Accepted Manuscript

Assessing clustering strategies for Gaussian mixture filtering a subsurface contaminant model

B. Liu, M.E. Gharamti, I. Hoteit

PII: S0022-1694(16)00066-4

DOI: <http://dx.doi.org/10.1016/j.jhydrol.2016.01.048>

Reference: HYDROL 21007

To appear in: *Journal of Hydrology*

Received Date: 16 September 2015

Revised Date: 13 December 2015

Accepted Date: 19 January 2016

Please cite this article as: Liu, B., Gharamti, M.E., Hoteit, I., Assessing clustering strategies for Gaussian mixture filtering a subsurface contaminant model, *Journal of Hydrology* (2016), doi: <http://dx.doi.org/10.1016/j.jhydrol.2016.01.048>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Assessing clustering strategies for Gaussian mixture filtering a subsurface contaminant model

B. Liu^a, M.E. Gharamti^{b,c}, I. Hoteit^{a,b,*}

^a*Computer, Electrical and Mathematical Sciences and Engineering, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia*

^b*Earth Sciences and Engineering, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia*

^c*Mohn-Sverdrup Center for Global Ocean Studies and Operational Oceanography, Nansen Environmental and Remote Sensing Center (NERSC), Bergen 5006, Norway*

Abstract

An ensemble-based Gaussian mixture (GM) filtering framework is studied in this paper in term of its dependence on the choice of the clustering method to construct the GM. In this approach, a number of particles sampled from the posterior distribution are first integrated forward with the dynamical model for forecasting. A GM representation of the forecast distribution is then constructed from the forecast particles. Once an observation becomes available, the forecast GM is updated according to Bayes' rule. This leads to (i) a Kalman filter-like update of the particles, and (ii) a Particle filter-like update of their weights, generalizing the ensemble Kalman filter update to non-Gaussian distributions. We focus on investigating the impact of the clustering strategy on the behavior of the filter. Three different clustering methods for constructing the prior GM are considered: (i) a standard kernel density estimation, (ii) clustering with a specified mixture component size, and (iii) adaptive clustering (with a variable GM size). Numerical experiments are performed using a two-dimensional reactive contaminant transport model in which the contaminant concentration and the heterogeneous hydraulic conductivity fields are estimated within a confined aquifer using solute concentration data. The experimental results suggest that the performance of the GM filter is sensitive to the choice of the GM model. In particular,

*Corresponding author, Tel: +966(12)808-0344

Email address: ibrahim.hoteit@kaust.edu.sa (I. Hoteit)

URL: <http://assimilation.kaust.edu.sa/> (I. Hoteit)

increasing the size of the GM does not necessarily result in improved performances. In this respect, the best results are obtained with the proposed adaptive clustering scheme.

Keywords: Ensemble Kalman filtering, Particle filtering, Gaussian mixtures, Clustering, Subsurface contaminant model.

1. Introduction

Contamination of subsurface aquifers with hazardous substances such as oil spills, pesticides and excessive use of fertilizers may pose dangerous problems not only to the environment but also to humans. According to the US Environmental Protection Agency (Belk et al., 1993), the major consequences of groundwater contamination usually cover poor-quality of drinking water, loss of water supply, high cleanup costs, high costs for alternative water supplies, and potential health problems including Hepatitis, Cholera, and Methemoglobinemia (a.k.a blue baby syndrome). It is therefore crucial for field experts to understand contamination problems and find effective feasible solutions for remediating it. From a modeling perspective, hydrologists have been working on addressing contamination problems by building numerical models that reflect the groundwater flow physics and the dynamics of the migrating contaminants (e.g., Javandel et al., 1984; Bedient et al., 1994; Bear and Cheng, 2009). These models, although very explicit, suffer from various limitations, including the use of physical assumptions and approximations in addition to poorly understood parameter fields (e.g., permeability and porosity). These limit the accuracy of subsurface models and affect their capability to reliably simulate and predict the responses of the subsurface (Gharamti et al., 2014a). Through inverse methods and Bayesian filtering approaches, hydrologists assess the quality of these models and address their uncertainties by constraining the models outputs with available data, a process known as Data Assimilation (DA) (Hendricks-Franssen and Kinzelbach, 2008; Gharamti et al., 2014b).

In Bayesian filtering, the distribution of the state of a dynamical system given a set of observations, called posterior distribution, is proportional to the product of the fore-

cast distribution of the system state and the likelihood of the observations given the state (Gustafsson et al., 2002). In general, it is not possible to derive the analytical solution of the nonlinear Bayesian filter and the only way to implement this approach is to compute, instead, a discrete numerical solution (Hoteit et al., 2008; van Leeuwen, 2009; Luo et al., 2012).

The Particle filter (PF) discretizes the state probability distribution by a set of weighted particles (Liu, 2008). The predictive evolution of the probability density function is then carried by propagating all particles forward in time with the dynamical model. When observations become available, the relative likelihood of the forecast particles is computed by updating their weights. Although the PF recently became popular in subsurface hydrology (Chang et al., 2012; Leisenring and Moradkhani, 2011), it still suffers from the degeneracy of the particles and the required large number of particles to accurately sample the filter distribution in large scale applications (Montzka et al., 2011; Moradkhani et al., 2012; Hoteit et al., 2012).

A more common approach for data assimilation in subsurface hydrology is the ensemble Kalman filter (EnKF) (Evensen, 1994; Burgers et al., 1998). The EnKF represents the first two moments (i.e., mean and covariance) of the Kalman filter (KF) by an ensemble of particles with uniform weights. The evolution of the KF statistics in time is then carried out, as in the PF, by integrating the ensemble members with the dynamical model. The ensemble is then updated with incoming observations as in the KF, by approximating it as a Gaussian distribution (Hoteit et al., 2012). This avoids the linearization of the system and makes the computations feasible even with high-dimensional nonlinear systems (Hendricks-Franssen and Kinzelbach, 2008, 2009; Gharamti et al., 2013, 2015). This success was attributed to the Kalman update of the particles which focuses the sampling of the particles to the area indicated by the observations, allowing the filter to efficiently operate with small ensembles (Kivman, 2003; Hoteit et al., 2008, 2012).

The fundamental advantage of nonlinear filtering is the ability to capture non-Gaussian

distributions. The single Gaussian representation of the state distribution in the EnKF reduces the computational burden, but it largely impoverishes this advantage (Anderson and Anderson, 1999; Hotcit et al., 2012). One way to compromise between the computational feasibility and the ability of representing non-Gaussian features is to use a convex sum of Gaussian densities; namely a Gaussian mixture (GM) model, to approximate the filter probability distributions (Alspach and Sorenson, 1972). Anderson and Anderson (1999) introduced the kernel-based GM representations of the prior distributions and presented a comparison with a single Gaussian representation in the Bayesian filter. Hoteit et al. (2008) developed a kernel particle Kalman filter for large-scale data assimilation problems in which a low-rank kernel GM is used to approximate the filter's distributions. Sun et al. (2009) proposed a clustering-based GM filter in which the forecast ensemble is first projected onto a low dimensional subspace, and then a prior GM is constructed based on the Expectation-Maximization (EM) technique. The number of Gaussian components is optimised by using a Bayesian information criterion (BIC). More recently, Hoteit et al. (2012) investigated an ensemble-based GM scheme in which, each component of the GM was updated through an ensemble of particles. The resulting filter was viewed as a set of weighted EnKFs operating in parallel. Sondergaard and Lermusiaux (2013) proposed a clustering-based GM filter in which the GM is propagated with the nonlinear model using the stochastic Dynamically Orthogonal field equations.

In this work, we follow the GM filtering framework proposed by Anderson and Anderson (1999) and focus on the construction of prior GMs using different clustering strategies. In this filter, an ensemble of particles sampled from the posterior distribution are first integrated forward with the model to approximate the forecast distribution. Three different clustering strategies are then presented and applied to construct a prior GM from the forecast particles. The first method follows the standard Kernel density estimation approach and treats the samples as the centres of the Gaussian components, all with the same covariance (Silverman, 1986). In this scheme, the GM mixture size is fixed and equal to the ensemble size. In the

second method, the particles are clustered into a predefined number of groups, each representing a Gaussian component. The third method uses an adaptive number of components, dynamically estimated according to a squared error criterion. Once observations become available, the analysis scheme of the GM filter is applied following two steps: a Kalman-type analysis step to update the center of each Gaussian component, and a PF-type analysis step to update the weights of the GM components (Alspach and Sorenson, 1972). We conduct numerical experiments in a 2D domain in which the flow of pure groundwater into a fully contaminated aquifer zone is simulated for 150 days. Contaminant concentration data from different subsurface wells are assimilated to estimate the concentration and the conductivity of a subsurface contaminant model. The goal is to assess the performance of the GM filter and study its sensitivity to the clustering strategy, and to evaluate its performance against that of the EnKF. In particular, we also investigate the impact of the number of components in the mixture and the clustering parameters on the GM filter behaviour. The rules of thumb for choosing those tuning parameters are discussed.

This paper is organised as follows. Section 2 briefly presents the ensemble GM filter. Three clustering methods for constructing the GM from the forecast ensemble are then presented and discussed in Sections 3. The setup of the numerical experiments is described in Section 4. Assimilation results are analysed in Section 5. Concluding remarks are given in Section 6.

2. Ensemble Gaussian mixture filtering

Consider an n -dimensional discrete-time dynamical system with a linear measurement model:

$$x_{k+1} = \mathcal{M}_k(x_k) + \eta_k, \quad (1)$$

$$y_k = H_k x_k + v_k, \quad (2)$$

where x_k represents an n -dimensional state vector consisting of both the model dynamical variables and the static parameters at time k , $\mathcal{M}_k$ is a nonlinear dynamical operator, and η_k represents a zero mean Gaussian white noise with covariance Q_k . The linear mapping operator H_k represents the measurement model and the random vector v_k denotes the measurement noise, assumed uncorrelated in time with zero-mean and covariance R_k . It is also assumed that η_k and v_k are uncorrelated. The case of nonlinear observations is commonly handled by augmenting the state vector with $H_k(x_k)$. In this case, the new state vector becomes $z_k = [x_k^T \ H_k(x_k)^T]^T$ and the observation is again linear composed of the last components of z_k [Hoteit et al. \(2002\)](#).

Bayesian filtering consists of computing the posterior distribution of x_k given available observations up to the estimation time $Y_k = \{y_i, i = 1, 2, \dots, k\}$, i.e., $p(x_k|Y_k)$. Bayes' rule is recursively applied to compute the posterior distribution as follows:

$$p(x_k|Y_k) = \frac{p(y_k|x_k)p(x_k|Y_{k-1})}{p(y_k|Y_{k-1})}, \quad (3)$$

where $p(y_k|x_k)$ is the likelihood of observation y_k given x_k . The prior distribution $p(x_k|Y_{k-1})$ is obtained from the posterior density at the previous assimilation cycle as:

$$p(x_k|Y_{k-1}) = \int p(x_k|x_{k-1})p(x_{k-1}|Y_{k-1})dx_{k-1}. \quad (4)$$

We start from a posterior distribution of the state at time k modeled as a GM, as follows:

$$p(x_k|Y_k) = \sum_{i=1}^m w_{k|k}^i \mathcal{N}(x_k; \mu_{k|k}^i, \mathbf{P}_{k|k}^i), \quad (5)$$

where $w_{k|k}^i$, $\mu_{k|k}^i$, and $\mathbf{P}_{k|k}^i$ represent the conditional weight, mean and covariance of the i th Gaussian component with respect to the measurements Y_k . Integer m denotes the number of Gaussian components in the mixture and $\mathcal{N}(\cdot|\mu, \mathbf{P})$ refers to a Gaussian distribution with mean μ and covariance $\mathbf{P}$. The positivity and normalization constraints impose:

$$\sum_{i=1}^m w_{k|k}^i = 1, \quad \text{and} \quad w_{k|k}^i \geq 0, \forall k. \quad (6)$$

The difference between the clustering based ensemble GM filter presented here and the analytical GM filter of [Alspach and Sorenson \(1972\)](#) is in the forecast step. Analytically propagating a GM with linearized models can be computationally very demanding for high dimensional nonlinear systems ([Hoteit et al., 2008, 2012](#)). A resampling procedure is thus implemented following [Anderson and Anderson \(1999\)](#) to approximate the posterior GM by an ensemble of state vectors as follows:

$$\{x_{k|k}^i\}_{i=1}^N \sim p(x_k|Y_k) = \sum_{i=1}^m w_{k|k}^i \mathcal{N}(x_k; \mu_{k|k}^i, \mathbf{P}_{k|k}^i). \quad (7)$$

Each ensemble member is then integrated with the nonlinear model to approximate the forecast distribution following a Monte Carlo approach, exactly as in the PF and EnKF:

$$\{x_{k+1|k}^i\}_{i=1}^N = \{\mathcal{M}_k(x_{k|k}^i) + \eta_k^i\}_{i=1}^N, \quad (8)$$

where η_k^i is a realization from $\mathcal{N}(\cdot; 0, Q_k)$. To re-construct a GM representation, a clustering

scheme is applied to model the forecast ensemble distribution as a GM:

$$p(x_{k+1}|Y_k) \approx \sum_{i=1}^m w_{k+1|k}^i \mathcal{N}(x_{k+1}, \mu_{k+1|k}^i, \mathbf{P}_{k+1|k}^i). \quad (9)$$

Once a new observation y_{k+1} becomes available, the posterior probability distribution with respect to Y_{k+1} is also a GM, given as:

$$p(x_{k+1}|Y_{k+1}) = \sum_{i=1}^m w_{k+1|k+1}^i \mathcal{N}(x_{k+1}; \mu_{k+1|k+1}^i, \mathbf{P}_{k+1|k+1}^i), \quad (10)$$

in which the means and the covariances of the mixture components are updated as in a Kalman filter (Alspach and Sorenson, 1972), i.e.,

$$\mathbf{G}_{k+1}^i = \mathbf{P}_{k+1|k}^i H_{k+1}^T (H_{k+1} \mathbf{P}_{k+1|k}^i H_{k+1}^T + R_k)^{-1}, \quad (11)$$

$$\mu_{k+1|k+1}^i = \mu_{k+1|k}^i + \mathbf{G}_{k+1}^i (y_{k+1} - H_{k+1} \mu_{k+1|k}^i), \quad (12)$$

$$\mathbf{P}_{k+1|k+1}^i = (\mathbf{I} - \mathbf{G}_{k+1}^i H_{k+1}) \mathbf{P}_{k+1|k}^i. \quad (13)$$

This step is similar to the analysis scheme of the EnKF without observation perturbations but may use a different Kalman gain for each ensemble member. The weights are also updated based on a PF-like update step:

$$w_{k+1|k+1}^i = \frac{w_{k+1|k}^i \mathcal{N}(y_{k+1} - H_{k+1} \mu_{k+1|k}^i | 0, H_{k+1} \mathbf{P}_{k+1|k}^i H_{k+1}^T + R_k)}{\sum_{j=1}^c w_{k+1|k}^j \mathcal{N}(y_{k+1} - H_{k+1} \mu_{k+1|k}^j | 0, H_{k+1} \mathbf{P}_{k+1|k}^j H_{k+1}^T + R_k)}. \quad (14)$$

A posterior estimate of the state at time t_{k+1} is the expectation of the posterior probability density function (10), and is given by:

$$\hat{x}_{k+1|k+1} = \sum_{i=1}^m w_{k+1|k+1}^i \mu_{k+1|k+1}^i. \quad (15)$$

To proceed with the forecast cycle, a set of new particles, $\{x_{k+1|k+1}^i\}_{i=1}^N$, needs to be resam-

pled from the analysis GM (10). We follow Anderson and Anderson (1999) to implement this in two steps. First, generate N random scalars l_j , $j = 1, \dots, N$ with probabilities $p(l_j = i) = w_{k+1|k+1}^i$. Then draw the new samples $x_{k+1|k+1}^{(j)}$ from the Gaussian distribution $\mathcal{N}(x_{k+1} | \mu_{k+1|k+1}^{l_j}, \mathbf{P}_{k+1|k+1}^{l_j})$.

3. Construction of the prior GM

Three clustering strategies for constructing the prior GM are presented and discussed in this section.

3.1. Kernel ensemble GM filter (KEnGMF)

The most straightforward approach to construct a GM from the forecast samples, $\{x_{k+1|k}^i\}_{i=1}^N$, is to use the kernel density estimation (Silverman, 1986). We thus consider:

$$p(x_{k+1} | Y_k) \approx \sum_{i=1}^N \frac{1}{N} \mathcal{N}(x_{k+1}; x_{k+1|k}^i, \mathbf{B}_{k+1|k}), \quad (16)$$

where $\mathbf{B}_{k+1|k}$ is a symmetric and positive definite bandwidth matrix. The number of Gaussian components in this GM kernel estimation is equal to the number of ensemble members ($m = N$). It further has uniform weights ($w_{k+1|k}^i = 1/N$), and uniform covariances ($\mathbf{P}_{k+1|k}^i = \mathbf{B}_{k+1|k}$). Once the observation y_{k+1} becomes available, equations (11)-(14) are applied to compute the posterior distribution (10).

The bandwidth matrix $\mathbf{B}_{k+1|k}$ characterizes how the approximated probability is continuously assigned to the samples and their neighbor space. Generally speaking, a small $\mathbf{B}_{k+1|k}$ assigns more probability to the sample points and less probability to their neighbor space. The bandwidth matrix is designed to be proportional to the sample covariance of the forecast ensemble as follows:

$$\mathbf{B}_{k+1|k} = \beta \cdot \mathbf{P}_k^f, \quad (17)$$

where β is a non-negative scalar parameter and $\mathbf{P}_k^f$ is the sample covariance of the ensemble:

$$\mathbf{P}_k^f = \frac{1}{N-1} \sum_{i=1}^N (x_{k|k-1}^i - \bar{x}_k^f)(x_{k|k-1}^i - \bar{x}_k^f)^T. \quad (18)$$

$\bar{x}_k^f$ is the mean of the forecast ensemble $\left\{x_{k+1|k}^i\right\}_{i=1}^N$. Parameter β plays a key role in assigning the respective weights of the KF and PF updates in the analysis step. A large β should result in a stronger Kalman gain, which brings the particles closer to the observations. Meanwhile, since the weight is proportional to the Gaussian likelihood with covariance $\beta H_k \mathbf{P}_k^f H_k^T + R_k$, a large β may decrease the variance of the weights, or in other words, limits the impact of the PF update.

Since the scaled sample covariance of the forecast ensemble, $\beta \mathbf{P}_k^f$, is used to represent the covariance of each of the prior Gaussian components, the distribution may not be accurately characterized, and thus the analysis weights in (14) not well estimated. To mitigate this problem, Stordal et al. (2010) efficiently reduced the effect of the weights update by nudging the original analysis weights to an uniform distribution, i.e.,

$$\omega_{k|k}^i = \gamma \omega_{k|k}^i + (1 - \gamma) \frac{1}{N}, \quad (19)$$

where $\gamma \in [0, 1]$ is a tuning parameter. Although this introduces bias in the filter estimate, Stordal et al. (2010) argued that it still capture some of the non-Gaussian features left out by the KF update. Note also that the EnKF is actually a variant of the KEnGMF with bandwidth $\beta = 1$ and weights nudging $\gamma = 0$. The kernel density estimation (16) uses the same weight and covariance for all Gaussian components, offering little flexibility in approximating the local pdfs. Moreover, its number of Gaussian components is equal to the number of samples, which may overfit the prior pdf by too many Gaussian components.

175 3.2. Clustered ensemble GM filter (CEnGMF)

To better approximate the forecast distribution $p(x_{k+1}|Y_k)$ by a GM, a clustering technique is applied to classify the forecast samples $\{x_{k+1|k}^i\}_{i=1}^N$ into m clusters $s(m) = \{S_{m,1}, S_{m,2}, \dots, S_{m,m}\}$, and then each cluster is approximated by a Gaussian component. Several clustering methods have been proposed in the literature, for example, see (Scott, 1992; Härdle and Simarx, 2011). Here, we consider the most popular *k-means* method for its simplicity and proven robustness (Duda et al., 2001; MacKay, 2003). The *k-means* method can be summarized as follows. First, let $\mu_{k+1|k}^i$ be the mean of the samples classified in cluster $S_{m,i}$, then the Sum-of-Squared-Error (SoSE) of clustering $s(m)$ is defined as

$$J_e(m) = \sum_{i=1}^m \sum_{x_{k+1|k}^j \in S_{m,i}} \left\| x_{k+1|k}^j - \mu_{k+1|k}^i \right\|^2, \quad (20)$$

where $\|\cdot\|^2$ is the Euclidean norm. This SoSE can be simply interpreted: for a given cluster $S_{m,i}$, the mean vector $\mu_{k+1|k}^i$ is the best representative of the samples in $S_{m,i}$ in the sense that it minimizes the sum of the squared lengths of the “error” vectors $x_{k+1|k}^j - \mu_{k+1|k}^i$ in $S_{m,i}$. Thus, $J_e(m)$ measures the total squared error incurred in representing the N samples $x_{k+1|k}^1, x_{k+1|k}^2, \dots, x_{k+1|k}^N$ by the m cluster centers $\mu_{k+1|k}^1, \mu_{k+1|k}^2, \dots, \mu_{k+1|k}^m$ (Duda et al., 2001). The value of $J_e(m)$ depends on how the samples are grouped into clusters and the number of clusters. Clustering by minimizing $J_e(m)$ is called *k-means* clustering, which is formally defined as follows,

$$s^*(m) = \arg \min_{s(m)} J_e(m), \quad (21)$$

176 where $s^*(m) = \{S_{m,1}^*, S_{m,2}^*, \dots, S_{m,m}^*\}$ is the solution of (21). Given an initial set of clusters
 177 $s^0(m)$, searching for $s^*(m)$ is carried out iteratively by repeating the following two steps
 178 (MacKay, 2003):

- 179 1. *Mean Update step*: For the clusters $s^n(m)$, calculate the mean of each cluster to be the
 180 centroids of the samples in this cluster.

2. *Assignment step*: Assign each sample to the cluster whose mean has the shortest distance to the sample than any other mean. The resulting clusters are $s^{n+1}(m)$.

The algorithm converges when the assignments no longer change. The computational complexity of the algorithm is $O(Nn_xmI)$, where I is the number of iterations (Inaba et al., 1994). Let N_i denotes the number of samples classified into $S_{m,i}$, then the weights and covariances of the prior GM are computed as:

$$w_{k+1|k}^i = \frac{N_i}{N}, \quad (22)$$

$$\mathbf{P}_{k+1|k}^i = \frac{1}{N_i - 1} \sum_{x_{k+1|k}^j \in S_i} (x_{k+1|k}^j - \mu_{k+1|k}^i)(x_{k+1|k}^j - \mu_{k+1|k}^i)^T. \quad (23)$$

This leads to a GM approximation of the forecast distribution as in (9). This clustering based GM representation uses different weights and covariances for each Gaussian component in the mixture. This provides a more efficient framework for fitting the local variations of the pdf than the Gaussian kernel representation, and should further suggest more accurate estimates of the Gaussian components statistics and their weights. This allows to avoid the weights nudging in the CEnGMF, as confirmed in our numerical experiments (see Section 5.2).

3.3. Adaptive Clustered ensemble GM filter (ACEnGMF)

The theoretical foundation of the GM approximation, which states that any probability density function can be approximated as closely as desired by a GM, is based on the condition that a sufficient number of Gaussian components with appropriate covariances could be determined (Alspach and Sorenson, 1972). However, given a set of samples to be clustered, increasing the number of clusters may decrease the number of samples allocated to each cluster, which may in turn degrade the accuracy of the estimated covariances. Intuitively, a suitable choice of the number of components should balance between maximum compression of the samples using a single cluster, and maximum accuracy by assigning each sample to

a different cluster (MacQueen, 1967). In this section, we introduce an adaptive clustering technique in order to determine a balanced GM representation of the prior distribution.

By applying the *k-means* clustering technique, the forecast samples $\{x_{k+1|k}^i\}_{i=1}^N$ are classified into m clusters $s^*(m) = \{S_{m,1}^*, S_{m,2}^*, \dots, S_{m,m}^*\}$, and its associated SoSE $J_e^*(m)$ is calculated as:

$$J_e^*(m) = \sum_{i=1}^m \sum_{x_{k+1|k}^j \in S_{m,i}^*} \left\| x_{k+1|k}^j - \mu_{k+1|k}^i \right\|^2. \quad (24)$$

$J_e^*(m)$ is viewed as a measure of the covariances of the clusters (Mardia et al., 1979). Obviously, increasing the number of components m will always reduce the covariances of the resulting clusters (Duda et al., 2001). In the extreme scenario, zero covariance is obtained when each sample point is considered as its own cluster (i.e., $m = N$). We apply the idea of the *elbow method* (Goutte et al., 1996) to find the optimal cluster size m^* by gradually increasing m from $m = 1$ until the SoSE $J_e^*(m)$ drops significantly, which basically means that adding another cluster will hardly improve the cluster modeling of the samples. This *elbow method* can be illustrated by a graph of the number of clusters m versus its corresponding SoSE $J_e^*(m)$. The “elbow” (or “knee”) reflects the significant drop on the curve of $J_e^*(m)$. This scheme is recursively implemented as follows:

$$\text{If } \frac{J_e^*(m+1)}{J_e^*(m)} > r, \text{ then } m = m+1, \text{ else } m^* = m, \quad (25)$$

where r is a user defined threshold to determine whether the drop of the SoSE is significant, or not. The algorithm stops when the drop in the SoSE is lower than the prescribed threshold r . The computational complexity of this algorithm is $\mathcal{O}(Nn_xMI)$, where $M = \sum_{i=2}^{m^*+1} i$. Naturally, for a set of forecast samples the threshold r implicitly determines the number of clusters. A large threshold r leads to a small number of clusters, while a small threshold r leads to a large number of small clusters. Setting the value of the threshold is usually done by “trial-and-error” (Duda et al., 2001). Here, several experiments are implemented with different values of r and the value of r that achieves the lowest MAE at the end of the

assimilation period is selected. Other methods have been proposed to estimate the “elbow” of the SoSE curve without setting up a threshold. For example, [Salvador and Chan \(2004\)](#) introduced a straight line fitting based scheme to determine the number of clusters. In this work, the clustering procedure is repeated by enumerating m from 1 to a large enough number, say, m_L . Then, using a pair of straight lines to fit the curve of $J_e^*(m)$, the boundary between these lines is regarded as the “elbow”. Repeating the clustering procedure can be however computationally demanding.

This scheme provides an adaptive algorithm to construct GM representations of the prior distribution. By properly designing the threshold r , it allows to determine the component size of the prior GM, which balances the accuracy of the clustering model and the accuracy of the parameters of each component. However, recursively searching for an optimal component size requires repeating the clustering procedure m^* times, which may be computationally more demanding than the CEnGMF.

Note that from the aspect of constructing the prior GM, the EnKF is actually a CEnGMF with $m = 1$, or an ACEnGMF with a large value of r , which basically means that only a single Gaussian component is considered in the representation of the prior distribution. The observation perturbations in the EnKF is regarded as a way of resampling a set of new particles following the posterior Gaussian distribution $\mathcal{N}(\hat{x}_{k+1|k+1}, (\mathbf{I} - \mathbf{G}_{k+1}H_{k+1})\mathbf{P}_{k+1|k})$, where $\mathbf{G}_{k+1}$ is the Kalman filter gain and $\mathbf{P}_{k+1|k}$ is the forecast covariance ([Hoteit et al., 2015](#)).

4. Application to a reactive contaminant transport model

We consider a two-dimensional (2D) flow and transport modeling system in a confined aquifer. The system consists of two coupled partial differential equations, one for the groundwater flow and the other for the reactive solute transport. We aim at estimating the subsurface contamination state and the hydraulic conductivity parameter at every grid cell of the aquifer domain. To do so, we collect synthetic well measurements of the solute concentration

249 and assimilate them into the flow-transport model using the ensemble GM filters.

250 4.1. Model Formulation

The steady-state groundwater flow equation is obtained from the conservation of total fluid volume and Darcy's law (Bear and Cheng, 2009):

$$\nabla \cdot (-\mathbf{K}\nabla h) = \nabla \cdot \mathbf{U} = 0, \quad (26)$$

where $\mathbf{K}$ is the hydraulic conductivity tensor [LT^{-1}], h is the hydraulic head [L], and $\mathbf{U}$ is the Darcy velocity [LT^{-1}]. By considering only the aqueous-solid surface reaction (sorption), the general transport system can be modeled by (Bear and Cheng, 2009):

$$\mathcal{R} \frac{\partial (\phi C)}{\partial t} = \nabla \cdot (\phi \mathcal{D} \nabla C - \mathbf{U}C), \quad (27)$$

where C is the dissolved solute concentration [ML^{-3}] and ϕ is the porosity of the media. The hydrodynamic dispersion coefficient tensor $\mathcal{D}$ for an isotropic porous medium can be represented as follows (Bear and Cheng, 2009):

$$\mathcal{D} = \kappa_l |\mathbf{U}_x| + \kappa_t |\mathbf{U}_y| + \delta_m, \quad (28)$$

where κ_l and κ_t are respectively the longitudinal and transverse pore-scale dispersivity constants [L]. $\mathbf{U}_x$ and $\mathbf{U}_y$ are the Darcy velocity components in the xy plane and δ_m is the effective molecular diffusion [L^2T^{-1}]. Assuming a linear sorption isotherm, the retardation factor $\mathcal{R}$ is expressed as:

$$\mathcal{R} = 1 + \rho_b k_d \phi^{-1}, \quad (29)$$

251 where ρ_b is the bulk density of the solid rocks [ML^{-1}] and k_d is the sorption distribution
252 (partition) coefficient [L^3M^{-1}] describing the ratio between the sorbed and dissolved con-
253 centrations at equilibrium (Bear and Cheng, 2009).

4.2. Model Configuration

We consider a 2D aquifer zone with medium textured soil in a rectangular domain. The steady-state groundwater flow in equation (26) is discretized in space into 50×50 grid cells of $2\text{m} \times 2\text{m}$, using a cell-centered finite difference (CCFD) scheme. The aquifer is assumed initially fully contaminated and water is injected at two different locations of the left boundary as shown in the sketch diagram of Figure 1. The eastern boundary is assigned a prescribed constant water head and the north and south boundaries are assumed impermeable. The transport equation (27) is discretized on the same space domain using the upwind of the CCFD scheme to ensure a stable solution (Bear and Cheng, 2009). All model parameters and boundary conditions are given in Table 1. The time stepping in the coupled system is 12 hours and the total simulation time is 150 days.

Reference log-conductivity fields are generated with sequential Gaussian simulation using the GCOSIM3D code (Gómez-Hernández and Journel, 1993) with a mean of $-13 \log(\text{m/s})$, a variance of $1.5 \log(\text{m}^2/\text{s}^2)$ and a Gaussian variogram with a range equal to 25 and 50 m in the x - and y -directions, respectively (Figure 2).

4.3. Assimilation Experiments Setup

A reference run is first performed using the mentioned parameters and is considered as the reference solution, or the “truth”. Some perturbations are next imposed on the model and the goal is to estimate the model’s solute concentration and hydraulic conductivity using the perturbed model (for forecasting) and assimilating perturbed data extracted from the reference run. Following this twin experimental framework, all quantities are known by design, which allows for full evaluation of the filters performances even on the non-observed variables. The reference final contaminant concentration distribution in the domain is shown in Figure 2 (right panel).

In the forecast model, we impose the following perturbations. Initially, we start with a fully contaminated aquifer domain as in the reference run but with concentration equal to the mean of the long reference solution. To slow the solute migration, 30% perturbation is

imposed on the porosity of the domain, modifying it to $\phi = 0.325$. Solute concentration measurements are collected every 15 days from 9 hydraulic wells, distributed in the domain as shown in Figure 2 (right panel). These measurements are perturbed with Gaussian noise of zero mean and a standard deviation equal to 10% of the total variability. That is, for each observation point, we calculate its standard deviation over the entire simulation period and then take 10% of it as the standard deviation of added noise. For the parameters, we use the geostatistical simulation toolbox (Gómez-Hernández and Journel, 1993) and condition on two measured hydraulic conductivity values (indicated by square symbols in the left panel of Figure 2) obtained from the reference field in order to generate the initial ensemble. Initial contaminant concentration realisations are chosen by randomly selecting snapshots from an unconditional forecast model run. To illustrate, we used the perturbed forecast model to perform a spin-up run for a 2 years period and then selected the initial concentration realisations from different time instances. This practical sampling procedure guarantees enough variability in the initial ensemble. As a way to improve the physical consistency between the sampled initial state and parameter ensembles, one may also perform a long spin up period prior to assimilation to “warm-up” the ensemble members. This should provide more reliable state-parameters cross correlations, especially when the ensemble size is small (Simon and Bertino, 2012).

The proposed three ensemble GM filters are tested with this contaminant transport model and the filter estimates are evaluated based on their mean absolute errors (MAE) and average ensemble spread (AES):

$$\text{MAE} = (Nn)^{-1} \sum_{j=1}^N \sum_{i=1}^n |x_i^t - x_{j,i}^e|, \quad (30)$$

$$\text{AES} = (Nn)^{-1} \sum_{j=1}^N \sum_{i=1}^n |\hat{x}_i^e - x_{j,i}^e|, \quad (31)$$

where x_i^t is the reference “true” value of the state at cell i , $x_{j,i}^e$ and $\hat{x}_i^e$ denote the analyzed

ensemble value and ensemble mean of the variable at location i (after resampling), respectively. The AES measures the ensemble spread, which is often used as an indication of the confidence in the estimated variables and parameters. In all experiments, the estimates of the concentration values are monitored to make sure they are non-negative and physically meaningful. In case that the estimates values are negative, we simply set them to zero as in (Li et al., 2012; Gu and Oliver, 2005; Gharamti et al., 2013). We use the results of the EnKF as a benchmark for comparison.

5. Experimental results

5.1. The Kernel ensemble Gaussian mixture filter (KEnGMF)

This section examines the kernel ensemble Gaussian mixture filter (KEnGMF) using two ensemble sizes $N = 100$, and $N = 300$ with different bandwidth parameters, β . The EnKF is also applied for benchmarking using the same ensemble size. For each ensemble size, we conduct three experiments with the KEnGMF, each with a different bandwidth parameter, $\beta = 0.1, 0.2, 0.3$. The time evolution of MAEs over the simulation period for the concentration and conductivity are plotted in Figure 3. Figures 4-7 further plot the concentration and the conductivity fields at the end of the simulation (day 150) as estimated by the EnKF and the KEnGMF. In these experiments, the weights nudging factor in Eq. (19) was set to $\gamma = 0.1$ to reduce the variance of the analysis weights.

One can clearly notice from the experimental results that the KEnGMF with larger bandwidth parameters exhibits better behavior during the early assimilation period. In Figure 3, for example, the KEnGMF with the smaller bandwidth parameters leads to the largest MAE during the first 80 days. One possible interpretation for this result is that initially the uncertainties imposed on the concentration profile are large and thus the estimate is far from the reference trajectory. Setting a small β means less weight is assigned to the observation, and thus the decrease in the initial estimation error takes place at a slow pace. The EnKF performance is similar to that of the KEnGMF with large bandwidth

parameters during the early assimilation period. Although the KEnGMF with suitable bandwidth parameter outperforms the EnKF at the end of the assimilation window in some cases, for instance the concentration with $\beta = 0.2$ and $N = 100$, the overall performance of the KEnGMF is generally slightly worse than that of the EnKF. We argue that this is due to the kernel-type representation of the Gaussian mixture that uses the same weight and covariance for each Gaussian component.

Spatially, we observe a good agreement between the estimates of the concentration and the reference field after 150 days. Compared to the EnKF estimate, the KEnGMF estimate with $N = 100$ and $\beta = 0.2$ (Fig. 4) is better correlated with the reference solution, especially in the north-east corner of the domain. The EnKF underestimates the contaminant concentration values in that region. For $N = 300$ (Fig. 5), the recovered concentration fields better delineate the major patterns of the reference concentration. The KEnGMF, however, estimates high concentration values in the southern part of the domain. This behaviour becomes more prominent with increasing value of the bandwidth parameter. Regarding the estimates of the conductivity and using 100 ensemble members (Fig. 6), the KEnGMF performs poorly as shown from the estimate of the low-conductivity region located towards the north-west part of the domain, especially for small β values. As we increase the ensemble size to 300 (Fig. 7), the mapping of the conductivity field improves using both the EnKF and the KEnGMF. This is observed from the better recovery of the high and low-conductivity regions.

To summarise, the KEnGMF with large bandwidth parameter can efficiently handle large initial uncertainties, but the robustness of the estimates might deteriorate over time. This is due to the fact that more emphasis is put on the observations, which shrinks the dispersion of the particles. A smaller bandwidth helps the KEnGMF to better capture the nonlinearities over time and achieve better performance, however, at the cost of a slower convergence rate.

5.2. The Clustered Ensemble Gaussian mixture filter (CEnGMF)

We apply the clustered ensemble GM filter (CEnGMF) under the same conditions as the KEnGMF, using two ensemble sizes, $N = 100, 300$, and different number of Gaussian components, $m = 2, 3, 4$. The time evolution of MAEs over the simulation period for the concentration and conductivity are plotted in Figure 8. Concentration estimates at day 150 obtained with the different numbers of Gaussian components in the GM are shown in Figure 9 for $N = 100$ and Figure 10 for $N = 300$, respectively. Note that the CEnGMF with a single Gaussian component (i.e., $m = 1$) is actually the EnKF. The estimated hydraulic conductivity fields at day 150 are also shown in Figure 11 for $N = 100$ and Figure 12 for $N = 300$, respectively.

One can notice two clear patterns in the behaviour of the CEnGMF. First, the CEnGMF exhibits a strong sensitivity with respect to the number of Gaussian components, m . Using $N = 100$ ensemble members, the best CEnGMF results were obtained with $m = 2$. Increasing the number of components (e.g., $m = 4$) degrades the CEnGMF performance, especially for the hydraulic conductivity. Only when the ensemble size is large enough, here $N = 300$, the CEnGMF performance improves with more Gaussian components. This is consistent with the discussion in Section 3 and can be explained by the fact that for a given ensemble size, the filter performance may degrade when too many Gaussian components are considered, as this reduces the number of samples assigned to each Gaussian component. This indeed limits the accuracy of the estimated sample covariance matrices of the components. Secondly, for the small ensemble size ($N = 100$), less Gaussian components may be enough for the estimation of conductivity, but not for the concentration. For instance, the CEnGMF with $m = 2$ exhibits the best performance in terms of the concentration estimate. For the conductivity, the CEnGMF with $m = 1$, i.e., EnKF, provides the most accurate solution. This is due to the fact that large ensembles are generally needed to obtain reliable estimates of the parameters (i.e., conductivity). Unlike the concentration that is directly observed, the accuracy of the parameters estimation highly depends on how well the cross-correlations with the data are

constructed (Gharamti et al., 2014b).

A spatial assessment of the recovered fields suggests a better behavior of the EnKF for the conductivity estimates when using $N = 100$ ensemble members. Clearly, using fairly small ensembles causes an over-shooting in the CEnGMF estimates. When the ensemble size is increased to $N = 300$, both the concentration and the conductivity estimates resulting from the CEnGMF with different number of Gaussian components in the GM significantly improve and the CEnGMF with $m = 2, 4$ even exhibits higher correlations with the reference solutions than the CEnGMF with $m = 1$ (EnKF) at the end of the experiment.

5.3. The adaptive clustered ensemble Gaussian mixture filter (ACEnGMF)

We also implemented the adaptive clustered ensemble Gaussian mixture filter (ACEnGMF) using $N = 100$ and $N = 300$ ensemble members. We use exactly the same set-up as in the previous experiments. As discussed in Section 3.3, a too large value of r may result in a single Gaussian representation of the forecast distribution ($m^* = 1$), and a too small value will likely lead to a prior GM with too many Gaussian components. A suitable range of the value of r should be case-specific and there is no obvious way to analytically compute a “best” value. Here, we examine two different values; $r = 0.3$ and 0.4 in which iteration (25) stops once the SoSE $J_e^*(m)$ is respectively less than 30% (i.e., $r = 0.3$) and 40% (i.e., $r = 0.4$) after adding one more Gaussian component into the cluster. The maximum number of Gaussian components is limited to 4 in this experiment. The time evolution of the MAEs over the entire simulation period are plotted for both the concentration and the conductivity in Figure 13. Figures 14 and 15 plot the snapshots of the estimated concentration and conductivity at the end of the assimilation (day 150), respectively.

Compared to the results of the CEnGMF, the ACEnGMF clearly suggests improved features in the estimated fields. As shown, in Figures 14 and 15, the dominant concentration and conductivity patterns are very well estimated especially for the case with $N = 300$. The ACEnGMF shows a more robust behavior with respect to the change in the clustering parameter, i.e., the dropping rate threshold ($r = 0.3$ and 0.4). Moreover, for a given r

the performance of the ACEnGMF is consistent for the estimation of the concentration and the conductivity. For instance, with $r = 0.4$ the ACEnGMF provides the most accurate estimates for both concentration and conductivity. Furthermore, for the conductivity, the behavior of the ACEnGMF is close to that of the EnKF during the early assimilation period. Examining more closely the results, the ACEnGMF initially tends to generate GMs with less Gaussian components, which imposes larger Kalman gains, allowing the filter to quickly respond to the large initial uncertainties. For large enough ensembles, the ACEnGMF seems to benefit from the non-Gaussianities captured by the GM as more data are assimilated into the system. One can see that when $N = 300$, the ACEnGMF with $r = 0.3$ and $r = 0.4$ leads to higher accuracy than that of the EnKF at the end of the assimilation period.

Overall, compared to the results of the CEnGMF, the ACEnGMF provides better estimates and is more robust with respect to changes in the clustering parameters. This is due to its adaptive formulation, which provides an efficient way to more accurately represent the prior distribution by a GM.

5.4. Analyzing the filters' average ensemble spread (AES)

The proposed three ensemble GM filters tested with the contaminant transport model are further evaluated based on their average ensemble spread (AES). The time-series evolution of AES obtained from the three filters are plotted in Figures 16, 17 and 18, respectively. As can be seen from the subplots of Figure 16, the confidence in the resulting KEnGMF estimates is quite low and this is reflected by the large ensemble spread values for both concentration and conductivity estimates. For instance, for an ensemble of 300 realizations the AES for conductivity after 150 days is 0.35 and 0.21 log(m/s) for the KEnGMF ($\beta = 0.1$) and the EnKF, respectively. As the value of the bandwidth parameter increases, the AES for both concentration and conductivity increases, suggesting large uncertainties in the estimates. This behavior is somehow consistent with the resulting MAE curves in Figure 3 in the sense that the EnKF is the most accurate and that the accuracy of the KEnGMF degrades with increasing values of the bandwidth parameter.

Concerning the estimates of the CEnGMF, on average the lowest AES values are obtained using a mixture component size equal to 4. This reflects more confidence in the resulting conductivity and concentration estimates, which is also in agreement with the results of Fig. 8, for $N = 300$. Nonetheless, for smaller ensemble size (i.e., $N = 100$), the AES estimates of CEnGMF (with $m > 1$) indicate large confidence while in fact the estimates of the states and parameters are not as good as those of the EnKF. In general, the plots suggest that using more Gaussian components decreases the spread of the ensemble and this is not surprising because the data are assimilated into clusters of smaller ensembles. The proposed ACEnGMF, as shown in Fig. 18, produces an AES that is quite comparable to that of the EnKF. One can also notice from the plot that using a smaller threshold value, r , lead to smaller ensemble spreads.

One can further use the AES information to assess filter inbreeding (Hendricks-Franssen and Kinzelbach, 2008). One possible way to check this is to evaluate the ratio of the ensemble skill, i.e., AAE to the AES. For a system that does not suffer from inbreeding, this ratio should be around one (De Lannoy et al., 2006), or in other words, both metrics are of the same order. This is generally true in all our experiments in which the ratio (using $N = 300$) was found to vary between 0.4 during the early assimilation steps and 1 by the end of the assimilation window. This indicates that the assimilation system using both the EnKF and the clustering-GM filters is well behaved and that the ensemble spread did not suffer from any collapse.

5.5. Assessing the non-Gaussianity of the prior pdf

In this section, we analyze the marginal/scalar prior pdf of one of the state variables as estimated by the different GM techniques. The EnKF is implemented with $n = 300$ members. We consider the scalar variable x_i , $i = 20$, which is the concentration of a point in the aquifer, at three different time instants, day 50, day 100 and day 150, for illustration. The true values of x_i and its forecast ensemble members at those three different times are respectively indicated by green squares and gray stars in Figure 19. The panels in the first,

second and third rows correspond to the estimates at days 50, 100 and 150, respectively.

The marginal forecast pdfs of x_i are computed as follows:

$$p_m(x_i) = \int_{x_1} \cdots \int_{x_{i-1}} \int_{x_{i+1}} \cdots \int_{x_{N_x}} p(x) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_{N_x}, \quad (32)$$

where $p(x)$ denotes the prior GM pdf constructed by applying the *k-means* clustering method, or the Gaussian kernel estimation, on the forecast ensemble. The scalar forecast pdfs of x_i , $p_s(x_i)$, is estimated by directly applying the *k-means* clustering method, or the Gaussian kernel estimation, on the forecast ensemble members of x_i . Note that $p_m(x_i)$ and $p_s(x_i)$ are different in the *k-means* clustering case (first and the second rows of Figure 19) and identical in the Gaussian kernel estimation case (the third row of Figure 19). The marginal pdf of x_i with one Gaussian component ($m = 1$) and the scalar pdf of x_i with one Gaussian component are identical. Some useful conclusions can be drawn from this experimental result. First, the GM pdfs constructed using the *k-means* clustering with $m > 1$ or the kernel estimation with $\beta < 1$ better represent the non-Gaussianity. For example, the last rows' panels of Figure 19 clearly show that the distributions of the forecast ensemble members at day 100 and 150 are asymmetrical and one side heavy-tailed. The GM pdfs better capture this pattern. Secondly, when more Gaussian components or smaller bandwidth are used, the resulting GM assigns larger likelihood to the heavily sampled areas. This implies that the GM estimated by the *k-means* clustering and kernel clustering is more consistent with the sampled ensemble than the single Gaussian. Third, the GM pdf with multiple Gaussian components ($m > 1$) or smaller bandwidth ($\beta < 1$) evaluated at the true value of x_i is generally higher than that of the single Gaussian pdf. This means that the GM representation of the prior distribution has smaller estimation error variance than that of the single Gaussian representation.

Compared to the kernel-based scheme, the clustering-based scheme seems to impose higher probability on negative concentration values. This could be related to the fact that the representation of the prior pdf, constructed by the clustering method, is more smooth than

that of the kernel method. This does not contradict the global performance evaluation. During the analysis step, the posterior pdf is obtained by locally updating the prior GM pdf component-by-component using available observations, instead of globally updating the prior pdf. Although the kernel-based EnGMF seems to recover a physically reasonable global representation of the prior pdf, its inability to capture the local probability as discussed earlier generally results in less accurate estimates than the clustered-based EnGMF.

6. Summary and discussion

We considered the Gaussian mixture (GM) approach to construct the prior distribution of the Bayesian filter for the problem of state (concentration)-parameter (conductivity) estimation of a subsurface contamination model. The study focused on the comparison of three different schemes: (i) kernel density estimation, (ii) clustering, and (iii) adaptive clustering, for constructing the prior GM from the forecast ensemble. An ensemble resampled from the posterior GM is then integrated with the dynamical model to propagate the prior probability following the common ensemble Monte Carlo approach. Numerical experiments were conducted to evaluate the behavior of the ensemble GM filter with respect to the different clustering strategies and to compare its performances with respect to the EnKF. Our main conclusions are as follows. The kernel ensemble GM filter (KEnGMF) provides limited performances. This is due to the kernel nature of its GM representation which provides limited ability to well represent the local pdfs of the forecast distribution. The clustered ensemble GM filter (CEnGMF) classifies the forecast ensemble members into fewer Gaussian components, which leads to a better representation of the forecast ensemble by a GM. This improved the GM filter's behavior, which remained, however, sensitive to the number of Gaussian components. The adaptive clustered ensemble GM filter (ACEnGMF) adjusts the number of Gaussian components at every assimilation cycle, while balancing the ability to well capture local nonlinearities using enough numbers of Gaussian components. This further allows to estimate more robust statistics of the GM model by identifying the relevant ensemble members in each Gaussian component. Our numerical results suggest that

the ACEnGMF is robust with respect to the choice of the clustering parameters and leads to the best filtering performances compared to all other tested schemes.

In summary, the results suggest that the ensemble GM filter may provide an efficient framework for state-parameter estimation of subsurface contamination problems. The ensemble GM filter is however strongly dependent on the strategy of fitting the forecast ensemble to a GM, with the best performance obtained using the proposed adaptive clustering approach.

Acknowledgements

The authors would like to thank two anonymous reviewers for their comments and suggestions that highly improved the quality of the paper. Research reported in this publication was supported by King Abdullah University of Science and Technology (KAUST).

References

- T. Belk, J. Smith, J. Trax, Wellhead Protection: A Guide for Small Communities, Chapter 3 “Groundwater Contamination”, US-EPA Seminar Publication, 1993.
- I. Javandel, C. Doughty, C.-F. Tsang, Groundwater transport: Handbook of mathematical models, vol. 10, American Geophysical Union, 1984.
- P. B. Bedient, H. S. Rifai, C. J. Newell, et al., Ground water contamination: transport and remediation., Prentice-Hall International, Inc., 1994.
- J. Bear, A. Cheng, Modeling Groundwater Flow and Contaminant Transport (Theory and Applications of Transport in Porous Media), Springer, 2009.
- M. Gharamti, A. Kadoura, J. Valstar, S. Sun, I. Hoteit, Constraining a compositional flow model with flow-chemical data using an ensemble-based Kalman filter, Water Resources Research 50 (3) (2014a) 2444–2467.

- 536 H. J. Hendricks-Franssen, W. Kinzelbach, Ensemble Kalman filtering versus sequential self-
537 calibration for inverse modeling of dynamic groundwater flow systems, *Journal of Hydrol-*
538 *ogy* 44.
- 539 M. Gharamti, J. Valstar, I. Hoteit, An adaptive hybrid EnKF-OI scheme for efficient state-
540 parameter estimation of reactive contaminant transport models, *Advances in Water Re-*
541 *sources* 71 (0) (2014b) 1 – 15, ISSN 0309-1708.
- 542 F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, R. Karlsson, P.-J.
543 Nordlund, Particle filters for positioning, navigation, and tracking, *IEEE Transactions*
544 *on Signal Processing* 50 (2) (2002) 425 –437, ISSN 1053-587X.
- 545 I. Hoteit, D.-T. Pham, G. Triantafyllou, G. Korres, A New Approximate Solution of the Op-
546 timal Nonlinear Filter for Data Assimilation in Meteorology and Oceanography, *Monthly*
547 *Weather Review* 136 (1) (2008) 317–334.
- 548 P. J. van Leeuwen, Particle Filtering in Geophysical Systems, *Monthly Weather Review*
549 137 (12) (2009) 4089–4114.
- 550 X. Luo, I. Hoteit, I. Moroz, On a nonlinear Kalman filter with simplified divided difference
551 approximation, *Physica D: Nonlinear Phenomena* 241 (6) (2012) 671–680, ISSN 0167-2789.
- 552 J. S. Liu, Monte Carlo Strategies in Scientific Computing, Springer Series in Statistics, 2008.
- 553 S. Chang, T. Chowhan, S. Latif, State and Parameter Estimation with an SIR Particle Filter
554 in a Three-Dimensional Groundwater Pollutant Transport Model, *Journal of Environmen-*
555 *tal Engineering* 138(11) (2012) 1114–1121.
- 556 M. Leisenring, H. Moradkhani, Snow water equivalent prediction using Bayesian data assim-
557 ilation methods, *Stochastic Environmental Research and Risk Assessment* 25 (2) (2011)
558 253–270, ISSN 1436-3240.

- C. Montzka, H. Moradkhani, M. Weihermuller, M. Canty, H. J. Hendricks-Franssen, H. Vereecken, Hydraulic parameter estimation by remotely-sensed top soil moisture observations with the particle filter, *Journal of Hydrology* 399(3-4) (2011) 410–421.
- H. Moradkhani, C. M. DeChant, S. Sorooshian, Evolution of ensemble data assimilation for uncertainty quantification using the particle filter-Markov chain Monte Carlo method, *Water Resources Research* 48.
- I. Hoteit, X. Luo, D.-T. Pham, Particle Kalman Filtering: A Nonlinear Bayesian Framework for Ensemble Kalman Filters, *Monthly Weather Review* 140 (2) (2012) 528–542.
- G. Evensen, Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *Journal of Geophysical Research: Oceans* C5 (99) (1994) 10143–10162.
- G. Burgers, P. Jan van Leeuwen, G. Evensen, Analysis Scheme in the Ensemble Kalman Filter, *Monthly Weather Review* 126 (6) (1998) 1719–1724.
- H. J. Hendricks-Franssen, W. Kinzelbach, Real-time groundwater flow modeling with the Ensemble Kalman Filter: Joint estimation of states and parameters and the filter inbreeding problem, *Water Resources Research* 365(3-4) (2009) 261–274.
- M. E. Gharamti, I. Hoteit, J. Valstar, Dual states estimation of a subsurface flow-transport coupled model using ensemble Kalman filtering, *Advances in Water Resources* 60 (0) (2013) 75 – 88.
- M. Gharamti, B. Ait-El-Fquih, I. Hoteit, An iterative ensemble Kalman filter with one-step-ahead smoothing for state-parameters estimation of contaminant transport models, *Journal of Hydrology* 527 (2015) 442–457.
- G. A. Kivman, Sequential parameter estimation for stochastic systems, *Nonlinear Processes in Geophysics* 10 (3) (2003) 253–259.

- 583 J. L. Anderson, S. L. Anderson, A Monte Carlo Implementation of the Nonlinear Filtering
584 Problem to Produce Ensemble Assimilations and Forecasts, *Monthly Weather Review*
585 127 (12) (1999) 2741–2758.
- 586 D. Alspach, H. Sorenson, Nonlinear Bayesian estimation using Gaussian sum approximations,
587 *IEEE Transactions on Automatic Control* 17 (4) (1972) 439–448.
- 588 A. Y. Sun, A. P. Morris, S. Mohanty, Sequential updating of multimodal hydrogeologic
589 parameter fields using localization and clustering techniques, *Water Resources Research*
590 45 (7) (2009) n/a–n/a, ISSN 1944-7973, w07424.
- 591 T. Sondergaard, P. F. J. Lermusiaux, Data Assimilation with Gaussian Mixture Models
592 Using the Dynamically Orthogonal Field Equations. Part I: Theory and Scheme, *Monthly*
593 *Weather Review* 141 (6) (2013) 1737–1760.
- 594 B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall,
595 New York, 1986.
- 596 I. Hoteit, D.-T. Pham, J. Blum, A simplified reduced order Kalman filtering and application
597 to altimetric data assimilation in Tropical Pacific, *Journal of Marine Systems* 36 (12)
598 (2002) 101–127, ISSN 0924-7963.
- 599 A. S. Stordal, H. A. Karlsen, G. Nævdal, H. J. Skaug, B. Vallès, Bridging the ensemble
600 Kalman filter and particle filters: the adaptive Gaussian mixture filter, *Computational*
601 *Geosciences* 15 (2) (2010) 293–305, ISSN 1420-0597.
- 602 D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley-
603 Interscience, 1992.
- 604 W. K. Härdle, L. Simarx, *Applied Multivariate Statistical Analysis*, Springer, 2011.
- 605 R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, John Wiley & Sons, 2001.

- 606 D. MacKay, Information Theory, Inference and Learning Algorithms, Cambridge University
607 Press, 2003.
- 608 M. Inaba, N. Katoh, H. Imai, Applications of Weighted Voronoi Diagrams and Randomiza-
609 tion to Variance-based K-clustering: (Extended Abstract), in: Proceedings of the Tenth
610 Annual Symposium on Computational Geometry, SCG '94, ACM, New York, NY, USA,
611 332–339, 1994.
- 612 J. MacQueen, Some methods for classification and analysis of multivariate observations, in:
613 Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,
614 Volume 1: Statistics, 281–29, 1967.
- 615 K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate Analysis, Academic Press, 1979.
- 616 C. Goutte, P. Toft, E. Rostrup, F. Nielsen, L. K. Hansen, On Clustering fMRI Time Series,
617 NeuroImage 9 (1996) 298–310.
- 618 S. Salvador, P. Chan, Determining the number of clusters/segments in hierarchical cluster-
619 ing/segmentation algorithms, in: Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th
620 IEEE International Conference on, 576–584, 2004.
- 621 I. Hoteit, D.-T. Pham, M. E. Gharamti, Mitigating observational error undersampling in the
622 stochastic EnKF, Monthly Weather Review In Press.
- 623 J. Gómez-Hernández, A. Journel, Joint sequential simulation of multiGaussian fields, Geo-
624 statistics Troia92 (1993) 85–94.
- 625 E. Simon, L. Bertino, Gaussian anamorphosis extension of the DEnKF for combined state
626 parameter estimation: application to a 1D ocean ecosystem model, Journal of Marine
627 Systems 89 (1) (2012) 1–18.
- 628 L. Li, H. Zhou, J. J. Gómez-Hernández, H. J. Hendricks-Franssen, Jointly mapping hydraulic

conductivity and porosity by assimilating concentration data via ensemble Kalman filter,
Journal of Hydrology 428-429 (2012) 152–169.

Y. Gu, D. S. Oliver, The Ensemble Kalman Filter for Continuous Updating of Reservoir
Simulation Models, Journal of Energy Resources Technology 128 (1) (2005) 79–87.

G. J. De Lannoy, P. R. Houser, V. Pauwels, N. E. Verhoest, Assessment of model uncer-
tainty for soil moisture through ensemble verification, Journal of Geophysical Research:
Atmospheres (1984–2012) 111 (D10).

Table Headings

Table 1: Boundary conditions and parameters of the coupled flow-transport model.

Figure Captions

Figure 1: A sketch diagram of the groundwater flow and contaminant transport configuration with boundary conditions. Major groundwater flow happens from the western to the eastern boundary. Shown inside the domain is the initial solute concentration in mg/L .

Figure 2: Reference hydraulic conductivity, $\mathbf{K}$ (left panel) and solute concentration, C (right panel) fields. The reference conductivity field is sampled from a Gaussian variogram model using the GCOSIM3D code. Black squares and circles denote the locations of the observations used to condition both fields.

Figure 3: Time evolution of the MAEs for the concentration (left panels) and the conductivity (right panels) as they result from the KEnGMF with ensemble sizes $N = 100, 300$ and different values of the bandwidth parameter β . The MAE of the EnKF is shown in blue for benchmarking.

Figure 4: Snapshots of the estimates of the concentration at the end of the assimilation (day 150) as they result from the EnKF and the KEnGMF with 100 ensemble members and different bandwidth parameters β .

Figure 5: Same as Figure 4, but for 300 ensemble members.

Figure 6: Snapshots of the estimates of the conductivity at the end of the assimilation (day 150) as they result from the EnKF and the KEnGMF with 100 ensemble members and different bandwidth parameters β .

Figure 7: Same as Figure 6, but for 300 ensemble members.

Figure 8: Time evolution of the MAEs for the concentration (left panels) and the conductivity (right panels) as they result from the CEnGMF with ensemble sizes $N = 100, 300$ and different number of Gaussian components. The MAE of the EnKF is shown for benchmarking.

Figure 9: Snapshots of the estimates of the concentration at the end of the assimilation (day 150) as they result from the EnKF and the CEnGMF with 100 ensemble members and different GM components number m .

Figure 10: Snapshots of the estimates of the concentration at the end of the assimilation (day 150) as they result from the EnKF and the CEnGMF with 300 ensemble members and different GM components number m .

Figure 11: Snapshots of the estimates of the conductivity at the end of the assimilation (day 150) as they result from the EnKF and the CEnGMF with 100 ensemble members and different GM components number m .

Figure 12: Snapshots of the estimates of the conductivity at the end of the assimilation (day 150) as they result from the EnKF and the CEnGMF with 300 ensemble members and different GM components number m .

Figure 13: Time evolution of the MAEs for the concentration (left panels) and the conductivity (right panels) as they result from the ACEnGMF with ensemble sizes $N = 100, 300$ and different values of the clustering threshold r . The MAE of the EnKF is shown for benchmarking.

Figure 14: Snapshots of the estimates of the concentration at the end of the assimilation (day 150) as they result from the EnKF and the ACEnGMF with different ensemble members ($N = 100, 300$) and different dropping rate thresholds r .

Figure 15: Snapshots of the estimates of the conductivity at the end of the assimilation

(day 150) as they result from the EnKF and the ACEnGMF with different ensemble members ($N = 100, 300$) and different dropping rate thresholds r .

Figure 16: Time evolution of the AESs for the concentration (left panels) and the conductivity (right panels) as they result from the KEnGMF with ensemble sizes $N = 100, 300$ and different values of the bandwidth parameter β . The AES of the EnKF is shown for benchmarking.

Figure 17: Time evolution of the AESs for the concentration (left panels) and the conductivity (right panels) as they result from the CEnGMF with ensemble sizes $N = 100, 300$ and different number of Gaussian components. The AES of the EnKF is shown for benchmarking.

Figure 18: Time evolution of the AESs for the concentration (left panels) and the conductivity (right panels) as they result from the ACEnGMF with ensemble sizes $N = 100, 300$ and different values of the clustering threshold r . The AES of the EnKF is shown for benchmarking.

Figure 19: Prior pdf of concentration variable x_{20} as constructed using the k-means clustering method and the Gaussian kernel method at day 50 (upper panels), day 100 (middle panels), and day 150 (lower panels). The ensemble size is $N = 300$. The number of Gaussian components is denoted by m in the k-means clustering, and the bandwidth parameter is denoted by β in the kernel estimation.

Figure 1

Figure 2

Figure 3

Figure 4

Figure 5

Figure 6

Figure 7

Figure 8

Figure 9

Figure 10

Figure 11

Figure 12

Figure 13

Figure 14

Figure 15

Figure 16

Figure 17

Figure 18

Figure 19

1.

TABLE. 1:

Parameter	Definition	Value
$U_{w,1}$ (m ³ /day)	Velocity at injection well 1	30
$U_{w,2}$ (m ³ /day)	Velocity at injection well 2	30
h_e (m-water)	Hydraulic head at east boundary	0
ϕ (%)	Porosity	25
ρ_b (g/cm ³)	Bulk density	1.22
K_d (ml/g)	Sorption partition coefficient	0.75
κ_L (m)	Longitudinal dispersivity	0.05
κ_T (m)	Transverse dispersivity	0.01
δ_m (m ² /s)	Molecular diffusion	0.0