Comput Geosci (2011) 15:307–323
DOI 10.1007/s10596-010-9205-3

ORIGINAL PAPER

Multimodal ensemble Kalman filtering using Gaussian mixture models

Laura Dovera · Ernesto Della Rossa

Received: 27 November 2009 / Accepted: 5 August 2010 / Published online: 18 August 2010

Abstract In this paper we present an extension of the ensemble Kalman filter (EnKF) specifically designed for multimodal systems. EnKF data assimilation scheme is less accurate when it is used to approximate systems with multimodal distribution such as reservoir facies models. The algorithm is based on the assumption that both prior and posterior distribution can be approximated by Gaussian mixture and it is validated by the introduction of the concept of finite ensemble representation. The effectiveness of the approach is shown with two applications. The first example is based on Lorenz model. In the second example, the proposed methodology combined with a localization technique is used to update a 2D reservoir facies models. Both applications give evidence of an improved performance of the proposed method respect to the EnKF.

Keywords Data assimilation · Ensemble Kalman filter · Gaussian mixture

1 Introduction

In literature, many application examples of data assimilation problems can be found in different domains: weather forecast, oceanic and atmospheric modeling, and oil/gas reservoir models. The main problem related to sequential data assimilation is the estimation of the hidden states of a dynamical system as a set of new observations becomes available. The recursively Bayesian updating is a probabilistic solution to the filtering problem that gives the posterior probability density function (PDF) of the system state given new observations. However, for real problems, the Bayesian recursion is intractable because of the dimensions of the model system. Consequently, the main challenge in this domain is to find efficient methods that provide approximate solutions.

A promising approach for high-dimensional problems is the Ensemble Kalman filter (EnKF) [1, 7]. The EnKF is a Monte Carlo method: an ensemble of prior models is generated to represent the mean and the covariance of variables and it is updated when data are sequentially assimilated. The method has been applied in ocean dynamical models [16], in meteorology [17] and, recently, in reservoir simulation models [23]. These examples show that the EnKF can handle different types of complex and nonlinear model systems.

The EnKF was introduced to overcome some of the problems of the extended Kalman filter [21] for strongly nonlinear models. In particular, the EnKF has the advantage that the explicit linearization of the model function is not required but it uses an ensemble of models from which all the necessary statistics can be computed. At each assimilation step each member of the ensemble is updated by linearly integrating new observations. The updating equation is based on Bayesian conditional distribution with a Gaussian prior whose mean and covariance are estimated using the ensemble. It can be shown [4] that, for linear dynamics, linear measurements, and Gaussian likelihood, the resulting

L. Dovera (✉) · E. Della Rossa
Eni Exploration & Production, San Donato Milanese, Milan, Italy
e-mail: laura.dovera@eni.com

E. Della Rossa
e-mail: ernesto.dellarossa@eni.com

updated ensemble approximates the theoretical posterior distribution if the ensemble size is sufficiently large. If a Gaussian prior model is assumed at each assimilation step, the EnKF updating of variables characterized by multimodal distributions may introduce some problems, [6, 27]. There are mainly two types of multimodality: multimodal features can be found in the initial dataset or they can be introduced by strong nonlinearities of the dynamical model. Both these cases represent a limit in the range of applicability of the EnKF. In reservoir modeling data assimilation problems data often have complex non-Gaussian distributions and the dynamic of the system is nonlinear. Most reservoir models are characterized by a distribution of facies that is generally guided by a spatial correlation. The spatial distribution of petrophysical properties within each facies can be modeled by a Gaussian random function. If the facies are very different in terms of petrophysical properties distributions, a unique Gaussian model may be not adequate to represent the natural heterogeneity of the reservoir. A suitable ensemble based data assimilation technique should be able to deal with these facies types of reservoir models. In fact, several approaches in this direction have been proposed: an overview on the subject is given in [1]. One of these methods has been proposed by us and it is presented in a preliminary form in [6].

Our purpose is to extend the EnKF and to develop an assimilation technique applicable to multimodal distributions. In particular, we assume Gaussian mixture models (GMMs) to approximate multimodal priors and we modify the EnKF in order to update Gaussian mixture (GM) distributions. The classical result of the conditional multivariate Gaussian can be extended to a multivariate Gaussian mixture [2] and it can be shown that, in the linear case with linear measurements and with Gaussian likelihood, the posterior distribution with a GM prior is again a GM. In our work, we use this result to reformulate the ensemble update equations when the prior ensemble is considered as a GMM sample. Following EnKF scheme, it is shown that, for a linear forward model, a linear measurements operator and a large ensemble, the proposed updating approach correctly approximates the posterior PDF. After each step, the adjusted ensemble is taken through the forward model to the next observation time. When the model dynamics are nonlinear, the parameters of the GMM prior have to be re-estimated using the ensemble. The expectation–maximization (EM) algorithm, [5, 15], is proposed to solve the estimation problem, and it provides an alternative for maximum likelihood estimation of the GM parameters.

GMM have been used in previous works in the context of particle filtering in different domains, for example, [26]. EnKF combined with GM have been used in [3] but the presented problem perspective is not specifically oriented to multimodal priors and the proposed method does not involve the use of the EM algorithm. The EM algorithm combined with mixture priors has been used in [25], but also in that case the updating equations are different because the posterior distribution is assumed Gaussian; hence the ensemble sampling associated to mixture components in the posterior mixture is not considered. Whereas, in our method the main focus is the modeling of the multimodal distribution in the context of reservoir facies modeling. Differently from [3], we are interested in maintaining the link between the GM components and facies and differently from [25] we aim at sampling a Gaussian mixture posterior. For this reason, even if some aspects of the method have been used in other fields, the updating procedure presented here and the application to spatially correlated random field are new. In addition, respect to [6], we added the validation of the posterior equations introducing the notion of *finite ensemble representation*. Using this concept, we show that our updating is consistent with the one given by the EnKF.

The paper is organized as follows. Section 2 contains the background and a general introduction to data assimilation problem and EnKF. In Section 3, we illustrate the EnKF for GMM that is the main contribution of the work: in particular, we derive the EnKF equations to update GM distributions and we explain the use of the EM algorithm in the method. In Section 4, we apply the method to the Lorenz model and in Section 5, we present a 2D single phase reservoir data assimilation problem with multimodal permeability distribution associated to facies. In both the examples we compare the results obtained with those of the standard EnKF data assimilation. Conclusions are given in Section 6, and technical details are given in Appendix.

2 Background and notation

2.1 Data assimilation

In data assimilation problem, the goal is to estimate the hidden model states $\left[\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_T\right]$ of a system by integrating available observations $[\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_T]$ where $t = 0, \ldots, T$ is a discrete time index. The exact solution to the problem would be obtained by evaluating the probability of $\mathbf{y}_T$ given $[\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_T]$. If the

process is first order Markovian, assuming conditional independence and single-state dependence, Bayesian formulation allows to obtain the posterior probability model:

$$\begin{aligned}\left[\mathbf{y}_0,\ldots,\mathbf{y}_T|\mathbf{d}_0,\ldots,\mathbf{d}_T\right] &\sim f(\mathbf{y}_0,\ldots,\mathbf{y}_T|\mathbf{d}_0,\ldots,\mathbf{d}_T)\\ &= \text{const}\times f(\mathbf{y}_0)\,f(\mathbf{d}_0|\mathbf{y}_0)\\ &\times\left[\prod_{t=0}^{T-1} f(\mathbf{d}_{t+1}|\mathbf{y}_{t+1})\,f(\mathbf{y}_{t+1}|\mathbf{y}_t)\right] \quad (1)\end{aligned}$$

being *const* the normalizing constant.

In sequential data assimilation, this PDF in Eq. 1 is generally computed with recursive algorithms that consist in a forecast step and an updating step. Once given, the initial PDF $f(\mathbf{y}_0)$, in the forecast step the model PDF $f(\mathbf{y}_t)$ is forward propagated using the system dynamics to give a new forecast $f(\mathbf{y}_{t+1}|\mathbf{y}_t)$, in the updating step the PDF is corrected to honor observations on the basis of the likelihood $f(\mathbf{d}_{t+1}|\mathbf{y}_{t+1})$. In particular, for each assimilation time step t, we can define the forecast state as $\mathbf{y}^f_{t+1}=\mathbf{y}_{t+1}|\mathbf{y}_t$ and the updated state as $\mathbf{y}^u_{t+1}=\mathbf{y}_{t+1}|\mathbf{d}_{t+1}$ (in the following, the time index t will be omitted and we will indicate the forecast state as $\mathbf{y}^f$ and the updated state as $\mathbf{y}^u$). The filtering process consists in a time recursive assessment of the forecast prior $f(\mathbf{y}^f)$, guided by the system dynamics, followed by a Bayesian updating of the prior to get the updated posterior $f(\mathbf{y}^u)$. An approximate solution of the data assimilation can be obtained by Kalman filter update equation [18] that is a Bayesian Least Squares estimator chosen in a linear class of functions (of measurements). If the exact posterior distribution is required, in the linear and Gaussian case, the analytical solution of Eq. 1 is known.

2.2 The ensemble Kalman filter

The EnKF is a powerful technique to solve nonlinear dynamics data assimilation problems in large scale systems. The basic idea of the EnKF is to approximate the forecast probability distributions by a finite set of model realizations called the ensemble. These realizations are updated according to the likelihood when data are integrated. The system dynamic is then used to forward the corresponding ensemble at the next time step.

At a given time t, the forecast distribution $f(\mathbf{y}^f)$ is explicitly represented by an ensemble of state variables $\{\mathbf{y}^f_j\}_{j=1,\ldots,N_e}$, being N_e the ensemble size. In the EnKF, the ensemble is used to compute the mean $\overline{\mathbf{y}^f}$ and the covariance $\widehat{\mathbf{C}}_{\mathbf{y}^f}$ estimators:

$$\overline{\mathbf{y}^f}=\frac{1}{N_e}\sum_{j=1}^{N_e}\mathbf{y}^f_j; \quad (2)$$

$$\widehat{\mathbf{C}}_{\mathbf{y}^f}=\frac{1}{N_e-1}\sum_{j=1}^{N_e}(\mathbf{y}^f_j-\overline{\mathbf{y}^f})(\mathbf{y}^f_j-\overline{\mathbf{y}^f})^T. \quad (3)$$

Suppose that a linear operator $\mathbf{H}$ relates the unobserved model $\mathbf{y}^f$ to the data $\mathbf{d}_{\text{obs}}$ as

$$\mathbf{d}_{\text{obs}}=\mathbf{H}\mathbf{y}^f+\boldsymbol{\varepsilon} \quad (4)$$

with $\boldsymbol{\varepsilon}$ random vector representing a measurement error with normal distribution $N(\boldsymbol{\varepsilon};\mathbf{0},\mathbf{C}_{\boldsymbol{\varepsilon}})$ independent of the model $\mathbf{y}^f$. Both the operator $\mathbf{H}$ and the data $\mathbf{d}_{\text{obs}}$ depend on the time. Here, we will omit this dependence since we are focusing on the single-time step filtering recursion. The EnKF update step consists of applying an update equation to each ensemble member $\{\mathbf{y}^f_j\}_{j=1,\ldots,N_e}$ as follows

$$\mathbf{y}^u_j=\mathbf{y}^f_j+\widehat{\mathbf{C}}_{\mathbf{y}^f}\mathbf{H}^T(\mathbf{H}\widehat{\mathbf{C}}_{\mathbf{y}^f}\mathbf{H}^T+\mathbf{C}_{\boldsymbol{\varepsilon}})^{-1}(\mathbf{d}_j-\mathbf{H}\mathbf{y}^f_j), \quad (5)$$

where $\mathbf{d}_j$ is the jth set of randomized measurements $\mathbf{d}_j=\mathbf{d}_{\text{obs}}+\boldsymbol{\varepsilon}_j$, $j=1,\ldots,N_e$, $\mathbf{d}_{\text{obs}}$ the measurement vector, $\boldsymbol{\varepsilon}_j\sim N(\boldsymbol{\varepsilon};\mathbf{0},\mathbf{C}_{\boldsymbol{\varepsilon}})$.

The importance of treating the observations $\mathbf{d}_j$ as random variables with a distribution with mean equal to the original observation $\mathbf{d}_{\text{obs}}$ and covariance equal to $\mathbf{C}_{\boldsymbol{\varepsilon}}$ was pointed out in [4]. With this additional assumption, it is proved in [4] that, in the linear case, the mean and covariance estimators of the updated ensemble $\{\mathbf{y}^u_j\}$ approximate the conditional mean and covariance of the posterior distribution given by the Kalman filter when the ensemble size is sufficiently large. A probabilistic treatment of the EnKF convergence to the Kalman filter is given in [20].

In our work, we consider the properties of the ensemble based estimators (mean and covariance) respect to the parameters of the conditional distribution of Kalman filter only in terms of the difference in Euclidean norm. For this reason we introduce the notion of *finite ensemble representation* of a Gaussian distribution and we discuss the approximation property of EnKF equations using this concept.

We consider a set of vectors $\{\mathbf{y}^f_j\}_{j=1}^{N_e}$ as a *finite ensemble representation* of a random Gaussian vector $\mathbf{y}^f$, with

mean $\boldsymbol{\mu}_{\mathbf{y}^f}$ and covariance $\mathbf{C}_{\mathbf{y}^f}$, with precision $(\eta_{\boldsymbol{\mu}}, \eta_{\mathbf{C}})$ when

$$\|\frac{1}{N_e}\sum_{j=1}^{N_e}\mathbf{y}_j^f - \boldsymbol{\mu}_{\mathbf{y}^f}\| \leq \eta_{\boldsymbol{\mu}}$$

$$\|\frac{1}{N_e}\sum_{j=1}^{N_e}\left(\mathbf{y}_j^f - \boldsymbol{\mu}_{\mathbf{y}^f}\right)\left(\mathbf{y}_j^f - \boldsymbol{\mu}_{\mathbf{y}^f}\right)^T - C_{\mathbf{y}^f}\| \leq \eta_{\mathbf{C}}.$$

The norm of vectors is the usual euclidean norm, i.e. if $\mathbf{a} \in R^N$, $\|\mathbf{a}\| = \sqrt{\mathbf{a}^T \cdot \mathbf{a}}$, while for a matrix A is the corresponding $\|A\| = \sup_{\mathbf{y}\neq\mathbf{0}} \frac{\|A\mathbf{y}\|}{\|\mathbf{y}\|}$.

If we consider a linear transformation $\mathbf{z} = L\mathbf{y}^f$ of the random vector $\mathbf{y}^f$ and the corresponding linear transform $\mathbf{z}_j = L\mathbf{y}_j^f$ of the ensemble, we can observe that $\{\mathbf{z}_j\}_{j=1}^{N_e}$ is again a *finite ensemble representation* of $\mathbf{z}$ with precision at least $(\|L\| \cdot \eta_{\boldsymbol{\mu}}, \|L\|^2 \cdot \eta_{\mathbf{C}})$. This allows to conclude that if $\{\mathbf{y}_j^f\}_{j=1}^{N_e}$ is a *finite ensemble representation* of $\mathbf{y}^f$, then $\mathbf{y}_j^u$, obtained in Eq. 5, is a *finite ensemble representation* of $\mathbf{y}^u \equiv \mathbf{y}|\mathbf{d}_{\text{obs}}$. The technical details about these results are given in the Appendix.

When the EnKF is used with nonlinear and non-Gaussian models, there are two levels of approximation. The first approximation is in the forward step where the model dynamic is used to propagate the ensemble members forward in time. In this way the estimated PDF can take into account non-Gaussian and nonlinear effect. The increase of the size of the ensemble improves the approximation of mean and covariance estimation, however the accuracy of ensemble Kalman filter methods depends on the size of the sample compared to the dimension of the parameters space. In real applications the number of variables often is orders of magnitude larger than the number of the models that can be computationally borne. In such cases, sampling error may result in spurious correlations which seriously deteriorate the results, producing a bias in the mean and a strong underestimation of the uncertainty. Several techniques have been proposed to reduce sampling errors for a finite ensemble size [10]. The second level of approximation is in the update. The data conditioning is in fact linear and Gaussian with weights (given by the Kalman gain) recursively estimated from the ensemble. The linearization suggests that for strongly nonlinear or non-Gaussian problems, the filter may provide unreliable results. On the other side these approximations make the EnKF easy to implement and computationally efficient. In fact the forecast step is intrinsically parallel, because each member of the ensemble can be simulated independently. Moreover, the linear algebra operations required in the updating are generally computationally cheaper than the forward simulation, because the updating requires the inversion of a symmetric positive definite matrix sized as the number of observations that in most of data assimilation problems are very small compared to the length of the state vector.

As pointed out in [27], multimodal systems are an example of non-Gaussian prior models where the EnKF assimilation updating encounters problems. Our target is to focus on this issue and to suggest an extension of the method to properly update also non-Gaussian priors, approximating multimodal distributions and updating multimodal systems. As in the Gaussian case, the goal is to derive an EnKF update procedure for GMM that approximates the posterior distribution corresponding to the linear case when the size of the ensemble is large.

3 Ensemble Kalman filter for Gaussian mixture models

3.1 Kalman filter for Gaussian mixture models

Non-Gaussian distributions and in particular multimodal distributions can be parametrically well described as weighted sums of Gaussian PDFs (Gaussian mixture models), [15]. For linear Bayesian problems with GMM prior and Gaussian likelihood, the posterior density is also a GMM and the Kalman filter update can be extended to a mixture of Gaussian distributions, [2]. Suppose that the PDF of the forecast model $\mathbf{y}^f$, at a given time t, is a Gaussian mixture. Then we can assume that the PDF of $\mathbf{y}^f$ is given by

$$f(\mathbf{y}^f) = \sum_{k=1}^{N_c}\pi_k f_k(\mathbf{y}^f), \quad \sum_{k=1}^{N_c}\pi_k = 1, \quad \pi_k \geq 0$$

$$f_k(\mathbf{y}^f) = N(\mathbf{y}^f; \boldsymbol{\mu}_{\mathbf{y}^f}^k, \mathbf{C}_{\mathbf{y}^f}^k), \quad k = 1, \ldots, N_c. \tag{6}$$

The π_k are the mixture weights and they can be interpreted as the probability of each component of the mixture.

From observations definition of Eq. 4, where the measurement operator is assumed linear and the error is assumed Gaussian and from the properties of GMM [2], it follows that the conditional $f(\mathbf{y}^u) = f(\mathbf{y}^f|\mathbf{d}_{\text{obs}})$ is again a GM given by

$$f(\mathbf{y}^f|\mathbf{d}_{\text{obs}}) = \sum_{k=1}^{N_c}\lambda_{\mathbf{y}^f|\mathbf{d}_{\text{obs}}}^k f_k(\mathbf{y}^f|\mathbf{d}_{\text{obs}}) \tag{7}$$

where each conditional component $f_k(\mathbf{y}^f|\mathbf{d}_{\text{obs}})$ is normal $N(\mathbf{y}^u; \boldsymbol{\mu}^k_{\mathbf{y}^f|\mathbf{d}_{\text{obs}}}, \mathbf{C}^k_{\mathbf{y}^f|\mathbf{d}_{\text{obs}}})$ with mean and covariance given by

$$\begin{aligned}\boldsymbol{\mu}^k_{\mathbf{y}^f|\mathbf{d}_{\text{obs}}} &= \boldsymbol{\mu}^k_{\mathbf{y}^f} + \mathbf{K}^k_{\mathbf{H}}(\mathbf{d}_{\text{obs}} - \mathbf{H}\boldsymbol{\mu}^k_{\mathbf{y}^f})\\ \mathbf{C}^k_{\mathbf{y}^f|\mathbf{d}_{\text{obs}}} &= \mathbf{C}^k_{\mathbf{y}^f} - \mathbf{K}^k_{\mathbf{H}}\mathbf{H}\mathbf{C}^k_{\mathbf{y}^f}\\ \mathbf{K}^k_{\mathbf{H}} &= \mathbf{C}^k_{\mathbf{y}^f}\mathbf{H}^T(\mathbf{H}\mathbf{C}^k_{\mathbf{y}^f}\mathbf{H}^T + \mathbf{C}_{\boldsymbol{\varepsilon}})^{-1}\end{aligned} \tag{8}$$

and the posterior weights $\lambda^k_{\mathbf{y}^f|\mathbf{d}}$ are computed as

$$\lambda^k_{\mathbf{y}^f|\mathbf{d}_{\text{obs}}} = \frac{\pi_k N(\mathbf{d}_{\text{obs}}; \boldsymbol{\mu}^k_{\mathbf{d}_{\text{obs}}}, \boldsymbol{\Sigma}^k_{\mathbf{d}_{\text{obs}},\mathbf{d}_{\text{obs}}})}{\sum_{\ell=1}^{N_c}\pi_\ell N(\mathbf{d}_{\text{obs}}; \boldsymbol{\mu}^\ell_{\mathbf{d}_{\text{obs}}}, \boldsymbol{\Sigma}^\ell_{\mathbf{d}_{\text{obs}},\mathbf{d}_{\text{obs}}})}$$

with

$$\begin{aligned}\boldsymbol{\mu}^k_{\mathbf{d}_{\text{obs}}} &= \mathbf{H}\boldsymbol{\mu}^k_{\mathbf{y}^f}\\ \boldsymbol{\Sigma}^k_{\mathbf{d}_{\text{obs}},\mathbf{d}_{\text{obs}}} &= \mathbf{H}\mathbf{C}^k_{\mathbf{y}^f}\mathbf{H}^T + \mathbf{C}_{\boldsymbol{\varepsilon}}, \quad k = 1, \ldots, N_c.\end{aligned} \tag{9}$$

If $N_c = 1$, the expressions are equivalent to the Kalman filter. When $N_c > 1$, the coefficients $\lambda^k_{\mathbf{y}^f|\mathbf{d}_{\text{obs}}}$ give the posterior probabilities of the GM components. Mean and covariance of each component are updated as in the Gaussian case. Similarly to the Gaussian case we can also define a Kalman gain matrix for each mixture component as

$$\mathbf{K}^k_{\mathbf{H}} = \mathbf{C}^k_{\mathbf{y}^f}\mathbf{H}^T(\mathbf{H}\mathbf{C}^k_{\mathbf{y}^f}\mathbf{H}^T + \mathbf{C}_{\boldsymbol{\varepsilon}})^{-1}, \quad k = 1, \ldots, N_c.$$

3.2 Ensemble Kalman filter for GMM

In the following we describe the new EnKF-GMM technique that represents the main contribution of the paper. As in the EnKF, the basic idea is to use discrete probability densities represented by an ensemble of model states. Each state of the ensemble is propagated forward in time using directly the model dynamics even with nonlinear forward functions. In the case of GM priors the approach meets two issues.

(a) As the ensemble is assumed to be sampled from a GM PDF, the GM parameters have to be directly estimated by the ensemble. Our proposal is based on the expectation–maximization algorithm, [5, 15, 25] as an alternative to estimate GM parameters i.e. means and covariances of the components and mixture weights.

(b) To update the ensemble members and to approximate the updated posterior distribution, as in the EnKF update, we provide an assimilation scheme that in the case of linear forward, linear measurement operator and gaussian error, samples the posterior Gaussian mixture distribution when the ensemble size is sufficiently large.

Suppose that the ensemble $\{\mathbf{y}^f_j\}_{j=1,\ldots,N_e}$ is sampled from a GM forecast prior $f(\mathbf{y}^f)$ as defined in Eq. 6. Ensemble estimators for the GM parameters, namely weights ($\widehat{\pi}_k$), means ($\widehat{\boldsymbol{\mu}}^k_{\mathbf{y}^f}$) and covariances ($\widehat{\mathbf{C}}^k_{\mathbf{y}^f}$), can be found using the expectation–maximization algorithm [5]. The EM algorithm can be used for maximum likelihood estimation from data sets with missing or hidden variables [15] and, as a special case, for Gaussian mixtures parameters estimation. EM is an iterative algorithm that allows us to find maximum likelihood estimates of parameters in probabilistic models in the presence of missing data. It is a two-step method. The expectation step computes an expectation of the log likelihood with respect to the current estimate of the distribution. The maximization step maximizes the expected log likelihood found in the previous step. The algorithm converges to the optimal solution in a number of steps, which depends on different factors such as distribution shape and data dimensions [15]. It is used here to estimate parameters means and covariance matrices and weights of Gaussian components of the mixture for joint distribution in the case of multimodality of the data. The problem of the identification of the number of components can be addressed in the context of model assessment and selection and it can be faced with Bayesian information criterium or other similar techniques [15, 25]. However, in several applications this number can be considered part of the prior knowledge of the system state. In reservoir models petrophysical properties, as porosity or permeability, are typically related to geological units (facies), as rock types or sedimentological classes. Variables inside facies are characterized by different Gaussian PDFs and the resulting underlying distribution becomes multimodal. In our work, we describe the multimodality with GMM. In this way, few classes with a different geological description are distinguished by a different Gaussian distribution. The proportion of each discrete variable given by the facies is instead guided by the mixture weights. A scalar example where GMM are used with this purpose is presented in [6]. The application of GMM and EM to identify petrophysical properties and litho fluid classes in real reservoirs is presented in [13] and [12]. The use of GMM is somewhat unusual in the context of data assimilation where GMM are typically adopted to approximate general forms of PDF using a lot of mixture components.

In the proposed method the EM algorithm is applied at each assimilation step on the forecasted ensemble

and it gives the membership probability $\widehat{\pi}_k$ for each element of the ensemble to belong to component k and the corresponding mean $\widehat{\boldsymbol{\mu}}^k_{\mathbf{y}^f}$ and covariance $\widehat{\mathbf{C}}^k_{\mathbf{y}^f}$ for each component $k = 1, \ldots, N_c$. The number of component is considered as an assigned prior information. Let now $\mathbf{d}_{\text{obs}}$ be the measurement vector as in Eq. 4 and $\mathbf{d}_j$, $j = 1, \ldots, N_e$, the set of measurements with $\mathbf{d}_j = \mathbf{d}_{\text{obs}} + \boldsymbol{\varepsilon}_j$, $j = 1, \ldots, N_e$, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{C}_{\boldsymbol{\varepsilon}})$. In order to obtain the samples of the posterior GMM, we perform a loop through the ensemble members, for each member select randomly an integer according to the discrete distribution of the weights (as computed by EM) and move the member to the new component according to the new mean and covariance of the new component. In detail we propose an updating procedure based on the following steps.

(1) For each component $\ell = 1, 2, \ldots, N_c$ compute $\lambda^{\ell}_{\mathbf{y}^f|\mathbf{d}_{\text{obs}}}$, as

$$\lambda^{k}_{\mathbf{y}^f|\mathbf{d}_{\text{obs}}} = \frac{\widehat{\pi}_k N(\mathbf{d}_{\text{obs}}; \mathbf{H}\widehat{\boldsymbol{\mu}}^k_{\mathbf{y}^f}, \mathbf{H}\widehat{\mathbf{C}}^k_{\mathbf{y}^f}\mathbf{H}^T + \mathbf{C}_{\boldsymbol{\varepsilon}})}{\sum_{\ell=1}^{N_c} \widehat{\pi}_\ell N(\mathbf{d}_{\text{obs}}; \mathbf{H}\widehat{\boldsymbol{\mu}}^{\ell}_{\mathbf{y}^f}, \mathbf{H}\widehat{\mathbf{C}}^{\ell}_{\mathbf{y}^f}\mathbf{H}^T + \mathbf{C}_{\varepsilon})}.$$

(2) Loop on ensemble members: for each $\mathbf{y}^f_j$ with $j = 1, 2, \ldots, N_e$:

(2.a) set k as the known component of the member $\mathbf{y}^f_j$;

(2.b) generate a random index of new component $\ell \in \{1, 2, \ldots, N_c\}$ according to the discrete distribution given by $\{\lambda^1_{\mathbf{y}^f|\mathbf{d}_{\text{obs}}}, \lambda^2_{\mathbf{y}^f|\mathbf{d}_{\text{obs}}}, \ldots, \lambda^{N_c}_{\mathbf{y}^f|\mathbf{d}_{\text{obs}}}\}$;

(2.c) compute the auxiliary vector $\mathbf{y}^{f'}_j$ according to

$$\mathbf{y}^{f'}_j = \boldsymbol{\mu}^{\ell}_{\mathbf{y}^f} + L^{\ell}(L^k)^{-1}(\mathbf{y}^f_j - \boldsymbol{\mu}^k_{\mathbf{y}^f})$$

where $L^{\ell}(L^{\ell})^T = \widehat{\mathbf{C}}^{\ell}_{\mathbf{y}^f}$ and $L^k(L^k)^T = \widehat{\mathbf{C}}^k_{\mathbf{y}^f}$;

(2.d) compute the updated vector $\mathbf{y}^u_j$ using the updating equation for the component ℓ on the auxiliary vector $\mathbf{y}^{f'}_j$:

$$\mathbf{y}^u_j = \mathbf{y}^{f'}_j + \widehat{\mathbf{K}}^{\ell}_{\mathbf{H}}(\mathbf{d}_j - \mathbf{H}\mathbf{y}^{f'}_j)$$

where $\widehat{\mathbf{K}}^{\ell}_{\mathbf{H}} = \widehat{\mathbf{C}}^{\ell}_{\mathbf{y}^f}\mathbf{H}^T(\mathbf{H}\widehat{\mathbf{C}}^{\ell}_{\mathbf{y}^f}\mathbf{H}^T + \mathbf{C}_{\boldsymbol{\varepsilon}})^{-1}$.

The validity of the updating scheme can be proved (see the Appendix for the details) using, as in the ensemble Kalman Filter, the notion of *finite ensemble representation* extended to GMM. A set of vectors $\{\mathbf{y}^f_j\}_{j=1}^{N_e}$ is considered a *finite ensemble representation* with precision η of the random GM vector $\mathbf{y}^f$ when there exists a partition of the set of indexes $J = \{1, 2, \ldots, N_e\}$, $J = I_1 \cup I_2 \cup \ldots \cup I_{N_c}$ such that $\{\mathbf{y}^f_{i_k}\}$, $i_k \in I_k$, is a *finite ensemble representation* of each Gaussian component k and

$$\mid \pi_k - \frac{n_k}{N_e} \mid \leq \eta$$

for some η, where n_k is the number of elements of the set I_k for every k. The application of the methodology is shown to Section 4 using Lorenz model as validation case.

Real filtering problems are typically characterized by high-dimensional systems where the state variable $\mathbf{y}$ contains a large number of parameters. Reservoir models typically require state systems with $10^5 - 10^6$ variables. This restriction poses two obstacles in the numerical implementation of the proposed method for large scale applications. The first problem is the application of the EM algorithm to the forecasted ensemble, which requires the evaluation of multivariate Gaussian PDFs on each state variables $\{\mathbf{y}^f_j\}_{j=1}^{N_e}$ and it becomes unfeasible when they contain a large number of parameters. Arithmetic underflow problems may occur with a parameter size greater than 100. Moreover the same problem arises in computing the updating weights, defined in step (1), in the case of high dimensions data space. The second problem is due to the covariance matrices factorization by Cholesky decomposition used in step (2.c), that cannot be addressed directly in a high-dimensional space.

The solution we adopted to face these problems is the modification of the proposed algorithm assuming the hypothesis of localized correlation as discussed in [8] and [3]. This alternative is equivalent to use only a subset of the model state variables belonging to a neighborhood of the observation to define the conditional distribution and to perform the updating. In our case, this localization have two advantages. Firstly, it overcomes the numerical issues in applying the EM and computing the covariance factorizations. Second, avoiding a unique global GM updating, it allows to get more flexibility in reproducing the local system multimodality through the updated weights. In this approximated form, the method can be also applied to update spatial correlated random field. An example of this approach to a single phase reservoir data assimilation problem is given in Section 5. Here, the EnKF-GMM is coupled with a local approximation to update a bimodal distribution of log-permeability on a 2D grid.

4 An example with Lorenz model

In order to show the methodology introduced in the previous Section 3.2, we propose an application to the well-known Lorenz model. A scalar example of the algorithm is given in [6]. The Lorenz model [19] is an example of chaotic and strongly nonlinear system. In the field of data assimilation, it became a benchmark for testing different algorithms on strongly nonlinear dynamics, see for example [3, 9] and [24]. However, according to our perspective, this model is viewed as source of multimodality and it is used here to illustrate the behavior of the EnKF-GMM. The results are evaluated with the aid of a reference posterior distribution obtained with a Monte Carlo method built with a large number of samples and without any parametric assumption. Finally we provide a comparison with standard EnKF results.

For all the cases treated, we performed a single step data assimilation. The forward model equations are given by the system of three coupled and nonlinear differential equations [19]:

$$\frac{dx}{dt} = \gamma(y - x), \quad \frac{dy}{dt} = \rho x - y - xz,$$

$$\frac{dz}{dt} = xy - \beta z,$$

with the commonly used coefficients $\gamma = 10$, $\rho = 28$ and $\beta = 8/3$, as in [9] and [24]. We assimilated data using three methods: EnKF, EnKF-GMM and a large scale Monte Carlo method. For all the three methods, an ensemble of initial model states was generated by adding Gaussian distributed noise with zero mean and variance equal to 1 to initial conditions given by

$$(x_0, y_0, z_0) = (1.508870, -1.531271, 25.46071).$$

The initial ensemble was propagated in time using numerical solution of Lorenz system and it was used to approximate the forecast distribution before the assimilation. An ensemble of 1,000 members was simulated for the EnKF and EnKF-GMM, while a large set of points (32, 000) was used for the nonparametric Monte Carlo approach. In all the tests, data were assimilated assuming a diagonal error covariance matrix with variances equal to 40 for all the three model parameters. Data assimilation was performed according to the three methods and the corresponding final posterior densities were estimated. In particular, for the standard EnKF we assimilated data using the updating Eq. 5 after estimating Gaussian prior statistics by means of Eqs. 2 and 3. The same equations have been used to estimate the posterior Gaussian density from the updated ensemble. In the case of EnKF-GMM, a GM prior model with two components is assumed and its parameters have been estimated using the EM algorithm. We chose this number of components to show the improvement respect to standard EnKF even with a low number of mixture components. By increasing this number, the flexibility of the model increases and better results can be obtained. However in this case we want to see if the algorithm can capture the system multimodality. We updated the ensemble using the procedure explained in Section 3.2 and we applied EM algorithm to the updated ensemble to provide the posterior GM density parameters. The full Monte Carlo solution consists of calculating the Bayes posterior density function as the product between the density (histogram) estimated from the forecasted ensemble and the Gaussian likelihood function on the three dimensional model parameters:

$$f(\mathbf{y}|\mathbf{d}) \propto f(\mathbf{d}|\mathbf{y})\, f(\mathbf{y})$$

with $\mathbf{y} = (x, y, z)$ and $\mathbf{d} = (d_x^0, d_y^0, d_y^0)$ which are the observed data.

The chaotic behavior of Lorenz model is well known: a set of points can be rapidly scattered across the two lobes, see Fig. 1. Even in a short time, Lorenz dynamics lead to non-Gaussian forecast distributions. Therefore the experiment was repeated with three different time steps, $t_1 = 0.2$, $t_2 = 0.3$, and $t_3 = 0.4$ assimilating data $\mathbf{d}_1 = (-5.5, -10, 11.5)$, $\mathbf{d}_2 = (-2.2, -3.9, 11.9)$ and $\mathbf{d}_3 = (0, 0, 15)$ respectively. The chosen time intervals give a range of conditions that allow the model to show different behaviors. For each experiment, we compared the three estimated posterior distributions

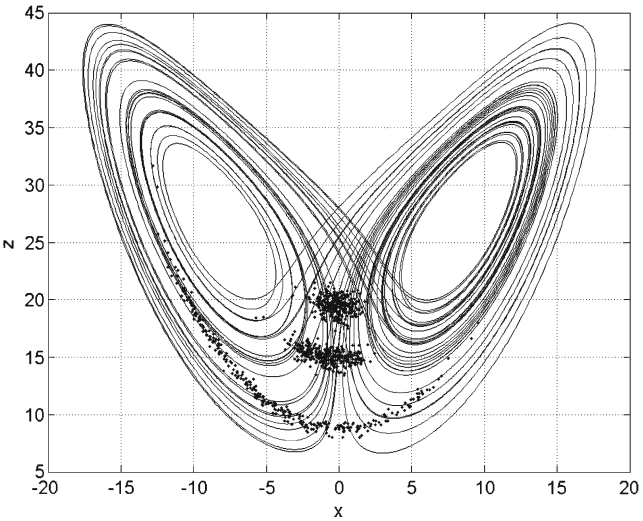

Fig. 1 The Lorenz attractor along plane xz. An ensemble of 300 points have been propagated till times 0.2, 0.3, 0.4

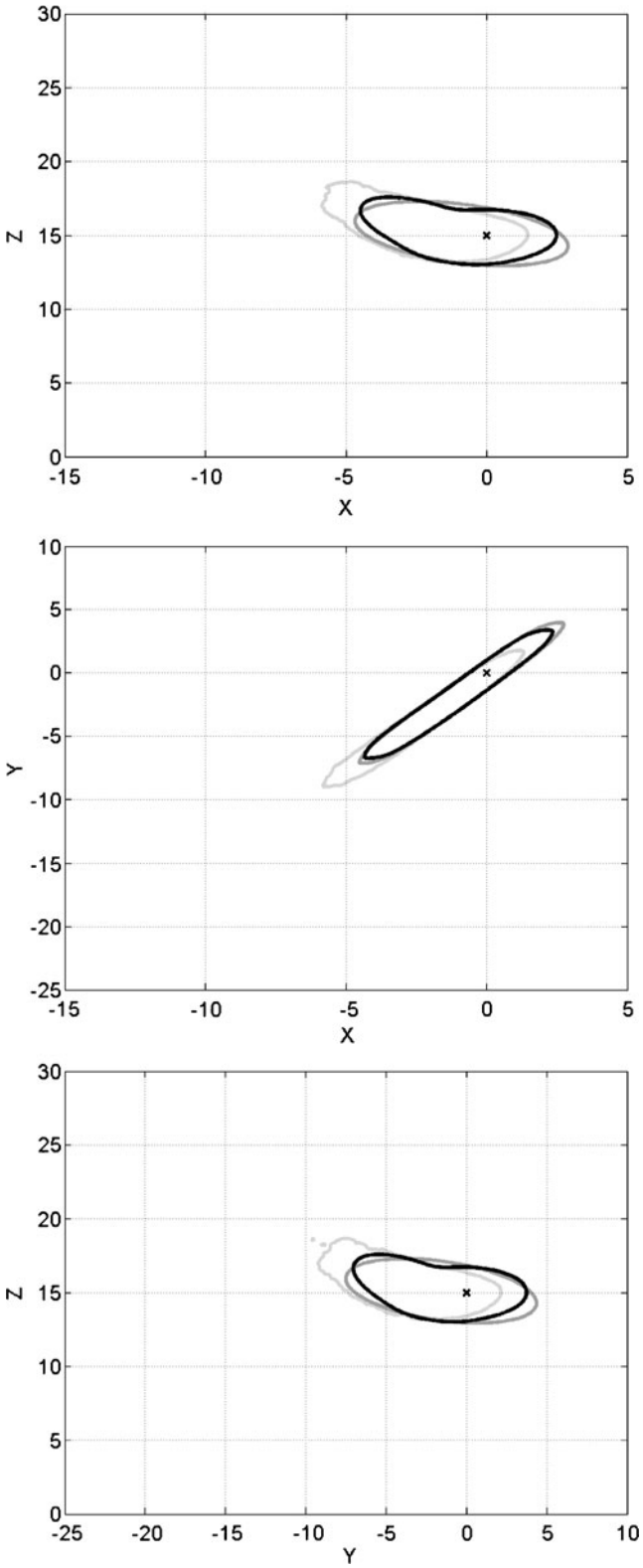

Fig. 2 0.99 confidence region of the 2D posterior marginal densities on plane xz, xy, yz (from *top* to *bottom*). Estimation at time step $t = 0.2$ integrating data $\mathbf{d}_1$ (*cross*) using full Monte Carlo method (*light grey*), EnKF (*grey*), and EnKF-GMM (*black*)

plotting the 0.99 confidence region of the 2D marginal densities.

Initially, at time $t_1 = 0.2$, the model is not far from linear and Gaussian model. In this case, as shown in Fig. 2, the solutions provided by EnKF and EnKF-GMM are very similar and they both have a reasonable match with the reference PDF obtained with the full Monte Carlo method. This result shows that when the prior model is well described by a single Gaussian PDF, the solution obtained by assuming a prior GMM is close to the EnKF one. At time step $t_2 = 0.3$, the model dynamic starts to produce a more complex forecast distribution and the nonparametric large sample Monte Carlo posterior distribution is non-Gaussian as it can be seen especially from the xz and yz 2D marginals (see Fig. 3). In this case, it is clear that a GM prior can better capture the behavior of the system than the EnKF solution. Along the plane xy, as shown in Fig. 3, the reference solution is approximately Gaussian and also in this case, the GMM posterior is similar to the Gaussian posterior of the EnKF solution. At time step $t_2 = 0.4$ the model dynamics are definitely non-Gaussian and in Figs. 4 and 5 we can see the clear advantages of the EnKF-GMM respect to EnKF. In fact the distribution obtained with the former is close to the reference, while the one obtained with the latter overestimates the uncertainty.

5 An example with a 2D single phase reservoir

In this section we test the EnKF-GMM methodology on a 2D synthetic reservoir example. The model considered is single phase slightly compressible fluid defined on a grid with 31×31 grid blocks and cell sizes of $100 \times 100 \times 50$ m. The geological model of the reservoir is described by a distribution of two facies corresponding to different petrophysical properties: one with low permeability and the other with higher permeability values. The porosity is considered constant. Within each facies, log-permeability is distributed according to a Gaussian PDF and the univariate distribution adopted for the geostatical modeling of the log-permeability field is a Gaussian mixture.

The model is based on a five-spot configuration: one injector and four producers. The injector is located at the center of the grid with a constant pressure equal to 105 bar. The four producers are located at the grid corners with a pressure of 90 bar. The initial pressure is constant and equal to 100 bar.

The example is organized as a data assimilation test followed by forecast. A realization of facies and a log-permeability grid has been chosen as "true" model

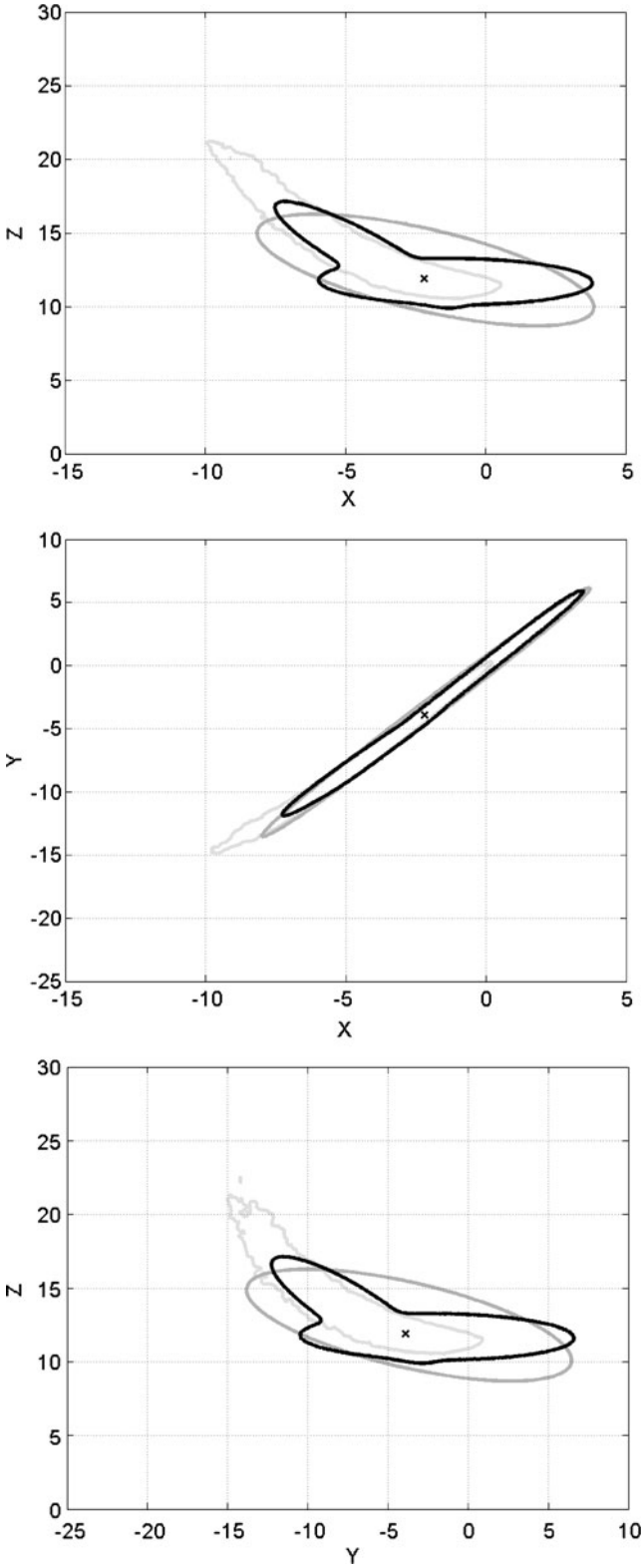

Fig. 3 0.99 confidence region of 2D the posterior marginal densities on plane xz, xy, yz (from *top* to *bottom*). Estimation at time step $t = 0.3$ integrating data $\mathbf{d}_2$ (*cross*) using full Monte Carlo method (*light grey*), EnKF (*grey*) and EnKF-GMM (*black*)

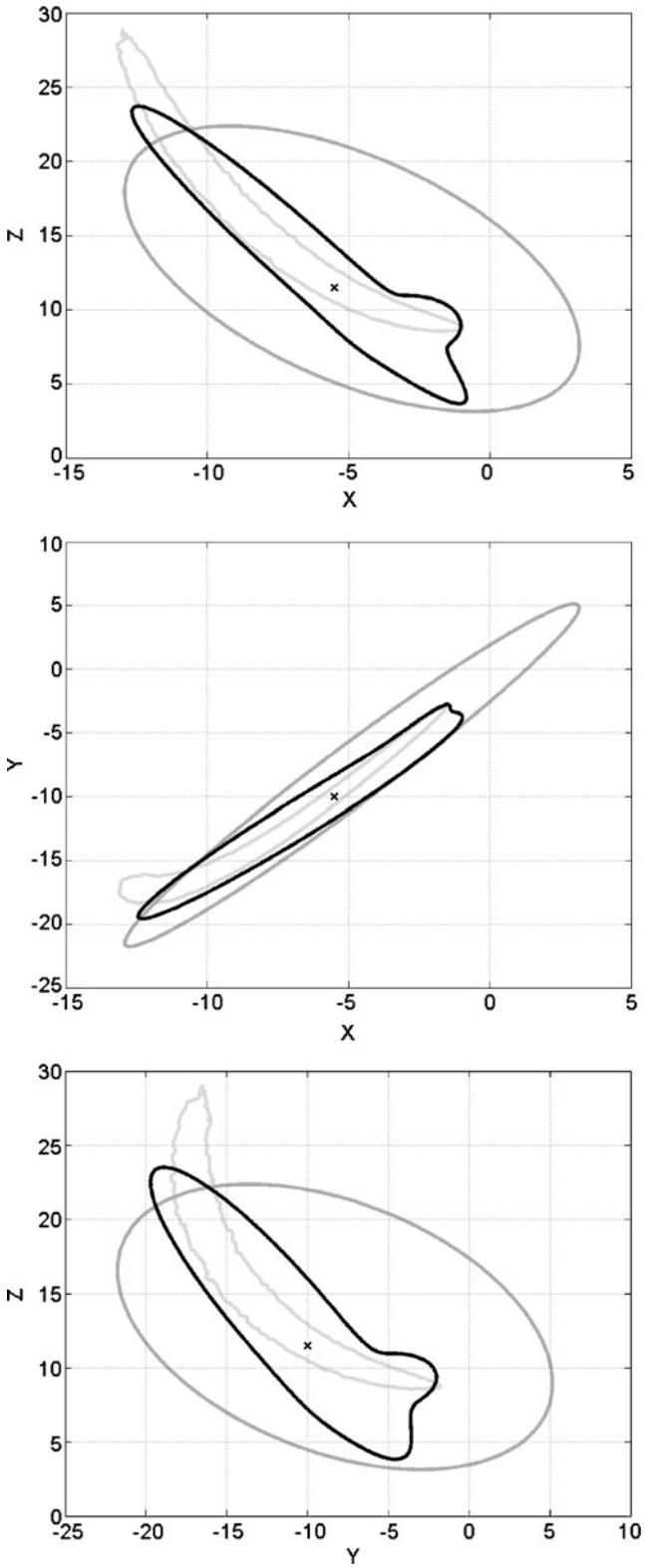

Fig. 4 0.99 confidence region of the 2D posterior marginal densities on plane xz, xy, yz (from *top* to *bottom*). Estimation at time step $t = 0.4$ integrating data $\mathbf{d}_3$ (*cross*) using full Monte Carlo method (*light grey*), EnKF (*grey*) and EnKF-GMM (*black*)

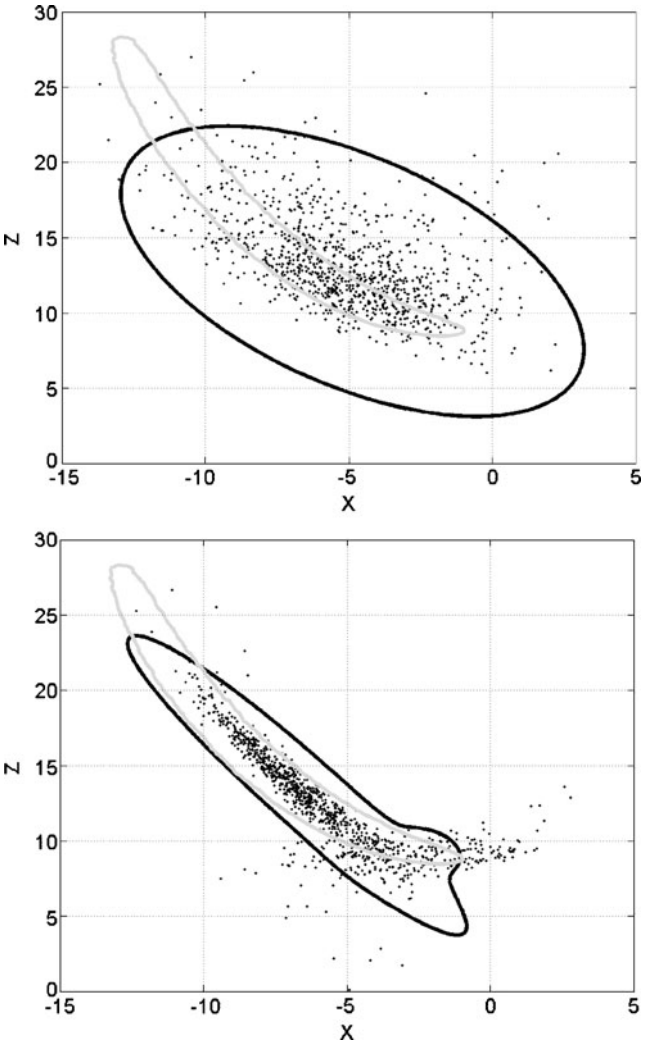

Fig. 5 Updated ensemble (*points*) and 2D marginal density confidence region 0.99 (*black*) on plane *xz* using EnKF (*top*) and EnKF-GMM (*bottom*). Estimation at time step $t = 0.4$. 0.99 confidence region from full Monte Carlo method plotted in *light grey*

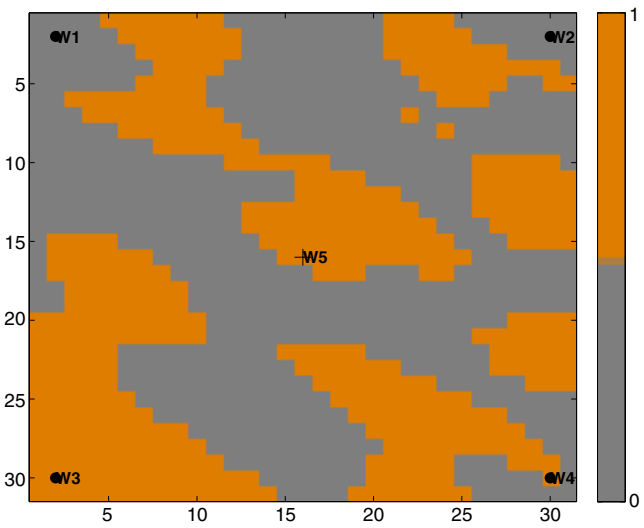

Fig. 6 Reference facies distribution. *Yellow* facies (component) has good permeability while *grey* facies has worse characteristics. The *black dots* are the producers and the *cross* is the injector well

and a single phase simulation is carried out to obtain the production data to be used as history matching measurements. The sequential data assimilation has been performed using both the standard EnKF and the EnKF-GMM (see Section 3) to update the log-permeability grid. The two ensembles of updated permeabilities have been used to simulate the production forecast with an additional well. The results have been compared with the "true" production of the reference model.

5.1 The reference model

We assume that the spatial distribution of log-permeability grids is essentially controlled by the spatial distribution of two facies. In agreement with our hypothesis on mixture model, the facies are identified with the two components of the mixture. We created the "true" model by means of random realization of a discrete variable (facies 1) using a sequential indicator simulation (SISIM, [14]) with an anisotropic spherical variogram (range of 250 m), anisotropy ratio of 0.5 and direction of anisotropy $\theta = 30°$ from north-west to south-east. We assigned a global proportions of the facies equal to $[0.54, 0.46]$. The choice of these parameters has been done in order to obtain a "true" model with a realistic spatial distribution of facies

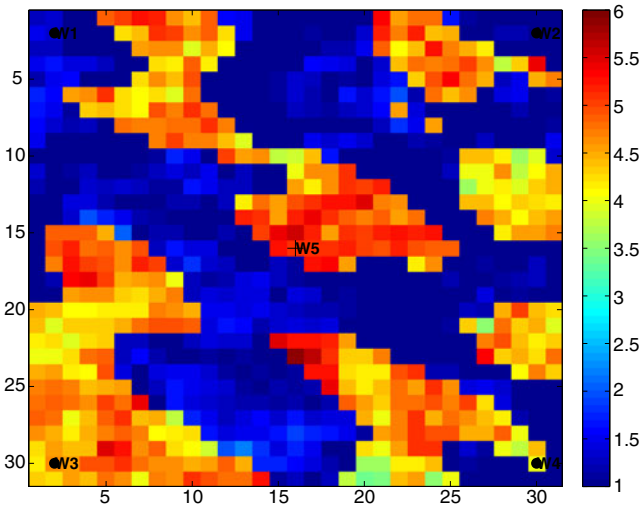

Fig. 7 Reference log-permeability distribution. The *black dots* are the producers and the *cross* is the injector well

shapes. Then, we distributed within each facies the log-permeability using the sequential gaussian simulation (SGSIM, [14]). The parameters required in this case are $\mu_1 = 1.0$ and $\sigma_1 = 0.39$ with isotropic spherical variogram and range equal to 250 m for facies 1 and $\mu_2 = 4.7$ and $\sigma_2 = 0.45$ again with isotropic spherical variogram and the same range for facies 2.

Flow simulation was carried out for 180 days with step length equal to 1 day. Flow rates for injector and producers were measured at each time step with additional independent measurement error. The measurement error is Gaussian distributed with mean equal to 0 and standard deviation equal to 10 m^3/ day. The reference facies grid is shown on Fig. 6 and the reference log-permeability is given on Fig. 7. Well locations are depicted in the same figures. The univariate distribution used within SISIM and SGSIM algorithms for the generation of the reference model is shown on Fig. 8: the two modes corresponding to the two mixture components are well separated. Figure 9 shows the measured flow rates for the five wells (without measurement error). The dynamic behavior of each well depends on the spatial distribution of facies. In fact the two producers in the "good" facies (bottom left and bottom right) have higher flow rates.

5.2 Data assimilation experiment and results

Data assimilation is performed integrating measurements every 30 days with a total of 6 updating steps. Both in the EnKF case and in the EnKF-GMM case, an initial ensemble of size $N_e = 100$ has been generated.

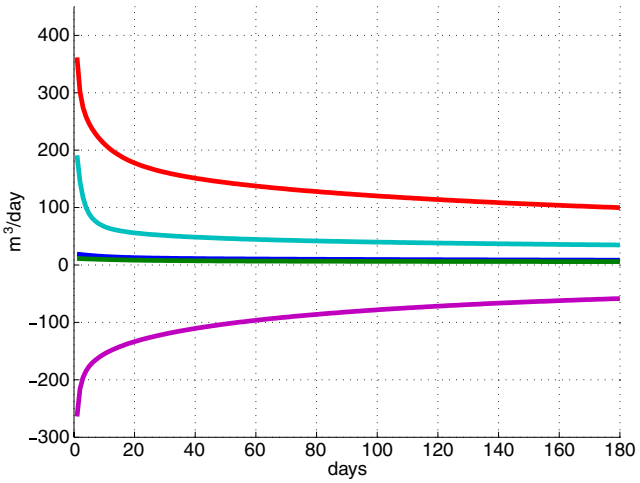

Fig. 9 Reference production curves for all the wells. *Blue* is the top left well, *green* is the top right well, *red* is bottom left well, *light blue* is bottom right well. *Magenta* is the injector well

For the EnKF, distance based localization has been adopted with a tapering function as in [11]. A local updating combined with a random regeneration of the log-permeability values has been implemented in the EnKF-GMM case. More in details, for each well a two levels neighborhood has been considered. Permeability and facies are updated with EnKF-GMM scheme in the inner neighborhood (disk shaped) around each single well. In the outer neighborhood (ring-shaped), permeability and facies are resampled, using sequential indicator and gaussian simulation, respectively. The re-simulated values are conditioned both to the updated log-permeability and facies values (inner side) and to the prior ensemble values (outer side). Far from the

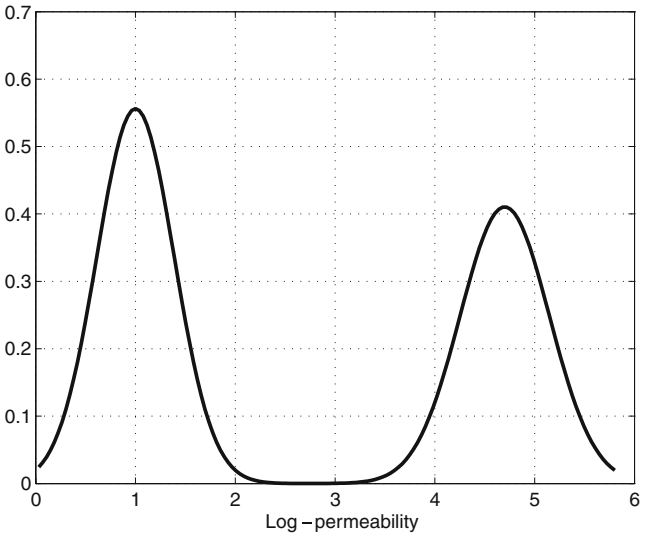

Fig. 8 Marginal probability distribution of log-permeability of the reference model

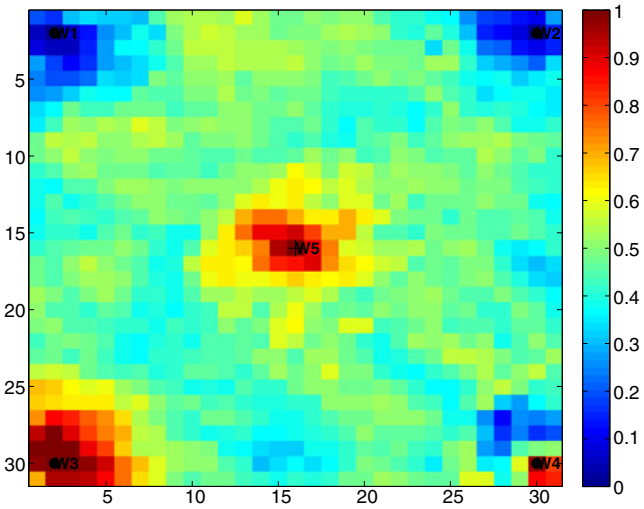

Fig. 10 Probability of facies 2 after facies updating performed with EnKF-GMM

wells, outside the outer neighborhood, the prior values are maintained. The EM algorithm is applied for each local update around the well separately. This approach allows to properly update facies and log-permeability where production data are informative and to preserve the global spatial correlation. The final output of the algorithm is an updated ensemble of log-permeability grids in association with a correspondent ensemble of facies realizations.

Figure 10 shows the facies probability distribution from the final updated ensemble obtained with the EnKF-GMM after six time steps. Comparing this probability with the reference facies map, it is possible to see that the facies around the wells are correctly identified. There is an underestimation around the lower right producer, well number 4, probably due to the small continuity of the facies in the region and consequently to the low production of that area. A similar information is provided with the log-permeability mean of the final ensemble obtained with the EnKF shown on Fig. 11. Also for the EnKF, the high and low log-permeability areas are correctly identified. However, looking at a single realization of the ensemble, see Fig. 12, it can be observed that the bimodal character of the facies is not maintained. In the lower left corner, around the well number 3, the shapes linked to the original two facies are only apparent. In the reality, the update produces a general increase of log-permeabilities toward very high values that does not preserve the bimodality. This behavior is more evident comparing Figs. 13 and 14. In these figures we show the histogram of the updated log-permeability values taken from a neighborhood of 5×5 cells around well 3 over the ensemble members.

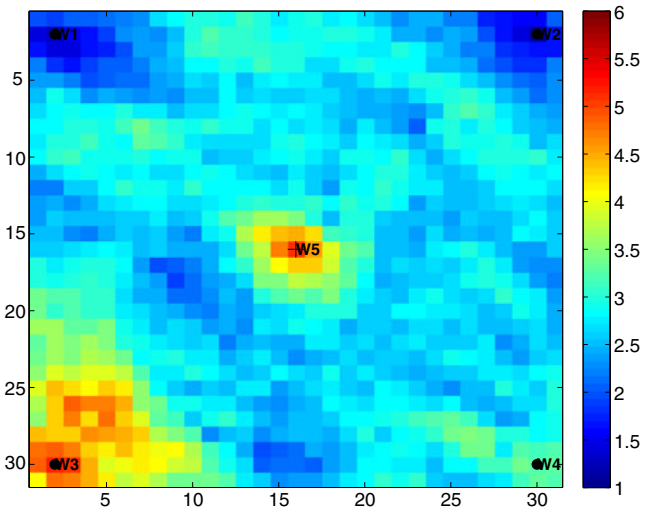

Fig. 11 Log-permeability mean obtained with EnKF

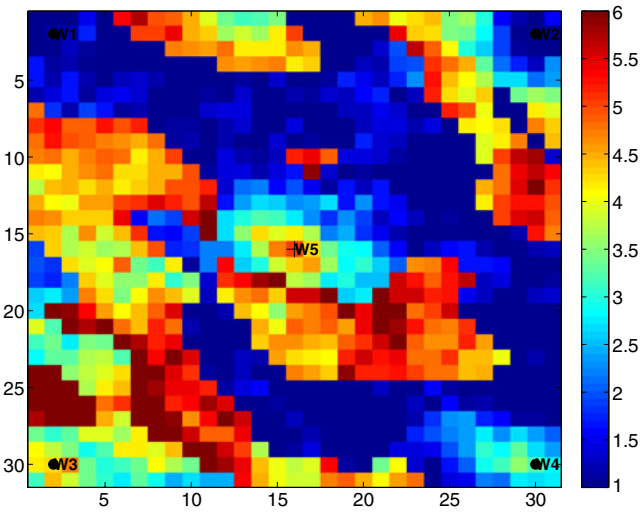

Fig. 12 A single log-permeability realization of the updated ensemble with EnKF

Figure 13 is obtained from the ensemble updated with the EnKF-GMM, Fig. 14 from the ensemble updated with the EnKF. The reason for choosing the 5×5 neighborhood is to analyze the bimodal character of the spatial histogram at the size of the neighborhood used for the update around the conditioning well 3. The red line corresponds to the reference log-permeability value at well location. It can be observed that the mixture update with EnKF-GMM preserves the bimodal character of the log-permeability distribution due to facies and one of the two mixture modes is centered on the "true" value. With the EnKF, the histogram is

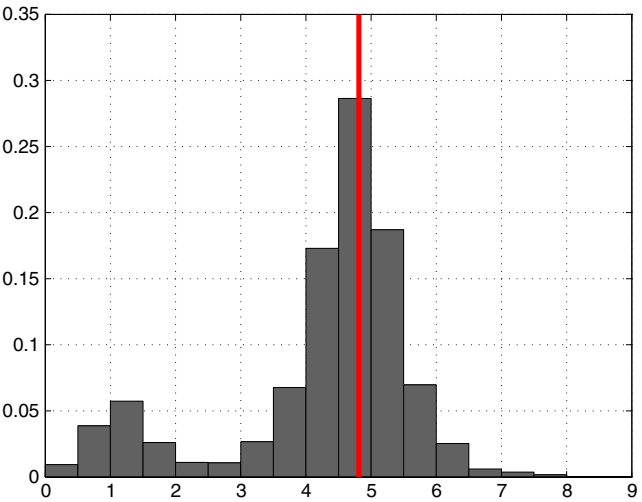

Fig. 13 Histogram of the EnKF-GMM updated permeability in a 5×5 neighborhood of the well 3 located in the *lower left* corner of the grid. The *red line* corresponds to the reference log-permeability value at well location

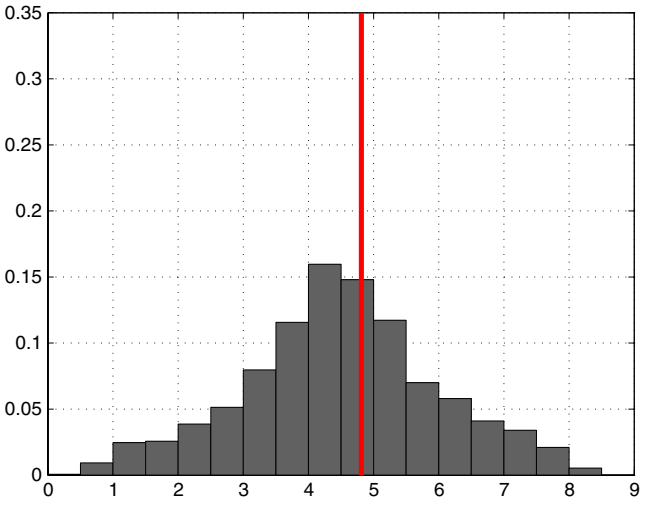

Fig. 14 Histogram of the EnKF updated permeability in a 5 × 5 neighborhood of the well 3 located in the *lower left* corner of the grid. The *red line* corresponds to the reference log-permeability value at well location

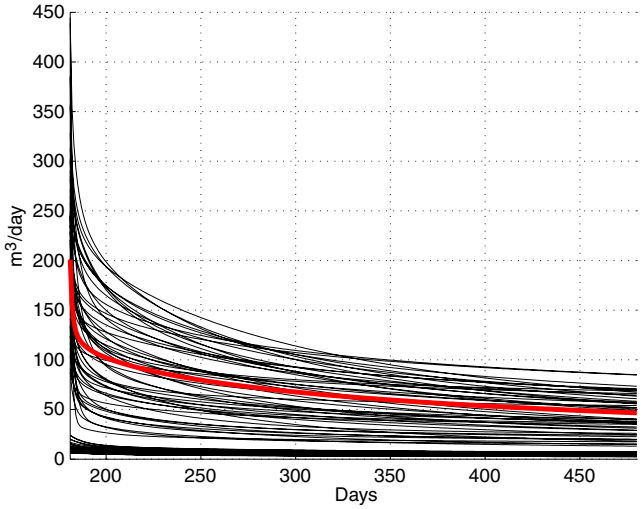

Fig. 16 Ensemble production forecast (*black curves*) of the additional well 6 with EnKF-GMM. The *red curve* is the reference production. Production is given in cubic meters. The *horizontal axis* is time in days

no more bimodal and the link between log-permeability and facies is lost. Note also the higher right tail of the distribution.

5.3 Forecast experiment and results

In order to evaluate the effect of the two different updates, we performed a forecast simulation with an additional well. The well is added close to the area that both EnKF-GMM (by means of the probability of facies map) and EnKF (by means of the log-permeability updated mean) identify as the best region. Figure 15 shows the location of the new well number 6. The forecast simulation is performed on the final updated ensemble obtained with EnKF and EnKF-GMM from time zero over a time interval of 300 days after the first 180 days of history matching. Figures 16 and 17 show the production forecast results for the EnKF-GMM and the EnKF, respectively. The two forecasts are compared with the reference production obtained with the "true" model. We can observe that both the

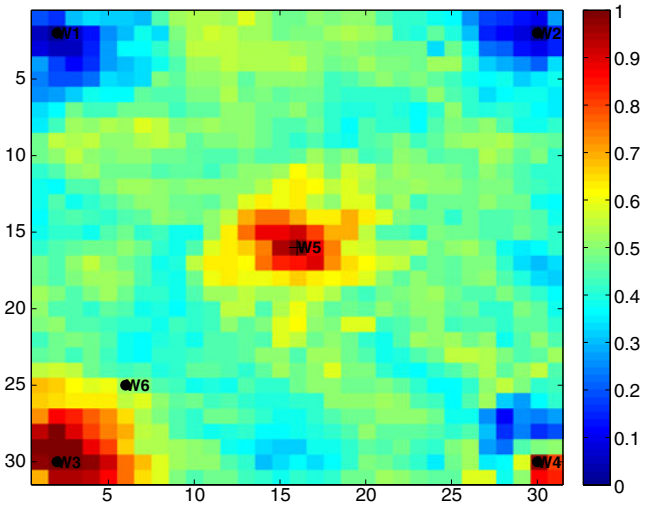

Fig. 15 Additional well location for the evaluation of forecast (*blue dot*). The color coded variable is again the probability of facies 2 from EnKF-GMM update

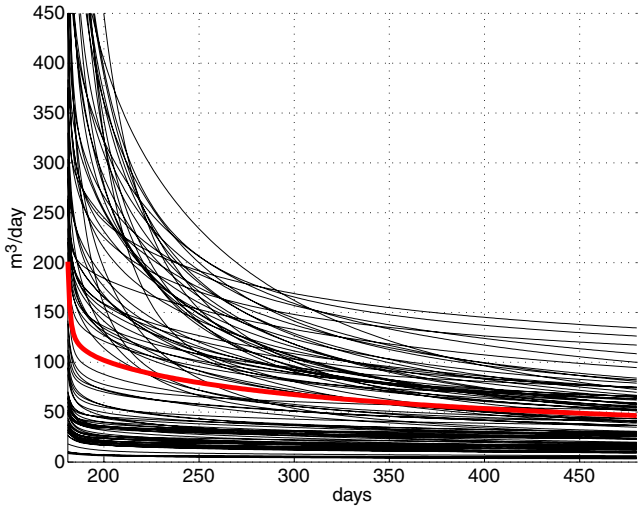

Fig. 17 Ensemble production forecast (*blue curves*) of the additional well 6 with EnKF. The *red curve* is the reference production. Production is given in cubic meters. The *horizontal axis* is time in days

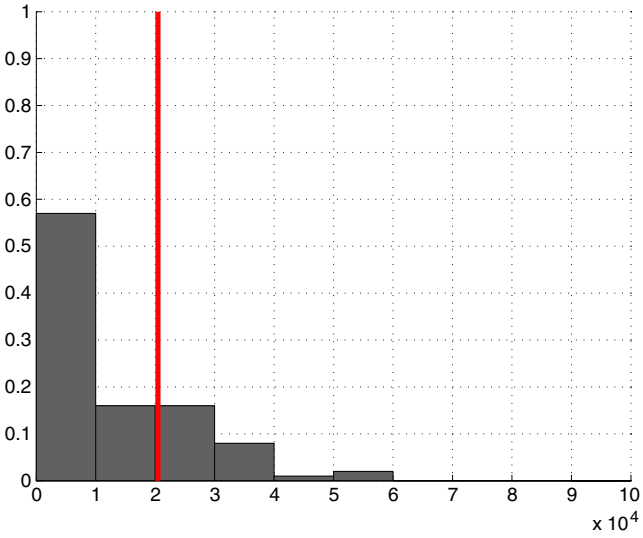

Fig. 18 Histogram of total production with EnKF-GMM. The *red curve* is the reference production. Data are given in cubic meters

assimilation schemes give a forecast uncertainty interval that contains the reference curve, but the EnKF overestimates the uncertainty. The same effect can be observed also in terms of total production forecast. Figures 18 and 19 show the histogram of the total production obtained with the EnKF-GMM and with the EnKF. In the EnKF case, the overestimation of log-permeability values gives also an higher probability to obtain very large production rate far from the "true" model.

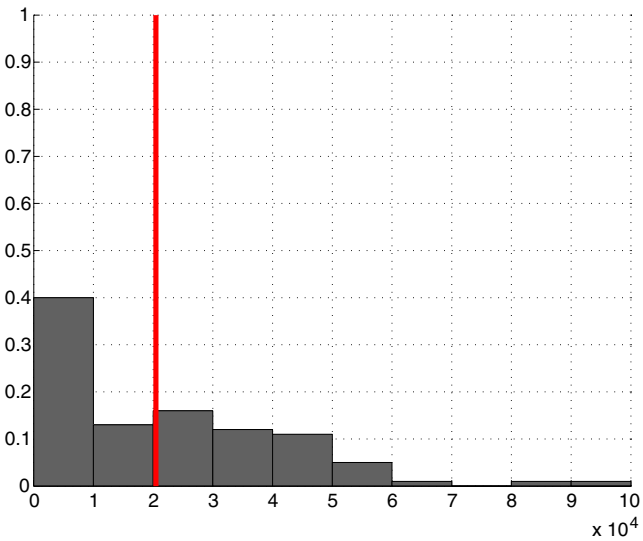

Fig. 19 Histogram of total production with EnKF. The *red curve* is the reference production. Data are given in cubic meters

6 Conclusions

In this work we presented a new ensemble data assimilation method to update systems characterized by multimodal distributions. The method is based on the ensemble Kalman filter and it makes use of Gaussian mixture models to describe multimodal priors and posteriors. Gaussian mixture models are a parametric model with an explicit analytical solution for the posterior probabilities in the case of linear models, linear measurements and gaussian likelihoods. We reformulated EnKF update equations assuming GM priors with the objective to provide a novel sequential updating technique able to reproduce multimodal behaviors with improved flexibility. As in the EnKF, forecast and updated PDFs are represented by an ensemble of state variables that is recursively used to retrieve the required statistics. For this purpose, the proposed method involves the use of the expectation–maximization algorithm. The EM algorithm allows to estimate the parameters of the Gaussian mixture PDF used to approximate the propagated forecasted ensemble at each assimilation step. Once the forecasted PDF parameters are estimated, each ensemble member is updated according to the proposed scheme. We introduced the *finite ensemble representation* and proved that the updated ensemble is a correct sample of the posterior Gaussian mixture distribution assuming linear model, linear measurements and Gaussian likelihood. The approach is consistent with the one given by the EnKF when the mixture is made by a single component.

As first example, the method has been tested on the Lorenz model. The results have been compared to those provided by the EnKF respect to a reference solution based on a nonparametric Monte Carlo Bayesian data assimilation with a large number of samples. The example shows that the method estimates better the posterior distribution and the uncertainty.

In the second example, the EnKF-GMM with localization is applied to a 2D single phase reservoir problem. The example gives evidence of the applicability of the method also to problems closer to typical spatial geosciences data assimilation context. In this case, the key point is the identification of the mixture components with the facies with different distributions of petrophysical properties. The obtained results allows to conclude that the EnKF-GMM compared with EnKF gives a better evaluation of posterior distribution and of uncertainty of the forecast production. The improvement is mainly due to the more accurate modeling of log-permeability multimodality.

Acknowledgements We would like to thank Alberto Cominelli and Dario Grana for helpful comments and discussions. We would also like to acknowledge eni exploration & production division for the permission to publish this paper.

Appendix

A.1 Finite ensemble representation of linear transformed vectors

The distribution of the linear transformed random vector $\mathbf{z} = L\mathbf{y}$ is Gaussian with mean $\boldsymbol{\mu}_\mathbf{z} = L\boldsymbol{\mu}_\mathbf{y}$ and covariance $C_\mathbf{z} = LC_\mathbf{y}L^T$ (see for example [22]). To prove that the set of vectors $\{\mathbf{z}_j\}_{j=1}^{N_e}$ is a *finite ensemble representation* of the linear transformed vector $\mathbf{z}$ we have to prove that

$$\|\frac{1}{N_e}\sum_{j=1}^{N_e}\mathbf{z}_j - L\boldsymbol{\mu}_\mathbf{y}\| \leq \eta_1$$

$$\|\frac{1}{N_e}\sum_{j=1}^{N_e}\left(\mathbf{z}_j - \boldsymbol{\mu}_\mathbf{z}\right)\left(\mathbf{z}_j - \boldsymbol{\mu}_\mathbf{y}\right)^T - LC_\mathbf{y}L^T\| \leq \eta_2.$$

for some precision (η_1, η_2). For the mean

$$\begin{aligned}\|\frac{1}{N_e}\sum_{j=1}^{N_e}\mathbf{z}_j - \boldsymbol{\mu}_\mathbf{z}\| &= \|L(\frac{1}{N_e}\sum_{j=1}^{N_e}\mathbf{y}_j - \boldsymbol{\mu}_\mathbf{y})\| \\ &\leq \|L\| \cdot \|\frac{1}{N_e}\sum_{j=1}^{N_e}\mathbf{y}_j - \boldsymbol{\mu}_\mathbf{y}\| \\ &\leq \|L\| \cdot \eta_{\boldsymbol{\mu}} = \eta_1.\end{aligned}$$

In a similar way for the covariance

$$\begin{aligned}&\|\frac{1}{N_e}\sum_{j=1}^{N_e}\left(\mathbf{z}_j - \boldsymbol{\mu}_\mathbf{z}\right)\left(\mathbf{z}_j - \boldsymbol{\mu}_\mathbf{y}\right)^T - C_\mathbf{z}\| \\ &= \|\frac{1}{N_e}\sum_{j=1}^{N_e}\left(L\mathbf{y}_j - L\boldsymbol{\mu}_\mathbf{y}\right)\left(L\mathbf{y}_j - L\boldsymbol{\mu}_\mathbf{y}\right)^T - LC_\mathbf{y}L^T\| \\ &= \|L\left(\frac{1}{N_e}\sum_{j=1}^{N_e}\left(\mathbf{y}_j - \boldsymbol{\mu}_\mathbf{y}\right)\left(\mathbf{y}_j - \boldsymbol{\mu}_\mathbf{y}\right)^T\right)L^T - LC_\mathbf{y}L^T\| \\ &\leq \|L\| \cdot \|\frac{1}{N_e}\sum_{j=1}^{N_e}\left(\mathbf{y}_j - \boldsymbol{\mu}_\mathbf{y}\right)\left(\mathbf{y}_j - \boldsymbol{\mu}_\mathbf{y}\right)^T - C_\mathbf{y}\| \cdot \|L^T\| \\ &\leq \|L\|^2 \cdot \eta_\mathbf{C} = \eta_2.\end{aligned}$$

A.2 Finite ensemble representation and ensemble Kalman filter

Firstly, we observe that the mean and the covariance of the conditional distribution of $\mathbf{y}^f|\mathbf{d}_{\text{obs}}$ are

$$\begin{aligned}\boldsymbol{\mu}_{\mathbf{y}^f|\mathbf{d}_{\text{obs}}} &= \boldsymbol{\mu}_{\mathbf{y}^f} + \mathbf{K_H}(\mathbf{d}_{\text{obs}} - \mathbf{H}\boldsymbol{\mu}_{\mathbf{y}^f}) \\ \mathbf{C}_{\mathbf{y}^f|\mathbf{d}_{\text{obs}}} &= \mathbf{C}_{\mathbf{y}^f} - \mathbf{K_H}\mathbf{H}\mathbf{C}_{\mathbf{y}^f}\end{aligned}$$

with, as usual, the Kalman gain matrix given by

$$\mathbf{K_H} = \mathbf{C}_{\mathbf{y}^f}\mathbf{H}^T(\mathbf{H}\mathbf{C}_{\mathbf{y}^f}\mathbf{H}^T + \mathbf{C}_{\boldsymbol{\varepsilon}})^{-1}.$$

The joint random vector $(\mathbf{y}^f, \mathbf{d}_{\text{obs}} + \boldsymbol{\varepsilon})$ is Gaussian distributed with mean and covariance given by

$$\boldsymbol{\mu}_{\mathbf{y}^f,\mathbf{d}} = \begin{bmatrix}\boldsymbol{\mu}_{\mathbf{y}^f} \\ \mathbf{d}_{\text{obs}}\end{bmatrix}, \quad \mathbf{C}_{\mathbf{y}^f,\mathbf{d}} = \begin{bmatrix}\mathbf{C}_{\mathbf{y}^f} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{\boldsymbol{\varepsilon}}\end{bmatrix}$$

and the joint set of vectors $\{(\mathbf{y}_j^f, \mathbf{d}_j)\}_{j=1}^{N_e}$ gives a *finite ensemble representation* of $(\mathbf{y}^f, \mathbf{d}_{\text{obs}} + \boldsymbol{\varepsilon})$.
Considering the linear operator

$$L = \left[\mathbf{1}_M - \mathbf{K_H}\mathbf{H}\ \mathbf{K_H}\right]$$

we have

$$\begin{aligned}\mathbf{y}_j^u = L\begin{bmatrix}\mathbf{y}_j^f \\ \mathbf{d}_j\end{bmatrix} &= \left[\mathbf{1}_M - \mathbf{K_H}\mathbf{H}\ \mathbf{K_H}\right]\cdot\begin{bmatrix}\mathbf{y}_j^f \\ \mathbf{d}_j\end{bmatrix} = \\ &= \mathbf{1}_M\mathbf{y}_j^f - \mathbf{K_H}\mathbf{H}\mathbf{y}_j^f + \mathbf{K_H}\mathbf{d}_j = \mathbf{y}_j^f + \mathbf{K_H}(\mathbf{d}_j - \mathbf{H}\mathbf{y}_j^f)\end{aligned}$$

where $j = 1, 2, \ldots, N_e$.
The inequalities given in the linear case allow to conclude that $\{\mathbf{y}_j^u\}_{j=1}^{N_e}$ is a *finite ensemble representation* of a random vector $\mathbf{y}^u$ with mean and covariance given by

$$\boldsymbol{\mu}_{\mathbf{y}^u} = L\begin{bmatrix}\boldsymbol{\mu}_{\mathbf{y}^f} \\ \mathbf{d}_{\text{obs}}\end{bmatrix}, \quad \mathbf{C}_{\mathbf{y}^u} = L\begin{bmatrix}\mathbf{C}_{\mathbf{y}^f} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{\boldsymbol{\varepsilon}}\end{bmatrix}L^T.$$

It is now necessary to show that these expressions give the conditional mean and covariance of the random vector $\mathbf{y}^f|\mathbf{d}_{\text{obs}}$. For the mean, we have

$$\boldsymbol{\mu}_{\mathbf{y}^u} = L\begin{bmatrix}\boldsymbol{\mu}_\mathbf{y} \\ \mathbf{d}_{\text{obs}}\end{bmatrix} = \boldsymbol{\mu}_{\mathbf{y}^f} + \mathbf{K_H}(\mathbf{d}_{\text{obs}} - \mathbf{H}\boldsymbol{\mu}_{\mathbf{y}^f}) \equiv \boldsymbol{\mu}_{\mathbf{y}^f|\mathbf{d}_{\text{obs}}}.$$

For the covariance, we have

$$\begin{aligned}\mathbf{C}_{\mathbf{y}^u} &= \left[\mathbf{1}_M - \mathbf{K_H H}\ \mathbf{K_H}\right]\cdot\begin{bmatrix}\mathbf{C}_{\mathbf{y}^f} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}_{\boldsymbol{\varepsilon}}\end{bmatrix}\cdot\begin{bmatrix}\mathbf{1}_M - \mathbf{H}^T\mathbf{K}_\mathbf{H}^T\\ \mathbf{K}_\mathbf{H}^T\end{bmatrix}\\ &= \left[\mathbf{C}_\mathbf{y}^f - \mathbf{K_H H}\mathbf{C}_\mathbf{y}^f\ \mathbf{K_H}\mathbf{C}_{\boldsymbol{\varepsilon}}\right]\cdot\begin{bmatrix}\mathbf{1}_M - \mathbf{H}^T\mathbf{K}_\mathbf{H}^T\\ \mathbf{K}_\mathbf{H}^T\end{bmatrix}\\ &= (\mathbf{C}_{\mathbf{y}^f} - \mathbf{K_H H}\mathbf{C}_{\mathbf{y}^f})(\mathbf{1}_M - \mathbf{H}^T\mathbf{K}_\mathbf{H}^T) + \mathbf{K_H}\mathbf{C}_{\boldsymbol{\varepsilon}}\mathbf{K}_\mathbf{H}^T\\ &= \mathbf{C}_{\mathbf{y}^f} - \mathbf{K_H H}\mathbf{C}_{\mathbf{y}^f} - \mathbf{C}_{\mathbf{y}^f}\mathbf{H}^T\mathbf{K}_\mathbf{H}^T\\ &\quad + \mathbf{K_H H}\mathbf{C}_{\mathbf{y}^f}\mathbf{H}^T\mathbf{K}_\mathbf{H}^T + \mathbf{K_H}\mathbf{C}_{\boldsymbol{\varepsilon}}\mathbf{K}_\mathbf{H}^T\\ &= \mathbf{C}_{\mathbf{y}^f} - \mathbf{K_H H}\mathbf{C}_{\mathbf{y}^f} - \mathbf{C}_{\mathbf{y}^f}\mathbf{H}^T\mathbf{K}_\mathbf{H}^T\\ &\quad + \mathbf{K_H}(\mathbf{H}\mathbf{C}_{\mathbf{y}^f}\mathbf{H}^T + \mathbf{C}_{\boldsymbol{\varepsilon}})\mathbf{K}_\mathbf{H}^T.\end{aligned}$$

Using the explicit expression for $\mathbf{K_H}$, we obtain

$$\begin{aligned}\mathbf{C}_{\mathbf{y}^u} &= \mathbf{C}_{\mathbf{y}f} - \mathbf{C}_{\mathbf{y}^f}\mathbf{H}^T(\mathbf{H}\mathbf{C}_{\mathbf{y}^f}\mathbf{H}^T + \mathbf{C}_{\boldsymbol{\varepsilon}})^{-1}\mathbf{H}\mathbf{C}_{\mathbf{y}^f}\\ &\quad - \mathbf{C}_{\mathbf{y}^f}\mathbf{H}^T((\mathbf{H}\mathbf{C}_{\mathbf{y}^f}\mathbf{H}^T + \mathbf{C}_{\boldsymbol{\varepsilon}})^{-1})^T\mathbf{H}\mathbf{C}_{\mathbf{y}^f}\\ &\quad + \mathbf{C}_{\mathbf{y}^f}\mathbf{H}^T(\mathbf{H}\mathbf{C}_\mathbf{H}\mathbf{H}^T + \mathbf{C}_{\boldsymbol{\varepsilon}})^{-1}(\mathbf{H}\mathbf{C}_{\mathbf{y}^f}\mathbf{H}^T + \mathbf{C}_{\boldsymbol{\varepsilon}})\\ &\quad ((\mathbf{H}\mathbf{C}_{\mathbf{y}^f}\mathbf{H}^T + \mathbf{C}_{\boldsymbol{\varepsilon}})^{-1})^T\mathbf{H}\mathbf{C}_{\mathbf{y}^f}\\ &= \mathbf{C}_{\mathbf{y}^f} - \mathbf{C}_{\mathbf{y}^f}\mathbf{H}^T(\mathbf{H}\mathbf{C}_{\mathbf{y}^f}\mathbf{H}^T + \mathbf{C}_{\boldsymbol{\varepsilon}})^{-1}\mathbf{H}\mathbf{C}_\mathbf{y}^f\\ &= \mathbf{C}_{\mathbf{y}^f} - \mathbf{K_H H}\mathbf{C}_{\mathbf{y}^f} \equiv \mathbf{C}_{\mathbf{y}^f|\mathbf{d}_{\text{obs}}}.\end{aligned}$$

This proves that the updated vectors $\{\mathbf{y}_j^u\}_{j=1}^{N_e}$ are a *finite ensemble representation* of the conditional random vectors $\mathbf{y}^f|\mathbf{d}_{\text{obs}}$.

A.3 Finite ensemble representation and linear transform of GMM

If $\{\mathbf{y}_j\}$ is a *finite ensemble representation* of the random vector $\mathbf{y}$ distributed according to a Gaussian mixture with weights $\{\pi_k\}_{k=1}^{N_c}$, means $\{\boldsymbol{\mu}_\mathbf{y}^k\}_{k=1}^{N_c}$ and covariances $\{\mathbf{C}_\mathbf{y}^k\}_{k=1}^{N_c}$ there is a partition such that each subset $\{\mathbf{y}_{i_k}\}_{i_k\in I_k}$ is a *finite ensemble representation* of the Gaussian component k. Then the linear property of Gaussian mixtures (see [2]) allows to conclude that $\mathbf{z} = A\mathbf{y}$ is distributed according to a Gaussian mixtures with PDF given by

$$f(\mathbf{z}) = \sum_{k=1}^{N_c}\pi_k N(\mathbf{z};\, A\boldsymbol{\mu}_\mathbf{y}^k,\, A\mathbf{C}_\mathbf{y}^k A^T).$$

The *finite ensemble representation* under linear transform property applied to each component leads to the conclusion that the transformed ensemble $\{\mathbf{z}_j\}$ is again a *finite ensemble representation* for the linear transformed random vector $\mathbf{z}$.

A.4 Finite Ensemble Representation for GMM updating algorithm

If the prior is a GM, the distribution of $\mathbf{y}^f|\mathbf{d}_{\text{obs}}$ in the linear case is again a Gaussian mixture with weights, means and covariances given in Section 3 [2]. If the index of components of the posterior distribution is sampled according to step (2.b), we get n_ℓ with $\ell = 1,\ldots,N_c$ such that

$$|\ \lambda^\ell_{\mathbf{y}^f|\mathbf{d}_{\text{obs}}} - \frac{n_\ell}{N_e}\ | \le \eta$$

where n_ℓ is the number of vectors $\mathbf{y}^u_{j_\ell}$ in the partition I_ℓ. In step (2.c), an element $\mathbf{y}^f_j$ of the *finite ensemble representation* is moved from component k to component ℓ. To prove this statement, we consider the equation

$$\Delta\mathbf{y}^{f\prime} = L^\ell(L^k)^{-1}\Delta\mathbf{y}^f$$

applied to the random vector $\Delta\mathbf{y}^{f\prime} = \mathbf{y}^f - \boldsymbol{\mu}^k_{\mathbf{y}^f}$ with zero mean and covariance $\mathbf{C}^k_{\mathbf{y}^f}$. According to the usual property of the linear transform we have that the mean of the Gaussian vector $\Delta\mathbf{y}^{f\prime}$ is

$$\boldsymbol{\mu}_{\Delta\mathbf{y}^{f\prime}} = L^\ell(L^k)^{-1}\mathbf{0} = \mathbf{0}$$

while the covariance is

$$\begin{aligned}\mathbf{C}_{\Delta\mathbf{y}^{f\prime}} &= (L^\ell(L^k)^{-1})\mathbf{C}^k_{\mathbf{y}^f}(L^\ell(L^k)^{-1})^T\\ &= (L^\ell(L^k)^{-1})(L^k(L^k)^T)(L^\ell(L^k)^{-1})^T\\ &= L^\ell(L^\ell)^T = \mathbf{C}^\ell_{\mathbf{y}^f}.\end{aligned}$$

Adding $\boldsymbol{\mu}^\ell_{\mathbf{y}^f}$ to the random vector $\Delta\mathbf{y}^{f\prime}$ we obtain the mean and the covariance of the required Gaussian component. The *finite ensemble representation* linear transform property applied to the transformation from $\mathbf{y}^f$ to $\mathbf{y}^{f\prime}$, allows to conclude that the vectors $\mathbf{y}^{f\prime}_j$ obtained in step (2.c) are a *finite ensemble representation* of the prior Gaussian component ℓ and the updating equations given in step (2.d) assure that $\mathbf{y}^u_j$ are a *finite ensemble representation* of the conditional component ℓ on the basis of the given property for ensemble Kalman filter on Gaussian distribution.

References

1. Aanonsen, S.I., Nævdal, G., Oliver, D.S., Reynolds, A.C.: The ensemble Kalman filter in reservoir engineering—a review. SPE J. **14**(3), 393–412 (2009)
2. Alspach, D.L., Sorenson, H.W.: Nonlinear Bayesian estimation using Gaussian sum approximation. IEEE Trans. Automat. Contr. **17**, 439–448 (1972)
3. Bengtsson, T., Snyder, C., Nychka, D.: Toward a nonlinear ensemble filter for high-dimensional systems. J. Geophys. Res. **108**(D24), 8775 (2002)

4. Burgers, G., van Leeuwen, P.J., Evensen, G.: Analysis scheme in the ensemble Kalman filter. Mont. Weather Rev. **126**, 1719–1724 (1998)
5. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. **39**, 1–38 (1977)
6. Dovera, L., Della Rossa, E.: Ensemble Kalman filter for Gaussian mixture models. Petroleum Geostatistics, A16, Eur. Assn. Geosci. Eng. (2007)
7. Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res. **99**, 10143–10162 (1994)
8. Evensen, G.: Data Assimilation—the ensemble Kalman filter. Springer, Berlin (2006)
9. Evensen, G., van Leeuwen, P.J.: An ensemble Kalman smoother for nonlinear dynamics. Mon. Weather Rev. **128**, 1852–1867 (1999)
10. Furrer, R., Bengtsson, T.: Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. J. Multivar. Anal. **98**, 227–255 (2007)
11. Gaspari, G., Cohn, S.E.: Construction of correlation functions in two and three dimensions. Q. J. Royal Meteorol. Soc. **125**, 723–757 (1999)
12. Grana, D., Della Rossa, E.: Probabilistic petrophysical-properties estimation integrating statistical rock physics with seismic inversion. Geophysics **75**(3), O21–O37 (2010)
13. Grana, D., Della Rossa, E., D'Agosto, C.: Petrophysical properties estimation in a Crosswell study integrated with statistical rock Physics. RC3.6 Soc. Exp. Geoph. (2009)
14. Goovaerts, P.: Geostatistics for natural resources evaluation. Oxford University Press, New York (1997)
15. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning. Springer, Berlin (2003)
16. Haugen, V.E., Evensen, G.: Assimilation of SST and SLA data into OGCM for the Indian Ocean. Ocean Dyn. **52**, 133–151 (2002)
17. Houtekamer, P.L., Mitchell, H.L., Pellerin, G., Buehner, M., Channon, L., Spacek, L., Hansen, B.: Atmospheric data assimilation with an ensemble Kalman filter: results with real observations. Mon. Weather Rev. **133**, 604–620 (2005)
18. Kalman, R.E.: A new approach to linear filtering and prediction problems. J. Basic Eng. **82**, 35–45 (1960)
19. Lorenz, E.N.: Deterministic nonperiodic flow. J. Atmos. Sci **20**, 130–141 (1963)
20. Mandel, J., Cobb, L., Beezley, J.D.: On the convergence of the ensemble Kalman filter. Available at: arxiv:0901.2951v1 (2009). Accessed 20 January 2009
21. Maybeck, P.S.: Stochastic Models, Estimation and Control, vol. 1. Academic Press, New York (1979)
22. Mardia, K., Kent, J., Bibby, J.: Multivariate Analysis. Academic Press, New York (1979)
23. Nævdal, G., Johnsen, L.M., Aanonsen, S.I., Vefring, E.H.: Reservoir monitoring and continuous model updating using ensemble Kalman filter. SPE J. **10**(1), 66–74 (2005)
24. Pham, D.T.: Stochastic methods for sequential data assimilation in strongly nonlinear systems. Mont. Weather Rev. **129**, 1194–1207 (2001)
25. Smith, K.W.: Cluster ensemble Kalman filter. Tellus **59A**, 749–757 (2007)
26. Van der Merwe, R., Wan, E.A.: Gaussian mixture sigma-point particle filters for sequential probabilistic inference in dynamics state-space models. In: Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Hong Kong (2003)
27. Zafari, M., Reynolds, A.: Assessing the uncertainty in reservoir description and performance predictions with the ensemble Kalman filter. SPE J. **12**(3), 382–391 (2007)