradient of $\mathcal{J}(x_j)$, i.e.,

$$\begin{aligned} \frac{\partial \mathcal{J}(x_j)}{\partial x_j} = &\,2C_{xx}^{-1}\big(x_j - x_j^{\mathrm{f}}\big) \\ &+2g'(x_j)C_{dd}^{-1}\big(g(x_j)-d_j\big). \end{aligned} \tag{19}$$

In iterative schemes, we also need the second derivative, or Hessian, of $\mathcal{J}(x_j)$ which becomes

$$\begin{aligned} \frac{\partial^2 \mathcal{J}(x_j)}{\partial {x_j}^2} = &\,2C_{xx}^{-1} + 2g'(x_j)C_{dd}^{-1}g'(x_j) \\ &+2g''(x_j)C_{dd}^{-1}\big(g(x_j)-d_j\big). \end{aligned} \tag{20}$$

3.1 ES

We can easily derive ES from the cost function (18). By setting the gradient (19) equal to zero, we obtain an equation for each updated (or analyzed) ensemble member x_j^{a} as

$$C_{xx}^{-1}\Big(x_j^{\mathrm{a}} - x_j^{\mathrm{f}}\Big) + g'\big(x_j^{\mathrm{a}}\big)C_{dd}^{-1}\Big(g\big(x_j^{\mathrm{a}}\big) - d_j\Big) = 0. \tag{21}$$

We start by defining the linearizations around x_j^{f}

$$g\big(x_j^{\mathrm{a}}\big) \approx g\big(x_j^{\mathrm{f}}\big) + g'\big(x_j^{\mathrm{f}}\big)\big(x_j^{\mathrm{a}} - x_j^{\mathrm{f}}\big), \tag{22}$$

$$g'\big(x_j^{\mathrm{a}}\big) \approx g'\big(x_j^{\mathrm{f}}\big) + g''\big(x_j^{\mathrm{f}}\big)\big(x_j^{\mathrm{a}} - x_j^{\mathrm{f}}\big), \tag{23}$$

where we will neglect the second derivative $g''(x_j^{\mathrm{f}})$ restricting our self to modest nonlinearity. We use these linearizations in the gradient (21) and multiply with $C_{xx}C_{dd}$ to get

$$\begin{aligned} &C_{dd}\Big(x_j^{\mathrm{a}} - x_j^{\mathrm{f}}\Big) \\ &+C_{xx}\Big(g\big(x_j^{\mathrm{f}}\big) + g'\big(x_j^{\mathrm{f}}\big)\big(x_j^{\mathrm{a}} - x_j^{\mathrm{f}}\big) - d_j\Big)g'\big(x_j^{\mathrm{f}}\big) = 0. \end{aligned} \tag{24}$$

Rearranging gives

$$\begin{aligned} &\Big(g'\big(x_j^{\mathrm{f}}\big)C_{xx}g'\big(x_j^{\mathrm{f}}\big) + C_{dd}\Big)\Big(x_j^{\mathrm{a}} - x_j^{\mathrm{f}}\Big) \\ &= g'\big(x_j^{\mathrm{f}}\big)C_{xx}\Big(d_j - g\big(x_j^{\mathrm{f}}\big)\Big), \end{aligned} \tag{25}$$

and we can now solve for x_j^{a} to get

ES with analytic gradient

$$\begin{aligned} x_j^{\mathrm{a}} &= x_j^{\mathrm{f}} + g'\big(x_j^{\mathrm{f}}\big)C_{xx} \\ &\qquad\times\Big(g'\big(x_j^{\mathrm{f}}\big)C_{xx}g'\big(x_j^{\mathrm{f}}\big) + C_{dd}\Big)^{-1}\Big(d_j - g\big(x_j^{\mathrm{f}}\big)\Big), \\ y_j^{\mathrm{a}} &= g\big(x_j^{\mathrm{a}}\big). \end{aligned} \tag{26}$$

The covariances C_{xx}, C_{yy}, and C_{yx}, are defined as the covariances around an ensemble means $x^{\mathrm{f}} = \overline{x_j^{\mathrm{f}}}$ and $y^{\mathrm{f}} = \overline{y_j^{\mathrm{f}}}$, with the overline denoting ensemble average, and we can write

$$C_{xx}^{\mathrm{e}} = \overline{\Big(x_j^{\mathrm{f}} - x^{\mathrm{f}}\Big)^2}. \tag{27}$$

We will also need to use an expansion of $g(x)$ around the ensemble mean

$$g\big(x_j^{\mathrm{f}}\big) \approx g\big(x^{\mathrm{f}}\big) + g'\big(x^{\mathrm{f}}\big)\big(x_j^{\mathrm{f}} - x^{\mathrm{f}}\big). \tag{28}$$

We can then write the following

$$\begin{aligned} C^{\mathrm e}_{xy} &= \overline{\left(x^{\mathrm f}_j - x^{\mathrm f}\right)\left(y^{\mathrm f}_j - y^{\mathrm f}\right)} \\ &= \overline{\left(x^{\mathrm f}_j - x^{\mathrm f}\right)\left(g\left(x^{\mathrm f}_j\right) - \overline{g\left(x^{\mathrm f}_j\right)}\right)} \\ &\approx \overline{\left(x^{\mathrm f}_j - x^{\mathrm f}\right)\left(g\left(x^{\mathrm f}\right) + g'\left(x^{\mathrm f}\right)\left(x^{\mathrm f}_j - x^{\mathrm f}\right)\right)} \\ &\quad \overline{-\left(x^{\mathrm f}_j - x^{\mathrm f}\right)\left(\overline{g\left(x^{\mathrm f}\right) + g'\left(x^{\mathrm f}\right)\left(x^{\mathrm f}_j - x^{\mathrm f}\right)}\right)} \\ &= g'\left(x^{\mathrm f}\right)\overline{\left(x^{\mathrm f}_j - x^{\mathrm f}\right)^2} \\ &= g'\left(x^{\mathrm f}\right)C^{\mathrm e}_{xx}, \end{aligned} \tag{29}$$

and

$$\begin{aligned} C^{\mathrm e}_{yy} &= \overline{\left(y^{\mathrm f}_j - y^{\mathrm f}\right)^2} \\ &= \overline{\left(g\left(x^{\mathrm f}_j\right) - \overline{g\left(x^{\mathrm f}_j\right)}\right)^2} \\ &\approx \overline{\Big(g\left(x^{\mathrm f}\right) + g'\left(x^{\mathrm f}\right)\left(x^{\mathrm f}_j - x^{\mathrm f}\right)} \\ &\qquad \overline{-\overline{g\left(x^{\mathrm f}\right) + g'\left(x^{\mathrm f}\right)\left(x^{\mathrm f}_j - x^{\mathrm f}\right)}\Big)^2} \\ &= g'\left(x^{\mathrm f}\right)\overline{\left(x^{\mathrm f}_j - x^{\mathrm f}\right)^2}\,g'\left(x^{\mathrm f}\right) \\ &= g'\left(x^{\mathrm f}\right)C^{\mathrm e}_{xx}g'\left(x^{\mathrm f}\right). \end{aligned} \tag{30}$$

When using these expressions, the update equation (26) becomes

ES with ensemble gradient

$$\begin{aligned} x^{\mathrm a}_j &= x^{\mathrm f}_j + C^{\mathrm e}_{xy}\left(C^{\mathrm e}_{yy} + C^{\mathrm e}_{dd}\right)^{-1}\left(d_j - g\left(x^{\mathrm f}_j\right)\right), \\ y^{\mathrm a}_j &= g\left(x^{\mathrm a}_j\right). \end{aligned} \tag{31}$$

Note that $d_j = d + \epsilon_j$, where we sample ϵ_j from the Gaussian distribution $\mathcal{N}(0, C_{dd})$, and we can use ϵ_j to compute and represent $C^{\mathrm e}_{dd}$.

3.2 IES

We can write a simple Gauss-Newton iteration as

$$x_{i+1} = x_i - \gamma\frac{\left.\frac{\partial J(x)}{\partial x}\right|_{x=x_i}}{\left.\frac{\partial^2 J(x)}{\partial x^2}\right|_{x=x_i}}. \tag{32}$$

Now we can use the gradient and Hessian from Eqs. 19 and 20 in this iteration. Note that the $g''(x)$ is normally assumed to be zero, since the term is anyway small when the nonlinearity is not too large and it does not impact the value of the gradient. Thus, we solve the quasi-Newton iteration

$$\begin{aligned} x_{j,i+1} &= x_{j,i} \\ &\quad - \gamma\frac{C^{-1}_{xx}\left(x_{j,i} - x^{\mathrm f}_j\right) + g'(x_{j,i})C^{-1}_{dd}\left(g(x_{j,i}) - d_j\right)}{C^{-1}_{xx} + g'(x_{j,i})C^{-1}_{dd}g'(x_{j,i})}. \end{aligned} \tag{33}$$

In the scalar case, we obtain a simler form by multiplying Eq. 33 with $1 = (C_{xx}C_{dd})/(C_{xx}C_{dd})$, and by replacing the covariances with their ensemble representations, i.e.,

IES with analytic gradient

$$\begin{aligned} x_{j,i+1} &= x_{j,i} \\ &\quad -\gamma\frac{C^{\mathrm e}_{dd}\left(x_{j,i} - x^{\mathrm f}_j\right) + g'(x_{j,i})C^{\mathrm e}_{xx}\left(g(x_{j,i}) - d_j\right)}{g'(x_{j,i})C^{\mathrm e}_{xx}g'(x_{j,i}) + C^{\mathrm e}_{dd}}, \\ y_{j,i+1} &= g(x_{j,i+1}). \end{aligned} \tag{34}$$

The IES minimization problem as defined in Eq. 34 correctly minimizes the cost functions (18) as long as the iterations converge to the global minimum for each realization.

In Eq. 34 we still need to compute $g'(x_{j,i})$ evaluated at the current iterate i. However, we can rewrite as follows

$$\begin{aligned} g'(x_{j,i})C^{\mathrm e}_{xx} &= g'(x_{j,i})C^{\mathrm{e},i}_{xx}\left(C^{\mathrm{e},i}_{xx}\right)^{-1}C^{\mathrm e}_{xx} \\ &\approx g'(\overline{x_i})C^{\mathrm{e},i}_{xx}\left(C^{\mathrm{e},i}_{xx}\right)^{-1}C^{\mathrm e}_{xx} \\ &\approx C^{\mathrm{e},i}_{xy}\left(C^{\mathrm{e},i}_{xx}\right)^{-1}C^{\mathrm e}_{xx}, \end{aligned} \tag{35}$$

where we assume that the inverse of the covariance $C^{\mathrm{e},i}_{xx}$ exists, and we as in ES evaluate g' at the ensemble mean. Then using the definition of the covariance (29) we can rewrite the "analytic" IES equation utilizing an ensemble approximation of the gradient.

IES with ensemble gradient

$$\begin{aligned} x_{j,i+1} =& x_{j,i} - \gamma\frac{C^{\mathrm e}_{dd}\left(x_{j,i} - x^{\mathrm f}_j\right)}{C^{\mathrm{e},i}_{yy} + C^{\mathrm e}_{dd}} \\ &- \gamma\frac{C^{\mathrm{e},i}_{xy}\left(C^{\mathrm{e},i}_{xx}\right)^{-1}C^{\mathrm e}_{xx}\left(g(x_{j,i}) - d_j\right)}{C^{\mathrm{e},i}_{yy} + C^{\mathrm e}_{dd}} \\ y_{j,i+1} =& g(x_{j,i+1}) \end{aligned} \tag{36}$$

Note that we change the expression for the gradient by introducing the ensemble representation given by Eq. 35, and thereby also alter the minimizing solutions defined by the gradient being equal to zero for each realization.

Here we also used the local iterate of the ensemble $C^{\mathrm{e},i}_{yy}$ in the denominator, but we could equally well use $C^{\mathrm e}_{yy}$ from the prior ensemble, since these choices do not change the gradient which defines the final solution, they only impact the step length used in the quasi-Newton iteration. Note also that the form of the equation requires the inversion of the covariance $C^{\mathrm{e},i}_{xx}$, which has a dimension equal to the number of parameters. However, in practical cases, the ensemble size is much smaller than the number of parameters and a pseudo-inversion can be computed from a singular-value decomposition of the ensemble. We must evaluate the gradient (and Hessian) in each iteration step. Thus, we must also integrate the ensemble in each iteration, and the total cost becomes equal to $N_\text{ies}+1$ ensemble integrations with N_ies being the required number of iterations to reach convergence.

3.3 ES–MDA

We can formally derive ES–MDA from the Bayesian formulation using a tempering procedure by [20, 25]. We rewrite the likelihood function for the measurements as

$$f(d|y) = f(d|y)^{\left(\sum_{i=1}^{N_\text{mda}} \frac{1}{\alpha_i}\right)} = \prod_{i=1}^{N_\text{mda}} f(d|y)^{\frac{1}{\alpha_i}}, \tag{37}$$

where

$$\sum_{i=1}^{N_\text{mda}} \frac{1}{\alpha_i} = 1. \tag{38}$$

For a Gaussian likelihood, we then get

$$\begin{aligned} f(d|y) &\propto \exp\left(-\tfrac{1}{2}(y-d)C_{dd}^{-1}(y-d)\right) \\ &= \textstyle\prod_{i=1}^{N_\text{mda}} \exp\left(-\tfrac{1}{2\alpha_i}(y-d)C_{dd}^{-1}(y-d)\right). \end{aligned} \tag{39}$$

Using Eq. 37 in Bayes' theorem (16), we obtain

$$f(x|d) \propto f(x) \prod_{i=1}^{N_\text{mda}} f\big(d|g(x_{i-1})\big)^{\frac{1}{\alpha_i}}. \tag{40}$$

This expression can be rewritten as a recursion starting with the prior $x=x_0$ leading to the posterior $x=x_{N_\text{mda}}$.

$$\begin{aligned} f(x_1|d) &\propto f(x_0)f(d|g(x_0))^{\frac{1}{\alpha_1}}, \\ f(x_2|d) &\propto f(x_1|d)f(d|g(x_1))^{\frac{1}{\alpha_2}}, \\ &\vdots \\ f(x_{N_\text{mda}}|d) &\propto f(x_{N_\text{mda}-1}|d)f(d|g(x_{N_\text{mda}-1}))^{\frac{1}{\alpha_{N_\text{mda}}}}. \end{aligned} \tag{41}$$

Maximizing each of the recursions corresponds to minimizing a cost function for each recursive step. Thus, ES–MDA solves a predefined sequence of N_mda minimization problems similar to the cost functions (18), written as

$$\begin{aligned} J(x_{j,i+1}) = {} & (x_{j,i+1}-x_{j,i})\left(C_{xx}^{\text{e},i}\right)^{-1}(x_{j,i+1}-x_{j,i}) \\ & + \left(g(x_{j,i+1})-d-\sqrt{\alpha_i}\epsilon\right) \\ & \times\left(\alpha_i C_{dd}^\text{e}\right)^{-1}\left(g(x_{j,i+1})-d-\sqrt{\alpha_i}\epsilon\right), \end{aligned} \tag{42}$$

where the initial $x_{j,i=1} = x_j^\text{f}$ and $(C_{xx}^\text{e})_{i=1} = C_{xx}^\text{e}$. In each step, we inflate the measurement errors by a factor $\sqrt{\alpha_i}$, which satisfies Eq. 38. There is no approximation introduced in this recursion, and this choice of α_i ensures that the N_mda recursive steps become precisely the ES solution in the linear case.

The sequence of cost functions (42) is in each step solved using the standard ES equations, which, with the inflated measurement errors, becomes when using the formulation with the "analytic gradient,"

ES–MDA with analytic gradient

$$\begin{aligned} x_{j,i+1} &= x_{j,i} \\ &\quad + g'(x_{j,i})C_{xx}^{\text{e},i}\left(g'(x_{j,i})C_{xx}^{\text{e},i}g'(x_{j,i})+\alpha_i C_{dd}^\text{e}\right)^{-1} \\ &\quad \times\left(d+\sqrt{\alpha_i}\epsilon_j - g(x_{j,i})\right), \\ y_{j,i+1} &= g(x_{j,i+1}), \end{aligned} \tag{43}$$

and with the ensemble gradient, we obtain

ES–MDA with ensemble gradient

$$\begin{aligned} x_{j,i+1} &= x_{j,i} + C_{xy}^{\text{e},i}\left(C_{yy}^{\text{e},i}+\alpha_i C_{dd}^\text{e}\right)^{-1} \\ &\qquad \times\left(d+\sqrt{\alpha_i}\epsilon_j - g(x_{j,i})\right), \\ y_{j,i+1} &= g(x_{j,i+1}). \end{aligned} \tag{44}$$

The final result after N_mda steps is $x_j^\text{a} = x_{j,N_\text{mda}}$. Note that the error covariances $C_{xy}^{\text{e},i}$ and $C_{yy}^{\text{e},i}$ are computed from the ensemble at step i, and are thus being updated recursively during the sequence of update steps. The benefit of this stepwise approach is that it uses many short linear steps with local linearization around $\overline{x_i}$ rather than one long ES step with linearization around x^f. The expectation is that this stepwise approach will lead to a better result than what is found using ES.

3.4 Remarks about ES, ES–MDA, and IES

It is clear that the ES, ES–MDA, and IES algorithms are similar in many aspects. ES is equivalent to ES–MDA with one step, and in the linear case the methods only differ in the choice of step lengths and the number of steps, and they converge to the same solution.

In the nonlinear case, there is, in addition to the use of different step lengths, also a difference related to a linearization of the nonlinear model and the evaluation of ensemble gradients.

- In the IES–analytic, there are no approximations introduced during the derivation starting from the cost function (18). Thus, by evaluating the gradient analytically, the iteration converges exactly to the minimum of the cost function for each realization. Unfortunately, as will be illustrated below the posterior IES ensemble is not sampling the posterior pdf defined by Bayes' in Eq. 16.
- In IES–ensemble, we replace the analytic expression for the gradient with an ensemble approximation. By using

an approximation for the gradient, we also change the minimizing solution for each realization. Thus, we will now sample another posterior pdf than the one obtained using IES–analytic.

- In the ES update scheme, we linearize the model around the first guess x^{f} and one single update step is computed using an approximate ensemble gradient. Thus, with a large update, it is likely that the solution will suffer from both an approximate direction and magnitude of the update. On the other hand, only two ensemble integrations are needed, one to generate the prior ensemble prediction, and one to compute the posterior ensemble.
- It is worth noting that the sequential EnKF computes many small recursive updates in time and the solution stays close to the measured state at each update step. Each local linearization is then likely to have less impact in EnKF than ES, and this property may explain the previous success when using non-iterative EnKF for reservoir history matching.
- In ES–MDA, we use the ES update equation with inflated measurement errors defined by the choices of α_i. Thus, we apply a local linearization of the update equations at each step (which is why we need to rerun the ensemble prediction at each step). The consistency and convergence to ES are proven for the linear case by Emerick and Reynolds [8]. However, for the nonlinear case, there is no proof of the convergence of ES–MDA. From the examples below, we will see that ES–MDA reduces the error in the final update compared to ES.
- The similarity of ES, ES–MDA, and IES, can be further illustrated by considering the IES iteration in Eq. 36, which for the first step with $\gamma = 1$ and $x_{j,1} = x_j^{\mathrm{f}}$, becomes identical to the ES update equation (31).
- Although none of the methods considered in this paper correctly samples the true posterior pdf from Bayes', the posterior ensembles can be used as proposal densities in a particle filter algorithm, and by assigning proper weights to the realizations, it is possible to sample the true posterior pdf.

4 Scalar example

We will use a simple scalar model to illustrate in some more detail the properties of the ES, ES–MDA, and IES methods. The example resembles the use of conditioning methods in history matching, i.e., there is a parameter x that serves as an input to a forward model to predict $y = g(x)$. We then observe y and try to estimate x, and then predict an updated y.

Ensemble methods are known to perform very well for weakly nonlinear dynamical models. However, it is more precise to say that the methods perform well with weakly nonlinear and monotonic models. By monotonic we mean that the derivative of the model with respect to the input parameter does not change its sign. Thus, for a model with a positive derivative, an increase in x will always lead to an increase in y. A monotonic model cannot support the multimodal behavior that is often associated with strongly nonlinear dynamics. Also, as was illustrated by Evensen [11, Chap. 10, Fig. 7], ES cannot consistently handle multimodal behavior. Reservoir models often exhibit a monotonic response, e.g., an increase in permeability leads to an associated increase in production and this property is mainly responsible for the success of ensemble methods in history matching. Thus, the discussion below considers a monotonic and weakly nonlinear scalar model, while we will study the highly nonlinear case in Section 5.

4.1 Scalar model

We assume an initial state x and a prediction y given by the model

$$\begin{aligned} y &= g(x) + q \\ &= x(1+\beta x^2) + q. \end{aligned} \tag{45}$$

Here, β is a parameter that determines the nonlinearity of the model. In the current example, we have used $\beta = 0.0$ for the linear case and $\beta = 0.2$ for the nonlinear case.

The model error variance is C_{qq} and q is a random variable sampled from $\mathcal{N}(0, C_{qq})$. We have assumed the model error to be zero although it would still be interesting to examine the impact of model errors on the inverse problem.

We sample the prior ensemble for x from a Gaussian distribution $\mathcal{N}(x^{\mathrm{f}} = 1, C_{xx} = 1)$ and the observation of y has the distribution $\mathcal{N}(d = -1, C_{dd} = 1)$. Thus, in the current example, x represents the initial state or the model parameter, while y is the prediction which is observed. The goal is to estimate x given an observation of y.

In this example, we use a sufficiently large number of samples, i.e., 10^7, to generate accurate estimates of the probability density functions, and this allows us to work directly with the pdfs and to examine the sampling properties of the methods.

The pdf for the model is given by the transition density

$$f(y|x) \propto \exp\left(-\frac{(y - x(1+\beta x^2))^2}{2C_{qq}}\right), \tag{46}$$

which in the limit of zero model errors becomes

$$f(y|x) \propto \delta\left(y - x(1+\beta x^2)\right). \tag{47}$$

4.2 Base-case experiments

In Fig. 1 we show the joint pdf $f(x, y)$ and joint conditional pdf $f(x, y|d)$ for the linear case with $\beta = 0$ and nonlinear case with $\beta = 0.2$ in the upper and middle plots respectively. Since we have set the model errors to zero, the joint pdfs have zero probability outside the curve defined by the model. In the linear case, the joint pdf has its probability mapped onto a line, with the highest probability around $x = 1$ as defined by the initial pdf for x. The conditioning on the datum $d = -1$ shifts the high probability towards $y = -1$ and on the same line. The same happens in the nonlinear case, but now the nonlinearity of the model is obvious in the mapping. For illustration, we also show the joint pdfs for the case when we include model errors in the bottom plots of Fig. 1. Then the joint pdf will be smooth in the y-direction taking into account that one value of x may be mapped to different values of y as defined by the stochastic forcing from the model errors.

4.2.1 ES–MDA scheme for α_i

We have defined a scheme for α_i, where we start by selecting any nonzero value for α'_1. Then α'_i is computed as

$$\alpha'_{i+1} = \alpha'_i / \alpha_{\text{geo}}, \tag{48}$$

where α_{geo} is a constant defining the change of step lengths from one step to the next. The final values for α_i are obtained by scaling the values from Eq. (48) as

$$\alpha_i = \alpha'_i \left(\sum_{i=1}^{N_{\text{mda}}} \frac{1}{\alpha'_i} \right). \tag{49}$$

With this scheme, $\alpha_{\text{geo}} = 1.0$ results in uniform step lengths. A positive $\alpha_{\text{geo}} < 1.0$ leads to increasing values for α_i and decreasing step lengths, and $\alpha_{\text{geo}} > 1.0$ leads to decreasing values for α_i and increasing step lengths. The scaling used in Eq. 49 ensures that the constraint on the sum of $1/\alpha_i$ in Eq. 38 is satisfied.

4.2.2 Definition of line legends

The line legends refer to different cases and are defined as follows: for ES and IES experiments we use typically IES_L_7_ENS where L denote linear case, 7 defines the ensemble size as 10^7, and ANA and ENS defines respectively an exact analytic gradient and an approximate ensemble gradient. For ES–MDA experiments we add the number of MDA steps e.g., 008 and the value used for α_{geo}, and an example is MDA_L_7_ENS_008_1.0 were we for the nonlinear cases just drop the L.

4.2.3 Linear-model results

In Fig. 2, we show the marginal pdfs from the linear case, where we compare ES, ES–MDA, and IES, using the ensemble gradients, with the exact Bayesian solution. We have used 8 ES–MDA steps with equal weights $\alpha = 8$, and the IES iterations used a step length of $\gamma = 0.5$.

The plots should be read as follows:

1. Start in the left plots with the red initial Gaussian pdf of x. We represent this prior pdf by a large ensemble of samples. Each realization of the prior ensemble is used as input to the model (45) and an ensemble of predictions, representing the distribution of y, is obtained and plotted as the red Gaussian pdf in the right plots.
2. We now compute the ES update of x to obtain the cyan ES posterior pdf for x in the left plots. Using the updated samples of the ES posterior pdf for x, we can compute the model prediction of the posterior for y, which we show in the right plots.
3. We then repeat this process for ES–MDA and IES by stepwise incremental updates of x followed by updates of y using the model.

The results from this experiment can be summarized as follows: In the linear case both ES–MDA and IES converges precisely to the ES solution, which also equals the true solution defined by the Bayesian update (black pdf), illustrating the consistency of the methods in the linear case. We will next discuss the methods in more detail for the nonlinear case.

4.3 ES experiments

In Fig. 3, we have plotted the ES solutions from the ES with an analytic gradient as defined by Eq. 26 and ES with an ensemble gradient as defined by Eq. 31. From these plots, it is clear that the use of the gradient to determine the update will only lead to an approximate solution in the nonlinear case. The introduction of an ensemble gradient introduces an additional approximation as seen from the differences in the two pdfs. Thus, ES is likely to give better results in cases with nearly linear models or when the updates are small.

4.4 IES experiments

In Fig. 4, we have plotted the IES solutions from the IES with an analytic gradient as defined by Eq. 34, and IES with an ensemble gradient as defined by Eq. 36. We exit the iterations for a realization as soon as the ratio between the gradient and Hessian is less than 0.0001. It is clear that there is a significant difference between the Bayesian

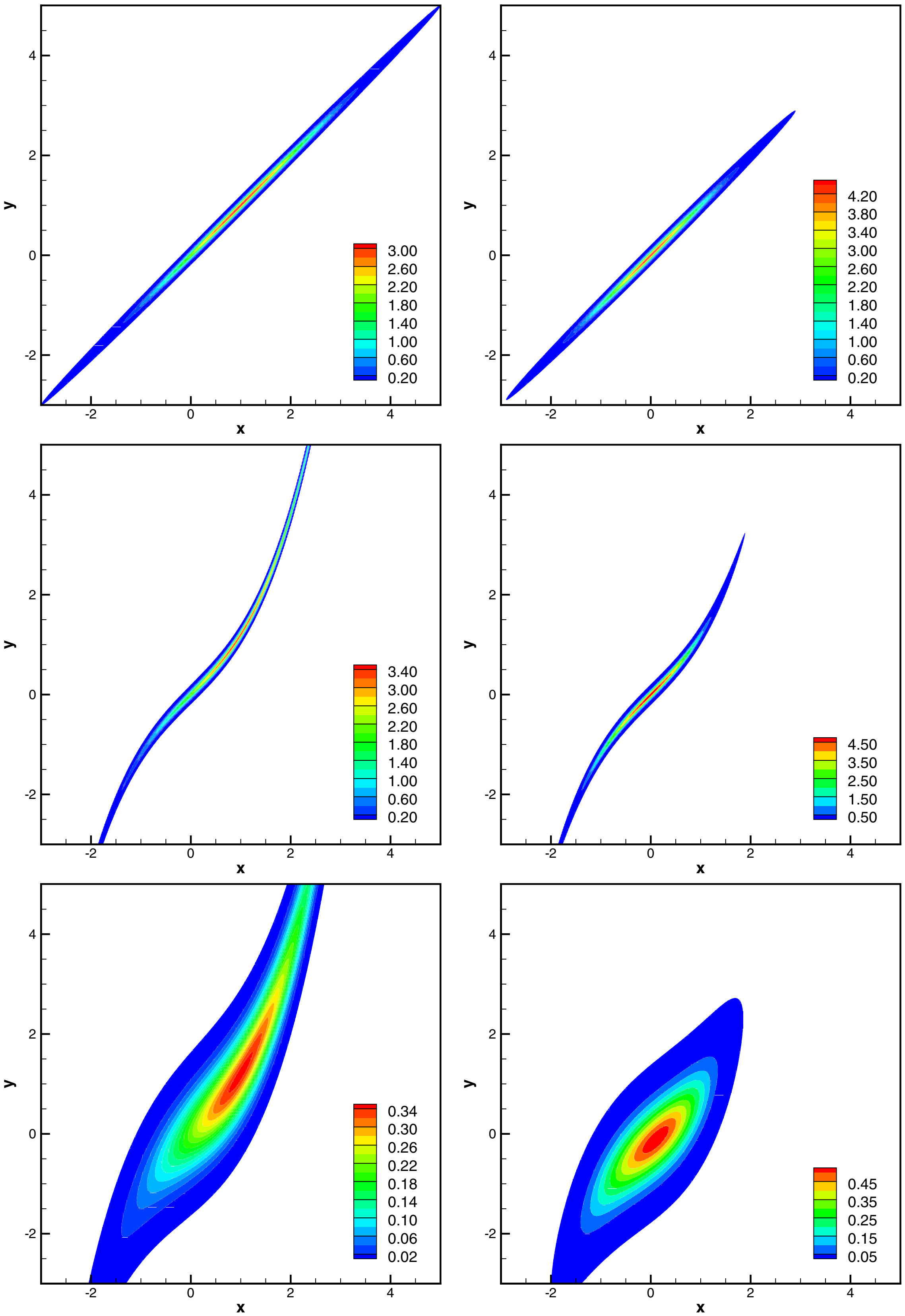

Fig. 1 The joint pdfs $f(x, y)$ (left) and joint conditional pdfs $f(x, y|d)$ from Bayes' (right), for the linear case (top), nonlinear case (middle), nonlinear case with finite model errors drawn from $\mathcal{N}(0, C_{qq} = 0.25)$ (bottom)

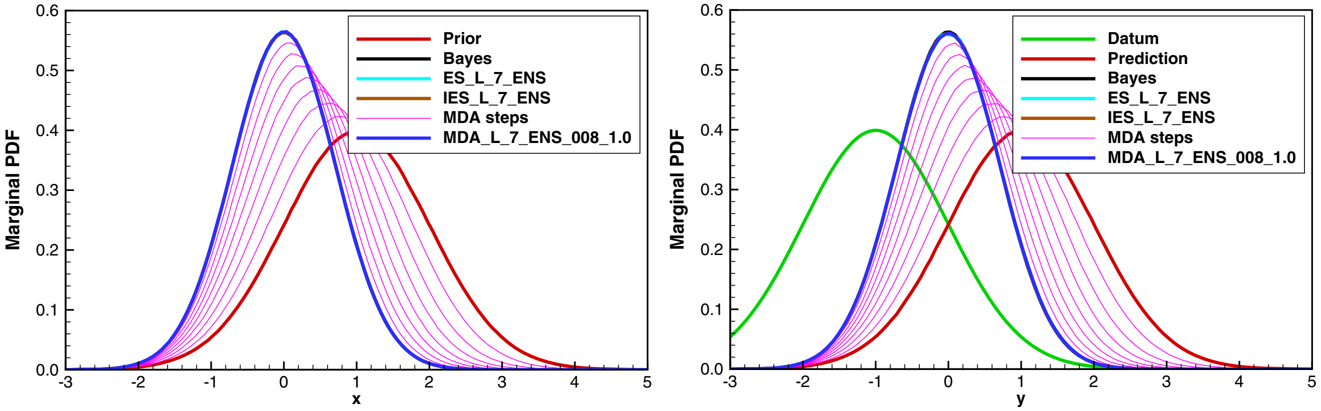

Fig. 2 We illustrate the consistency between ES, ES–MDA, and IES in the linear case. The red pdfs are the prior distribution for x (left plot) and the associated prediction for y (right plot). The green pdf gives the likelihood for the measurements of y. The cyan pdf is the ES estimate. The blue pdf is the posterior estimate from ES–MDA with eight uniform steps. The brown pdf is the IES result. We have only plotted results using the smoothers with the ensemble gradients. The magenta pdfs are the intermediate seven steps in ES–MDA (update of x and corresponding predictions of $y = g(x)$)

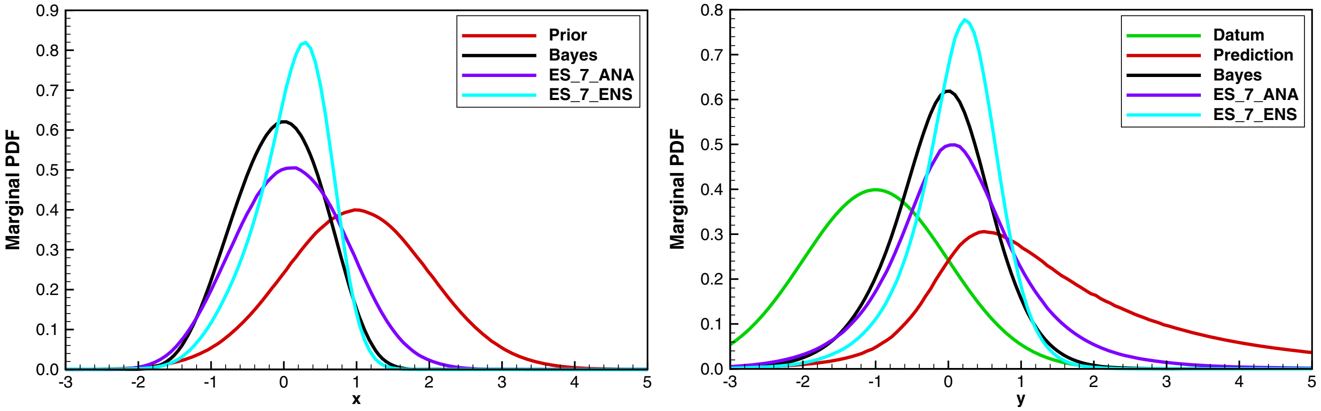

Fig. 3 The figure shows the marginal pdfs for x (left) and y (right) when using ES with the analytic gradient from Eq. 26 and the ensemble gradient from Eq. 31

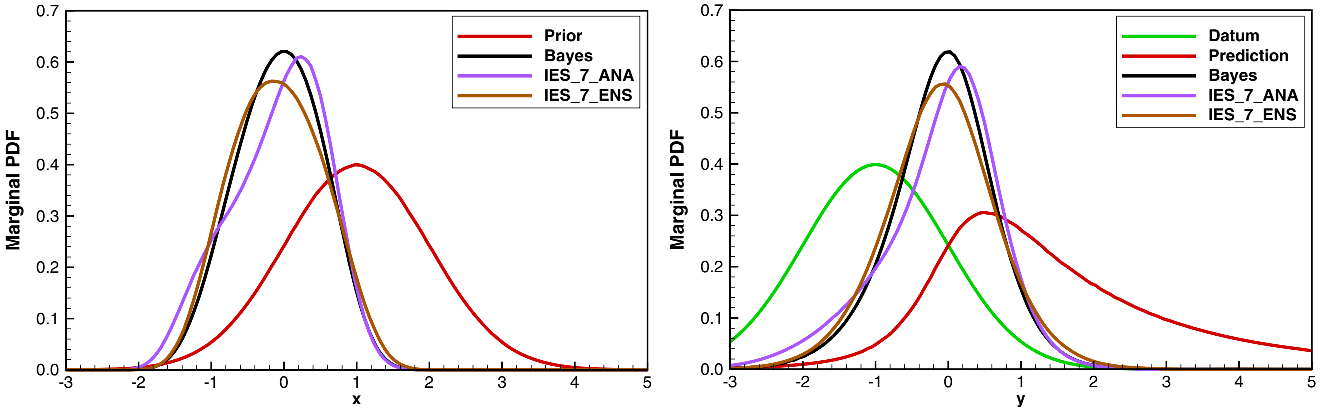

Fig. 4 The figure shows the marginal pdfs for x (left) and y (right) when using IES with the analytic and ensemble gradients from Eqs. 34 and 36

posterior and the approximate solutions found by IES with both analytic and ensemble gradients. It is unfortunate that the cost functions we minimize do not sample the Bayesian posterior in the nonlinear case, which means we are solving the wrong problem. Still, we observe that the IES estimate is a rather good approximation to the Bayesian posterior, and significantly better than ES.

In Fig. 5 we plot the first four moments of the distributions for x and y, i.e., the mean, variance, skewness, and kurtosis, as computed from the ensemble. Note that we plot a kurtosis where we have subtracted the value 3, to center it around zero. We have also included the results of an experiment with 100 realizations, to examine the impact of using a small ensemble size since this leads to a further approximation of the gradient. It is clear that for the mean the results from using an ensemble of 100 realizations are converging to a slightly larger value for x while the use of an analytic versus ensemble gradient does not make a big difference.

For the skewness and kurtosis, we observe that the initial prediction of y indicates strong deviations from Gaussianity, while during the iterations the skewness and kurtosis for y is significantly reduced and we converge towards a more Gaussian distribution. On the other hand, we also observe that the skewness and kurtosis for x, which were initially equal to zero, slowly deviate from the initial value, indicating that the distribution for the estimate of x is slightly non-Gaussian. This non-Gaussianity is clearly seen from the plots in Fig. 4. IES seems to work well for estimation of the posterior mean and variance even using only 100 realizations, while the plotting of higher order moments makes less sense with such a small ensemble.

Compared with ES–MDA, IES is relatively easy to analyze as long as the method converges to the global minimum of the cost function (18). The focus has therefore been more on the parameters of the iteration scheme to ensure fast convergence, rather than trying to understand precisely, which distribution IES is sampling.

4.5 ES–MDA experiments

ES–MDA has some parameters that actually will change the final solution. The solution will depend on the number of MDA steps used, and the sequence of values used for α_i. Thus, it is not apparent what ES–MDA should converge to, or how we should determine a converged result.

4.5.1 ES–MDA convergence with number of step lengths

To start, we will examine the convergence of ES–MDA with the number of MDA steps. We have run ES–MDA with 1, 2, 4, 8, 16, 32, 64 and 128 steps using a constant uniform value of α_i that equals the total number of steps in each case ($\alpha_{\text{geo}} = 1.0$). In Fig. 6, we plot the estimates of the pdfs for x and y. We see how ES–MDA with only one step (i.e., ES) is rather far from the correct Bayesian posterior. Then, using ES–MDA with 2, 4, 8, and 16 MDA steps gives a significant stepwise improvement, while when using 32, 64, and 128 MDA steps we needed to zoom the plots to see any difference, so we did not plot these results. It is also amazing how close the converged ES–MDA solution is to the Bayesian posterior in this case. However, we must run additional experiments with different nonlinear models, before we conclude anything about the general quality of the converged ES–MDA result.

In Fig. 7, we plot the statistical moments for the ES–MDA steps as a function of the sum $\sum_i 1/\alpha_i$. Thus, we can analyze how the step lengths and number of steps influence the convergence of the statistical moments. We see that, as soon as we use a certain number of steps (above 16 here), it is difficult to distinguish the results. Thus, we conclude that 16 or more steps, in this case, may be needed for ES–MDA to converge with infinite sample size. On the other hand, even ES–MDA with only two steps provides an improvement compared to ES.

Like for the IES, we also ran a case using only 100 realizations with ES–MDA, and we plot the results for mean and variance in Fig. 8. We see that for more than 16 iterations, the estimated means will contain sampling errors that are larger than the error reduction due to an increase in the number of steps. Thus, with a small ensemble size, there is no benefit of running very many steps, which is an essential result concerning practical use of ES–MDA.

4.5.2 ES–MDA convergence with non-uniform step lengths

Some publications (e.g., [6, 14, 15, 17, 18, 22]) have suggested the use of small initial steps to regularize the problem, and have referred to the Discrepancy Principle when deriving new optimal schemes for the sequence of α_i.

We initially believed that the use of small initial steps in ES–MDA (corresponding to large values of α) would lead to reduced errors, since the non-Gaussianity of the distribution for y is the largest in the early steps, and then the corresponding approximations in the linear update equations would be the largest. However, from the plots in Fig. 9, we note that the use of a geometrical reduction of α where $\alpha_{i+1} = \alpha_i/2$ does not result in a significant change of the results.

One could also suggest that the improvement with reduced step length is an effect of reducing truncation errors in the ES–MDA scheme, which is based on a linearization (22) and (23) around the local estimate of x, and a linearization (28) around the ensemble mean. We can also interpret the ES–MDA scheme as a time-stepping

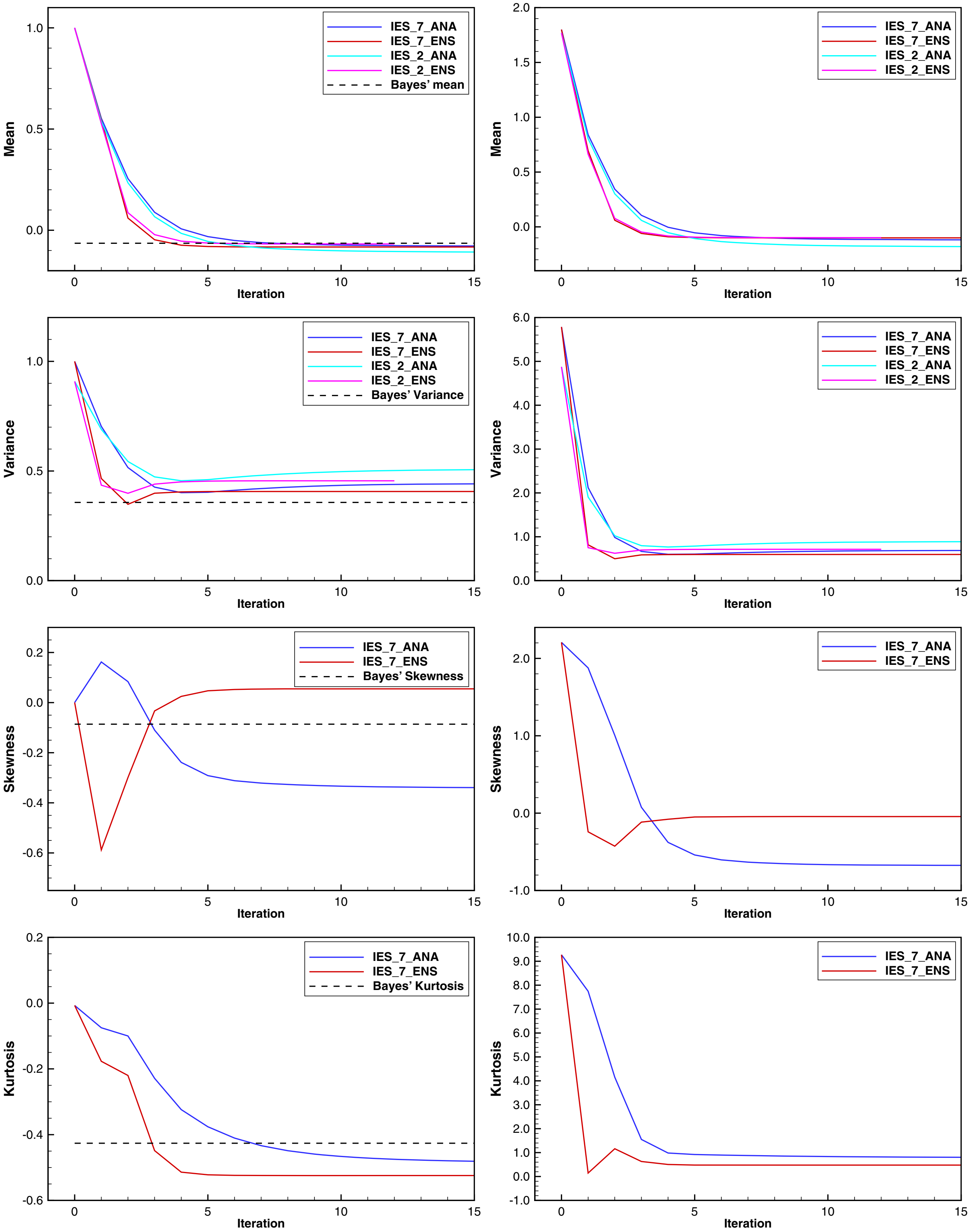

Fig. 5 Here, we illustrate the IES convergence of ensemble mean, variance, skewness, and kurtosis for x (left plots) and y (right plots). For the mean and variance, we also plot a solution using only 100 realizations. The dashed black lines are the theoretical values computed from the Bayesian posterior

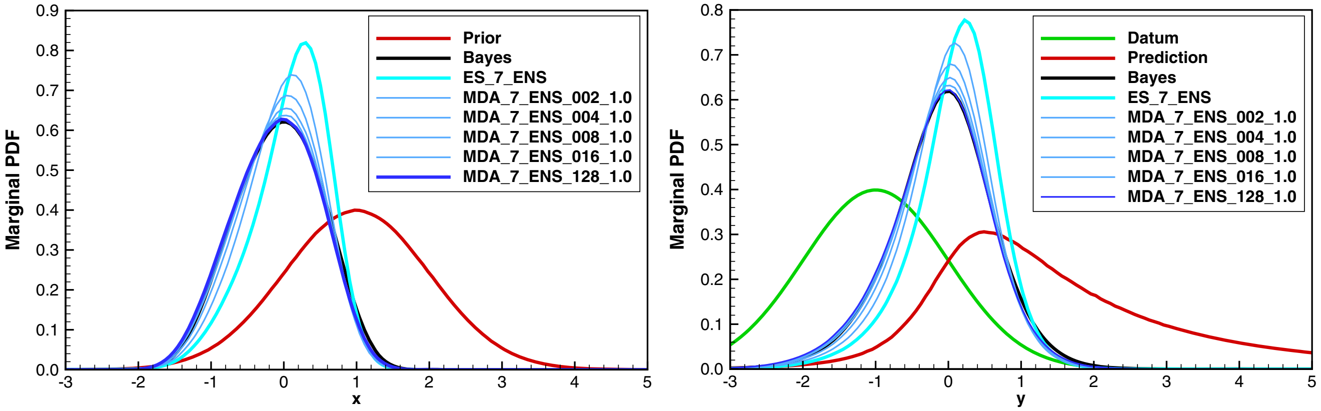

Fig. 6 The figure shows the marginal pdfs for x (left) and y (right) when using ES–MDA with 2, 4, 8, 16, and 128 steps

scheme in the pseudo-time $T_n = \sum_{i=1}^{n} 1/\alpha_i$. Then it is not clear how to implement a changing step length since it will depend on the truncated terms in the linearizations as well as the approximation resulting from introducing the ensemble covariance between x and y.

The use of a scalar example allowed for running very many experiments using a different number of update steps in ES–MDA and different geometrical factors α_{geo}. In Fig. 10 we have plotted statistics for cases with two steps in the left column and four steps in the right column, comparing geometrical factors $\alpha_{\text{geo}} = (0.5, 1.0, \ldots, 9.0)$. When evaluating the results, it is natural to compare with the solution that ES–MDA converges to as represented by ES–MDA with 128 uniform steps. We also plot the moments of the true Bayesian solution. We expect that an improvement of using a non-uniform sequence of MDA step lengths is best visualized using a small number of steps. Thus, we should probably not conclude from the case with eight steps shown in Fig. 9.

From the plots in Fig. 10, we observe that the estimate for the mean is improving with increasing value of the geometrical factor α_{geo}, and in particular in the case with two MDA steps the best result is obtained using $\alpha_{\text{geo}} = 9.0$ where the estimate is very close to the reference solution where 128 steps were used. For the case with four MDA steps, we find that any value of $\alpha_{\text{geo}} \geq 2.0$ gives equally good results. For the variance, we observe that the best result is obtained using a value of α_{geo} equal 5.0 in the case with two MDA steps and a value of 2.0 when we use four MDA steps. For the skewness the best result is obtained with $\alpha_{\text{geo}} = 4.0$ and $\alpha_{\text{geo}} = 2.0$ respectively for the two cases, while for the kurtosis the best result is obtained for $\alpha_{\text{geo}} = 2.0, 3.0$ in the two-step case and $\alpha_{\text{geo}} = 2.0$ in the four-step case.

In Fig. 11 we plot the statistics for the case with eight update steps. Here it is seen that the optimal step lengths follow a scheme $\alpha_{\text{geo}} = 2.0$. In all the cases, starting with a long step followed by shorter steps always lead to poorer results.

We also see in Fig. 10, that the change in the statistical moments is larger in the first step than the second step. By increasing α_1 we take a shorter first step and the relative changes in steps one and two become approximately the same. Thus, it seems that adjusting the α_i's such that the relative magnitude of the updates in the different MDA steps remains similar, may be beneficial.

To conclude, we are running ES–MDA in a clean setup for a scalar model using a single datum and using a vast ensemble size. The only factor that can have an implication on the choice of step length is then the nonlinearity of the model and possibly the number of update steps used. The effect of the nonlinearity is probably influencing ES–MDA through the linearizations (22) and (23) used in the derivation of the ES–MDA equations and the linearizaion (28) used when introducing the ensemble representation of the covariances. Thus, the impact of these approximations may be reduced by using smaller step lengths in the initial update steps. From the examples, we see an improvement of using a geometrical scheme for α, and the benefit is more significant the fewer MDA steps are used. For the cases with four and eight MDA steps, it seems like a value $\alpha_{\text{geo}} = 2.0$ is close to optimal, but it is likely that this factor is also model dependent.

5 Iterative smoothers with highly nonlinear dynamics

We will now study the iterative smoothers with a highly nonlinear and non-monotonic scalar model. The purpose is to examine if the iterative methods can handle problems with multimodal behavior. We use a simple model that leads

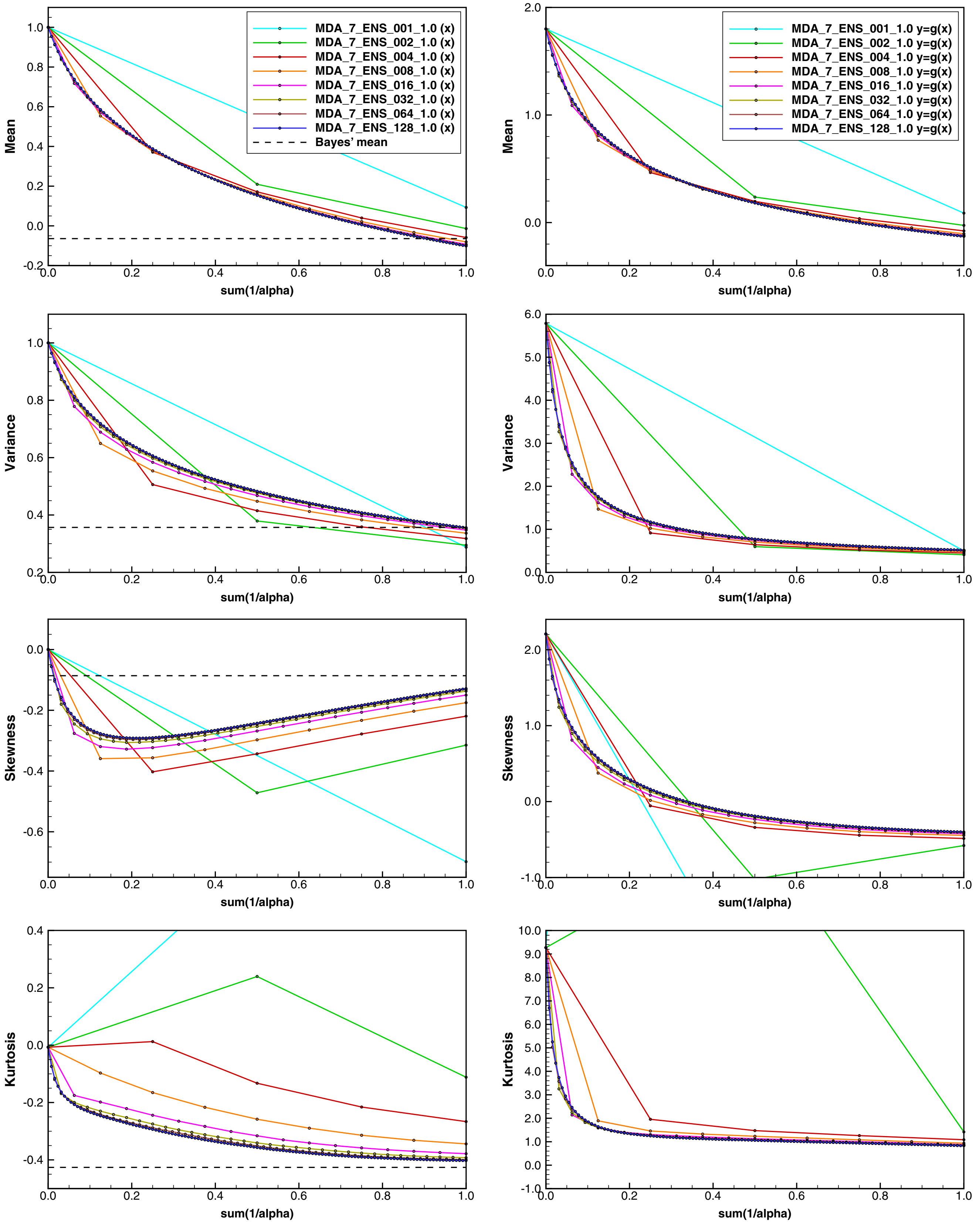

Fig. 7 The plots show the ES–MDA convergence for x (left plots) and y (right plots) of ensemble mean, variance, skewness and kurtosis. We are using 10^7 realizations and 2, 4, 8, 16, 32, 64, and 128 uniform steps. The dashed black lines are the theoretical values computed from the Bayesian posterior. The line legends given in the upper plots also apply for the remainder of the plots in the respective columns

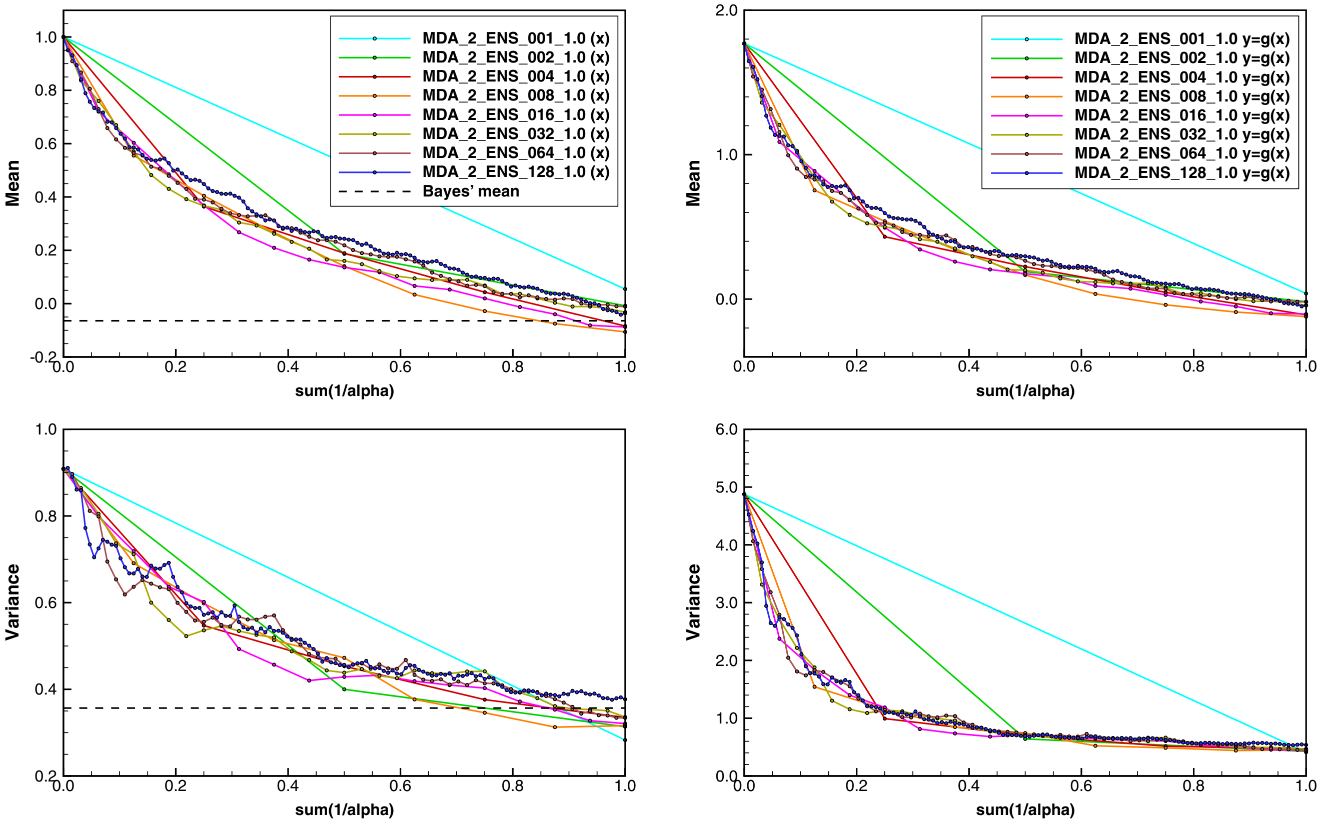

Fig. 8 The plots show the ES–MDA convergence for x (left plots) and y (right plots) of ensemble mean, variance, skewness and kurtosis. We are using 100 realizations and 2, 4, 8, 16, 32, 64, and 128 uniform steps. The dashed black lines are the theoretical values computed from the Bayesian posterior. The line legends given in the upper plots also apply for the lower plots

to inverse problems with arbitrarily many modes dependent on the width of the prior, i.e.,

$$y = 1 + \sin(\pi x) + q. \tag{50}$$

Depending on the prescribed width of the prior for x, the sampled x values can be mapped to arbitrarily many wavelengths of the $\sin(\pi x)$ function. We assume in all cases a measurement of y, $d = 1.0$, with standard deviation equal to 0.1. We then run ES, ES–MDA with 32 steps, and IES, using both the analytic and the ensemble representations of the gradient. We will consider three cases of increasing nonlinearity induced by selecting priors of different widths.

In the plots of the joint pdfs in the Figs. 12, 13 and 14, we have blanked values less than 0.1% of the maximum value in the plots. To better visualize the joint pdfs, we have added a small model error $q \leftarrow \mathcal{N}(0.0, C_{qq} = 0.0009)$ to the final prediction of y.

5.1 Case 1: $x_j^{\mathrm{f}} \leftarrow \mathcal{N}(0.0, C_{xx} = 0.01)$

The first case has a prior for x_j^{f} sampled from $\mathcal{N}(0.0, C_{xx} = 0.01)$. Thus 99.7% of the samples will be located in the interval $[-0.3 : 0.3]$, which is 0.6 of a wavelength of the functional mapping. Thus, the mapping of the prior model is monotonic, and this example becomes similar to the weakly nonlinear case studied in the previous section. In Fig. 12 we plot the prior joint pdf in the upper left plot and the Bayesian posterior pdf in the upper right plot. Additionally, we show the conditional pdfs from ES, ES–MDA, and IES, using ensemble gradients in the left column and the analytic expressions for the gradients in the right column. In this, almost linear case, all methods provide nearly identical results in good agreement with the Bayesian posterior, which we can also see from the marginal pdfs in the upper plot in Fig. 15. ES_ANA and ES_ENS match the Bayesian nearly perfectly. IES_ENS and IES_ANA provide roughly the same solutions, and together with ES–MDA_ENS they give slightly too low variance. Finally, ES–MDA_ENS and ES–MDA_ANA differ with the analytic formulation being a little better than the ensemble formulation.

5.2 Case 2: $x_j^{\mathrm{f}} \leftarrow \mathcal{N}(0.0, C_{xx} = 0.09)$

When we sample x_j^{f} from $\mathcal{N}(0.0, C_{xx} = 0.09)$, the interval $[-0.9 : 0.9]$ will contain 99.7% of the samples and covers almost a full wavelength of the functional mapping. With

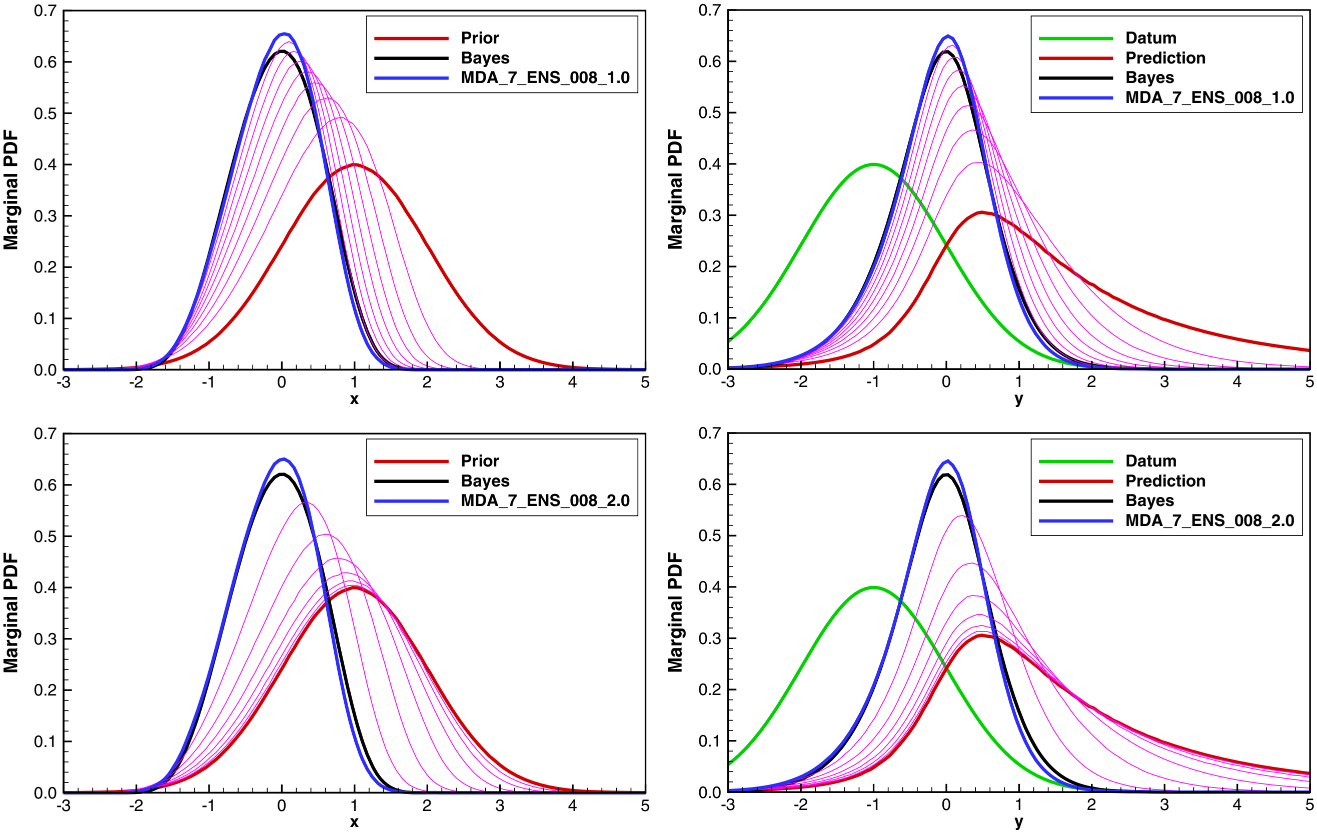

Fig. 9 The figure shows the marginal pdfs for x (left) and y (right) when using ES–MDA–8 with uniform α (upper plots) and geometrical $\alpha_{\text{geo}} = 2$ (lower plots)

the measurement of $y = 1.0$, we can expect some issues related to bimodality as is illustrated by some of the joint conditional pdfs in Fig. 13 and the marginal pdfs in Fig. 15.

We see that ES–MDA_ENS and IES_ENS give solutions very close to the Bayesian posterior. ES_ENS and ES_ANA both result in weak updates with too large variance. Thus, this case is too nonlinear for ES while the iterative smoothers are capable of resolving the nonlinearity.

IES_ANA and ES–MDA_ANA over-estimate a probability of x at $x = \pm 1$ corresponding to the two alternative modes. The ensemble representations of the smoothers give better results than the analytical versions, and it seems that the use of the same ensemble gradient for all realizations leads to a regularization that helps the methods converge correctly to the Bayesian posterior. For ES–MDA, we also made the same observation in the weakly nonlinear case in the previous section.

It is also interesting to see that ES–MDA_ANA does significantly worse than ES in this case. It appears that the inflated measurements with large values of α lead to a diffusion of the updates. Using ES–MDA with 32 steps and a uniform scheme for α, all measurement perturbations will be multiplied by $\sqrt{32}$. A measurement perturbation outside the interval $[-1/\sqrt{32} : 1/\sqrt{32}]$ leads to an inflated measurement located outside the range for the nonlinear mapping, i.e., $y \in [0 : 2]$. We used a standard deviation for the measurements of 0.1, so 99.7% of the measurement perturbations are located within the interval $[-0.3 : 0.3]$. Thus, with $1/\sqrt{32} \approx 0.176$ a substantial fraction of the realizations will be located outside the range of y. Also, in ES–MDA_ANA each realization will have its analytic gradient, and together with the excessive perturbations, this introduces a diffusion in the updates in the ES–MDA steps. We noticed that ES-MDA_ANA with four and eight update steps improved on ES, but then with 16 and 32 number of update steps, the results became worse. In fact, if we solve for α from $0.3\sqrt{\alpha} = 1.0$ we obtain $\alpha \approx 11$ which is the threshold where inflated measurements start exceeding the range of y. We also notice that the methods with analytic gradients have some realizations located in the secondary minima.

5.3 Case 3: $x_j^{\mathrm{f}} \leftarrow \mathcal{N}(0.0, C_{xx} = 0.36)$

In this final case, we sample x_j^{f} from $\mathcal{N}(0.0, C_{xx} = 0.36)$. Thus 99.7% of the samples will be located in the interval

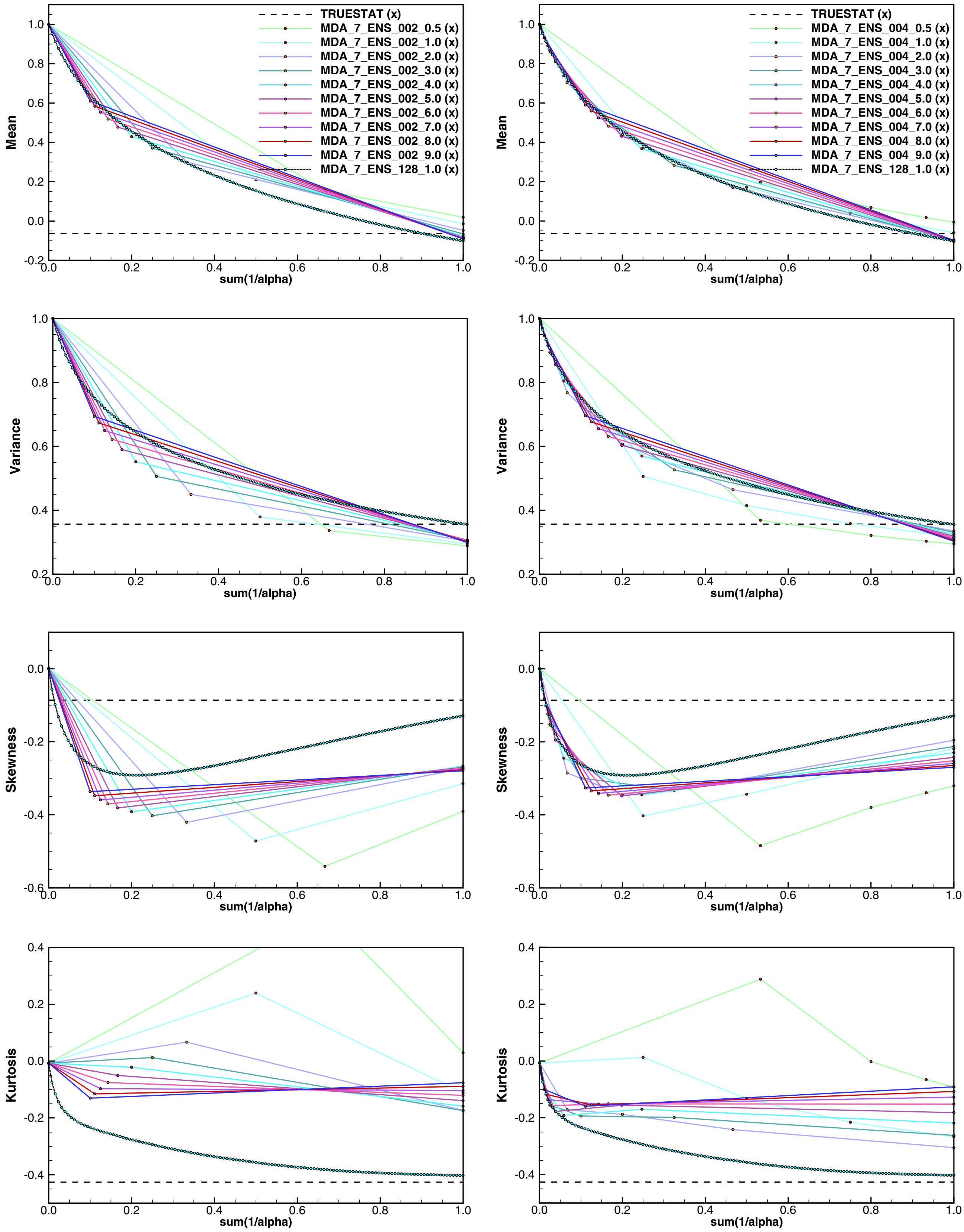

Fig. 10 The plots show the ES–MDA convergence for x with different geometrical schemes for α_i. We show results for the ensemble mean, variance, skewness, and kurtosis from using ES–MDA with two update steps in the left plots and four update steps in the right plots. The dashed black lines are the theoretical values computed from the Bayesian posterior. The line legends given in the upper plots also apply for the remainder of the plots in the respective columns

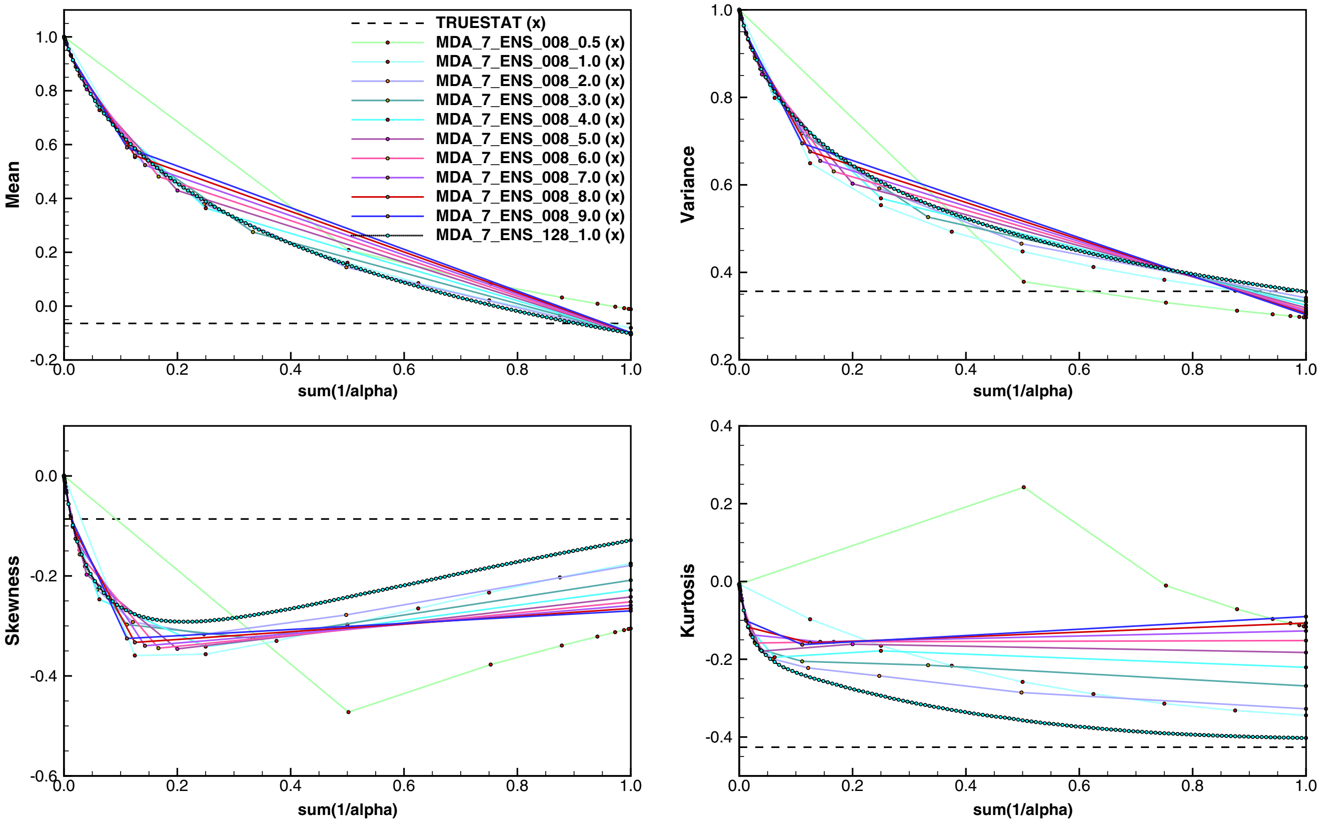

Fig. 11 The plots show ES–MDA Statistics for x with different geometrical schemes for α_i and eight update steps. We show results for ensemble mean, variance, skewness and kurtosis. The line legends given in the upper left plot applies to all the plots

$[-1.8 : 1.8]$ which covers two wavelengths of the functional mapping. The results are shown in in Figs. 14 and 15. Due to the non-monotonic and oscillating behavior of the model, there are five values of x that lead to a prediction of $y = 1$. We see from the Bayesian joint conditional pdf in Fig. 14 that the true solution has five modes corresponding to $x = (-2, -1, 0, 1, 2)$. We also see the multimodal solution in the marginal pdfs for x in Fig. 15.

It is clear that neither ES, ES–MDA or IES with ensemble gradients can reproduce the multimodal posterior solution. However, the methods with ensemble gradients recover the mode at $x = 0$, although in this example ES gives the best result, followed by ES–MDA and then IES.

On the other hand, IES_ANA gives results in very well agreement with the Bayesian posterior with realizations sampling the five significant modes of the system. Also, we noticed that IES_ANA had some realizations sampling additional modes at ±3 (not shown) that are not likely according to the Bayesian posterior. These realizations were probably trapped in local minima. We also observe similar results from ES–MDA_ANA although again with a diffusive behavior as in the previous case.

6 Summary

We have discussed the derivation of the Ensemble Smoother with Multiple Data Assimilation (ES–MDA) and the Iterative Ensemble Smoother (IES) and analyzed their performance with a simple nonlinear scalar model. The derivation provides insight into the approximations that are applied when deriving the two methods and this should help the user to know what to expect from the two iterative smoothers.

We have illustrated the connection between Bayes' theorem and the minimization of an ensemble of cost functions, one for each realization, which is exact in the linear case, and we have thus proved that for a linear model ES, ES–MDA, and IES give the same result and exactly sample the posterior distribution. For a nonlinear model, this connection is only approximate. We have illustrated that IES with an analytic gradient exactly minimizes the ensemble of cost functions and results in a solution that differs from the posterior Bayes' pdf. We have also illustrated how IES is implementing an approximate ensemble-based gradient, which changes the definition of the minima of the cost

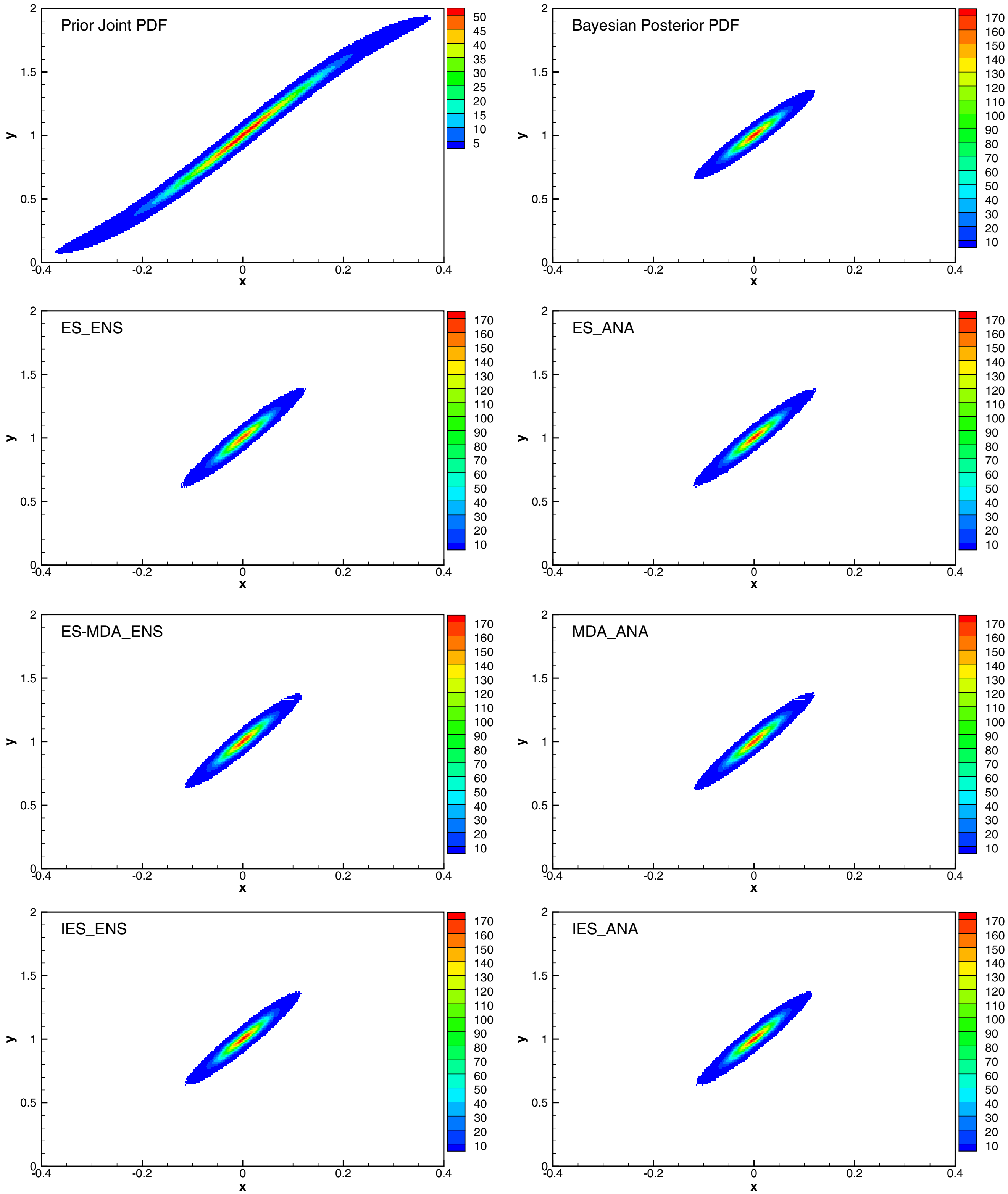

Fig. 12 The plots show joint pdfs for case 1 in the nearly linear case with a narrow prior

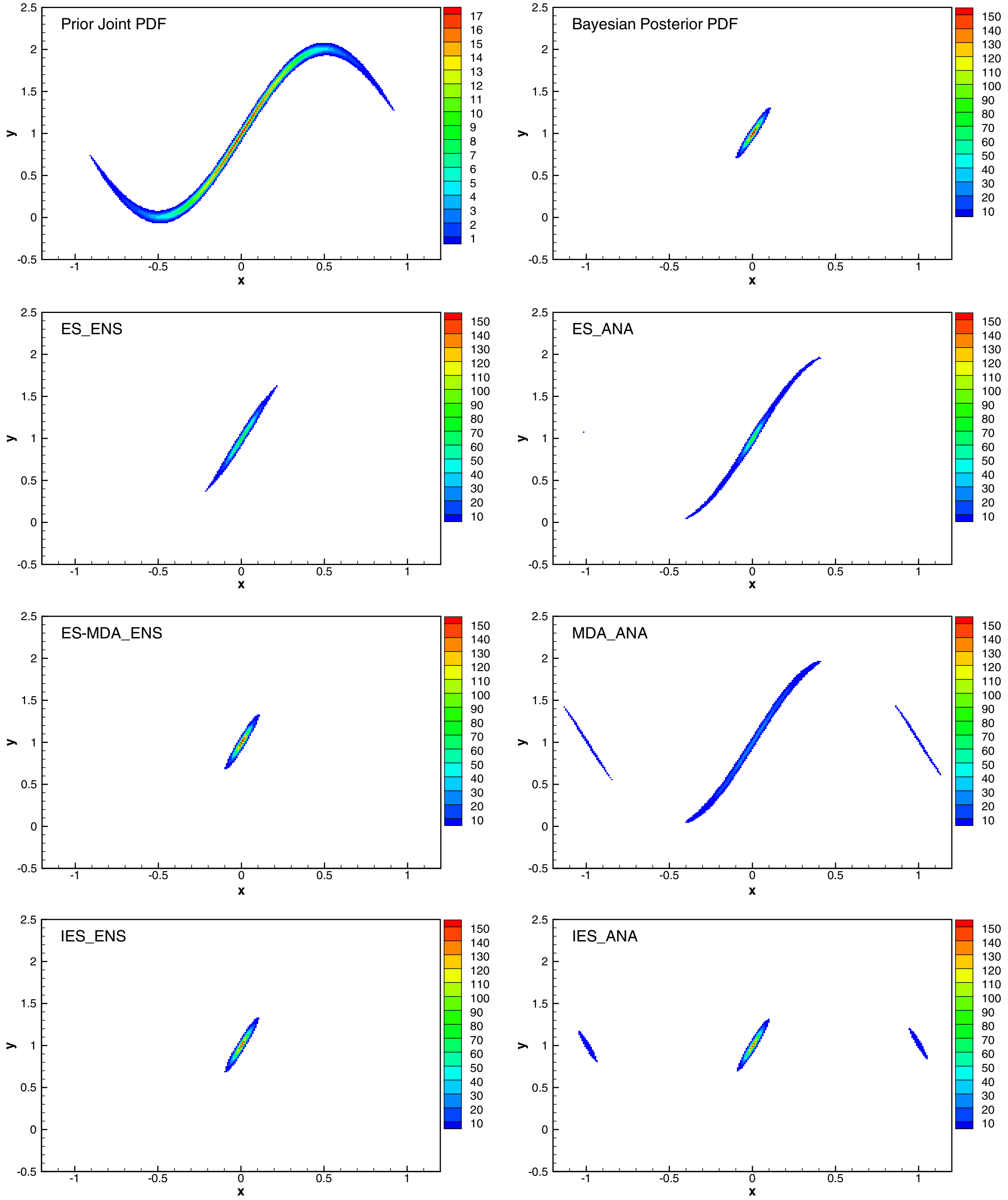

Fig. 13 The plots show joint pdfs for case 2 in the weakly nonlinear case

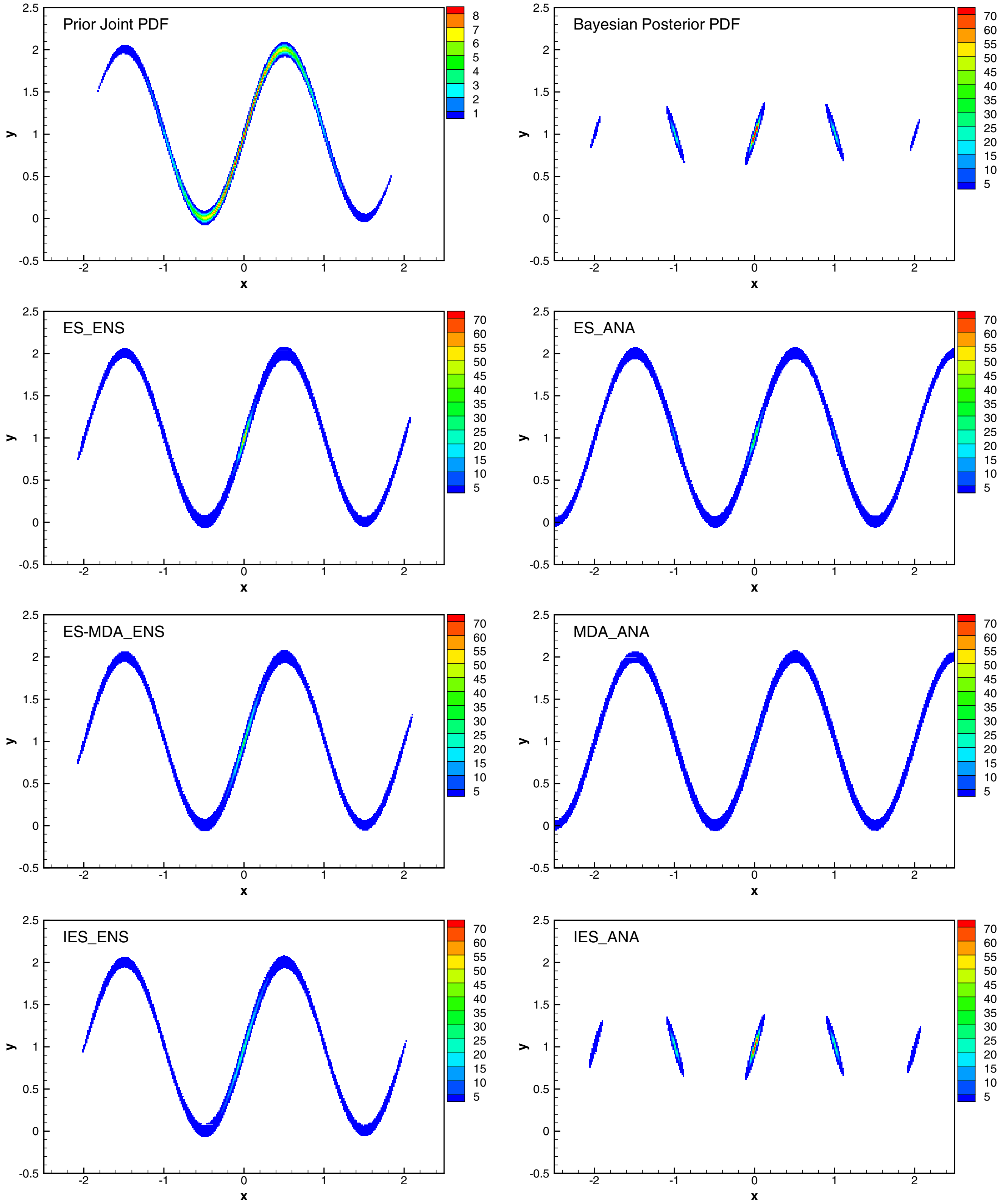

Fig. 14 The plots show joint pdfs for case 3 in the highly nonlinear case with a wide prior that includes multiple modes of the pdf

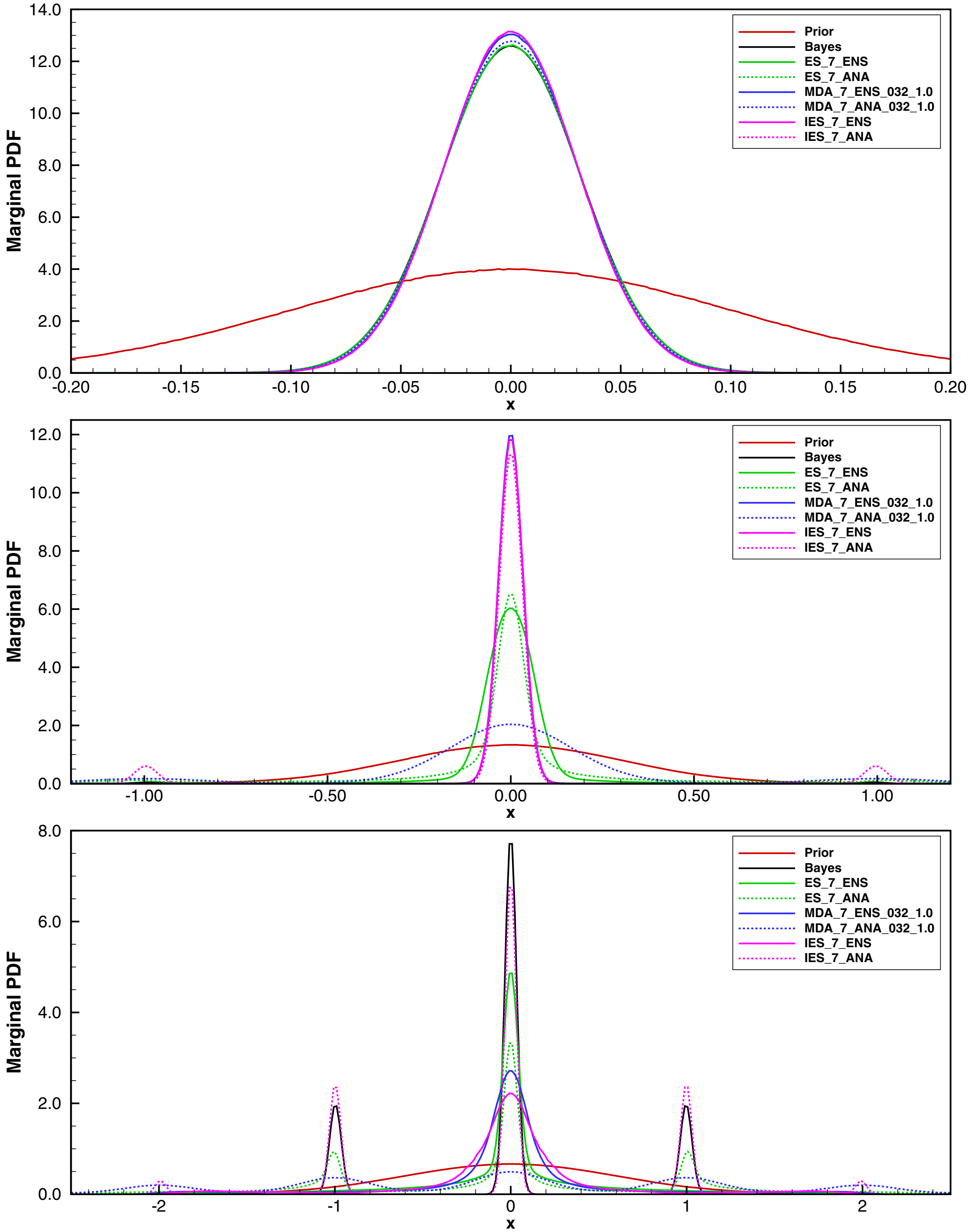

Fig. 15 The plots show the marginal pdfs for x from the cases 1–3 with a different priors from top to bottom

functions minimized for each realization. Thus, the further introduction of an ensemble gradient leads to an additional approximation of the results, but ES and ES–MDA apply similar approximations in the update equations.

We can derive ES–MDA as a direct solver for minimizing the ensemble of cost functions under the assumption of Gaussianity. In the nonlinear case, these equations will not solve for the minima of the of cost functions but instead result in a variance minimizing solution where the prior and likelihood are both assumed to be Gaussian. On the other hand, we can also derive ES–MDA as a solution method for the standard Kalman filter update, which is derived directly from Bayes under the assumption of a Gaussian likelihood.

It was shown that ES–MDA leads to better results with increasing number of MDA steps. The ES–MDA solution appears to converge at 16 to 32 MDA steps, furthermore, with a limited ensemble size of 100 realizations, the sampling errors are masking the improvement in accuracy obtained by running more than about 16 MDA steps.

For low number of ES–MDA steps we could show a benefit of changing the scheme for selecting values of α. This result supports an interpretation of the ES–MDA analysis as a time-stepping algorithm in the pseudo-time defined as $T_n = \sum_{i=1}^{n} 1/\alpha_i$, where the accuracy of the pseudo-time stepping is dependent on the nonlinearity of the model.

An additional highly nonlinear case, which exhibits multiple modes, was also run to examine the range of validity of smoother methods with nonlinearity, as well as to study their properties with non-monotonic mappings. An overall conclusion is that all the smoothers work well with weakly nonlinear models. Furthermore, the use of an ensemble gradient which is the same for all ensemble members, prevents that different realizations may converge to different modes of the pdf as is the case when we use the analytic gradient (see also Section 2.5 in [4]). For a highly nonlinear model with multiple modes, none of the smoothers can correctly solve for the conditional posterior. We noticed that using IES with an exact analytic representation of the gradient; it is possible to obtain an accurate representation of the posterior conditional pdf also in the multimodal case. However, for a practical implementation, the use of an analytic gradient will require the use of adjoints models, and the computational problem becomes immense. Also, when using ES–MDA with models that map the prior parameters into a bounded range of values, the method will have difficulties when inflated measurements exceed this range.

It is clear from the experiments that iterations in IES or multiple update steps in ES–MDA reduce the impact of weak nonlinearity and lead to better results than what can be obtained from ES. So which method to choose? For numerical efficiency, it is advised to use ES for all preliminary experiments until a final production simulation is ready to be run. Thereafter, we can use both ES–MDA and IES. ES–MDA has the advantage (or disadvantage) that one can predefine the number of steps, and also reuse the numerical implementation from ES, and the method is conceptually easy to understand and implement. However, a large number of steps may be needed to obtain a converged result. IES may require fewer iterations to converge, but the method requires a separate implementation, and convergence issues may show up if we choose poor values for the step length. For now, we conclude that neither ES–MDA or IES precisely sample the posterior pdf from Bayes', but it appears that the optimal choice of method will depend on the degree of nonlinearity and the properties of the model used. We have previously seen several examples of both methods giving consistent results, e.g., when history matching reservoir simulation models in the study by Evensen and Eikrem [Strategies for conditioning reservoir models on rate data using ensemble smoothers, under review].

Acknowledgements This work was supported by a research project funded by Statoil and has benefited from the interaction and collaborations with the Nordforsk Nordic center of excellence in data assimilation, EMBLA. The author is grateful for constructive comments by three anonymous reviewers that improved the readability of the paper.

References

1. Aanonsen, S.I., Naevdal, G., Oliver, D.S., Reynolds, A., Valles, B.: Ensemble Kalman filter in reservoir engineering – A review. SPE J. **14**(3), 393–412 (2009). https://doi.org/10.21188/117274-PA. SPE-117274-PA
2. Bocquet, M., Sakov, P.: An iterative ensemble Kalman smoother. Q. J. R. Meteorol. Soc. **140**, 1521–1535 (2014)
3. Burgers, G., van Leeuwen, P.J., Evensen, G.: Analysis scheme in the ensemble Kalman filter. Mon. Weather. Rev. **126**, 1719–1724 (1998)
4. Chen, Y., Oliver, D.S.: Ensemble randomized maximum likelihood method as an iterative ensemble smoother. Math. Geosci. **44**, 1–26 (2012)
5. Chen, Y., Oliver, D.S.: Levenberg-Marquardt forms of the iterative ensemble smoother for efficient history matching and uncertainty quantification. Comput. Geosci. **17**, 689–703 (2013)
6. Emerick, A.A.: Analysis of performance of ensemble-based assimilation of production and seismic data. J. Petrol. Sci. Eng. **139**, 219–239 (2016)
7. Emerick, A.A., Reynolds, A.C.: History matching time-lapse seismic data using the ensemble Kalman filter with multiple data assimilations. Comput. Geosci. **16**(3), 639–659 (2012)
8. Emerick, A.A., Reynolds, A.C.: Ensemble Smoother with multiple data assimilation. Comput. Geosci. **55**, 3–15 (2013)
9. Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res. **99**(C5), 10,143–10,162 (1994)
10. Evensen, G.: Sampling strategies and square root analysis schemes for the EnKF. Ocean Dyn. **54**, 539–560 (2004)
11. Evensen, G.: Data Assimilation: the Ensemble Kalman Filter, 2nd edn. Springer, Berlin (2009)
12. Evensen, G.: The ensemble Kalman filter for combined state and parameter estimation. IEEE Control. Syst. Mag. **29**(3), 83–104 (2009)
13. Evensen, G., van Leeuwen, P.J.: An ensemble Kalman smoother for nonlinear dynamics. Mon. Weather. Rev. **128**, 1852–1867 (2000)

14. Iglesias, M.A.: Iterative regularization for ensemble data assimilation in reservoir models. Comput. Geosci. **19**(1), 177–212 (2015)
15. Iglesias, M.A.: A regularizing iterative ensemble Kalman method for PDE-constrained inverse problems. Inverse Prob. **32**(2), (2016). https://doi.org/10.1088/0266-5611/32/2/025002
16. Kitanidis, P.K.: Quasi-linear geostatistical therory for inversing. Water Resour. Res. **31**(10), 2411–2419 (1995)
17. Le, D.H., Emerick, A.A., Reynolds, A.C.: An adaptive ensemble smoother with multiple data assimilation for assisted history matching. SPE Journal, SPE-173214-PA **21**(6), 2195–2207 (2016)
18. Luo, X., Stordal, A.S., Lorentzen, R.J., Nævdal, G.: Iterative Ensemble Smoother as an approximate solution to a regularized minimum-average-cost problem: Theory and applications. SPE Journal, SPE-176023-PA **20**(5), 962–982 (2015)
19. Nævdal, G., Johnsen, L.M., Aanonsen, S.I., Vefring, E.: Reservoir monitoring and continuous model updating using the ensemble Kalman filter. In: SPE Annual Technical Conference and Exhibition (SPE 84372) (2003)
20. Neal, R.M.: Sampling from multimodal distributions using tempered transitions. Stat. Comput. **6**(4), 353–366 (1996)
21. Oliver, D.S., He, N., Reynolds, A.C.: Conditioning Permeability Fields to Pressure Data. In: ECMOR – 5th European Conference on the Mathematics of Oil Recovery (1996)
22. Rafiee, J., Reynolds, A.C.: Theoretical and efficient practical procedures for the generation of inflation factors for ES–MDA. Inverse Problems **33**(11), 115003 (2017)
23. Sakov, P., Oliver, D.S., Bertino, L.: An iterative EnKF for strongly nonlinear systems. Mon. Weather. Rev. **140**, 1988–2004 (2012)
24. Skjervheim, J.A., Evensen, G., Hove, J., Vabø, J.: An ensemble smoother for assisted history matching. SPE 141929 (2011)
25. Stordal, A., Elsheikh, A.H.: Iterative ensemble smoothers in the annealed importance sampling framework. Adv. Water Resour. **86**, 231–239 (2015)
26. van Leeuwen, P.J., Evensen, G.: Data assimilation and inverse methods in terms of a probabilistic formulation. Mon. Weather. Rev. **124**, 2898–2913 (1996)