Accepted Manuscript

An Iterative Ensemble Kalman Filter with One-Step-Ahead Smoothing for State-Parameters Estimation of Contaminant Transport Models

M.E. Gharamti, B. Ait-El-Fquih, I. Hoteit

PII: S0022-1694(15)00339-X

DOI: <http://dx.doi.org/10.1016/j.jhydrol.2015.05.004>

Reference: HYDROL 20439

To appear in: *Journal of Hydrology*

Please cite this article as: Gharamti, M.E., Ait-El-Fquih, B., Hoteit, I., An Iterative Ensemble Kalman Filter with One-Step-Ahead Smoothing for State-Parameters Estimation of Contaminant Transport Models, *Journal of Hydrology* (2015), doi: <http://dx.doi.org/10.1016/j.jhydrol.2015.05.004>

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.

An Iterative Ensemble Kalman Filter with One-Step-Ahead Smoothing for State-Parameters Estimation of Contaminant Transport Models

M.E. Gharamti^{a,b}, B. Ait-El-Fquih^a, I. Hoteit^{a,*}

^a*King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia*

^b*Nansen Environmental and Remote Sensing Center (NERSC), Bergen, Norway*

Abstract

The ensemble Kalman filter (EnKF) is a popular method for state-parameters estimation of subsurface flow and transport models based on field measurements. The common filtering procedure is to directly update the state and parameters as one single vector, which is known as the Joint-EnKF. In this study, we follow the one-step-ahead smoothing formulation of the filtering problem, to derive a new joint-based EnKF which involves a smoothing step of the state between two successive analysis steps. The new state-parameters estimation scheme is derived in a consistent Bayesian filtering framework and results in separate update steps for the state and the parameters. This new algorithm bears strong resemblance with the Dual-EnKF, but unlike the latter which first propagates the state with the model then updates it with the new observation, the proposed scheme starts by an update step, followed by a model integration step. We exploit this new formulation of the joint filtering problem and propose an efficient model-integration-free iterative procedure on the update step of the parameters only for further improved performances.

Numerical experiments are conducted with a two-dimensional synthetic subsurface transport model simulating the migration of a contaminant plume in a heterogenous aquifer domain. Contaminant concentration data are assimilated to estimate both the contaminant state and the hydraulic conductivity field. Assimilation runs are performed under imperfect

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

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

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

modeling conditions and various observational scenarios. Simulation results suggest that the proposed scheme efficiently recovers both the contaminant state and the aquifer conductivity, providing more accurate estimates than the standard Joint and Dual EnKFs in all tested scenarios. Iterating on the update step of the new scheme further enhances the proposed filter's behavior. In term of computational cost, the new Joint-EnKF is almost equivalent to that of the Dual-EnKF, but requires twice more model integrations than the standard Joint-EnKF.

Keywords: State-parameter estimation, Subsurface contaminant transport, Ensemble Kalman filter, One-step-ahead smoothing, Bayesian estimation.

1. Introduction

Subsurface flow and transport models are subject to various input, parametric and structural uncertainties (Abbaspour et al., 1997; Dagan and Neuman, 2005). Thus, it is important to quantify and reduce the uncertainties of these models before using their outputs for decision making (Burris and Antworth, 1992; West and Harwell, 1992; Billings and Billings, 1993). The major sources of uncertainties are mainly associated with the poor characterization of three-dimensional subsurface parameters such as permeability and porosity, and processes such as sorption and biodegradation coefficients (Chen and Zhang, 2006; Bailey et al., 2012; Gharamti and Hoteit, 2014). Sequential and variational data assimilation techniques serve as efficient tools for addressing the uncertainties of such models by incorporating various in-situ measurements of the flow and the contaminant transport (Hendricks Franssen and Kinzelbach, 2008; Bailey and Baù, 2011; Altaf et al., 2013). The ensemble Kalman filter (EnKF) is a sequential technique that became popular in the field of hydrology because of its non-intrusive formulation, robustness, and reasonable computational requirements (Reichle et al., 2002; Vrugt et al., 2006; Zhou et al., 2011; Gharamti et al., 2013).

The EnKF is a Monte Carlo implementation of the traditional Kalman filter (e.g., Kalman, 1960; Gharamti et al., 2011), operating with cycles of forecast and update steps.

In the forecast step, a set (ensemble) of state realizations is integrated with the subsurface model forward in time. The sample mean and covariance of the forecasted ensemble are then used to approximate the Gaussian distribution of the state of the system. In the update step, the EnKF uses available observations and applies a Kalman update to each ensemble member (Evensen, 2003). This ensemble framework enables to account for model errors that are not only present in the uncertain parameters but also in the model structure and inputs, such as external forcings (McLaughlin, 2002). In practice, however, the EnKF is often affected by a deteriorating quality of the analyzed covariance matrix, as it is often implemented with limited ensembles and assuming error-free forecast models. This may cause the filter inbreeding problem in which the ensemble variance is increasingly under-estimated over time (Hendricks Franssen and Kinzelbach, 2008).

The most common strategy for tackling the state-parameters estimation problem in subsurface hydrology and reservoir applications is the joint filtering approach (Chen and Zhang, 2006; Hendricks Franssen and Kinzelbach, 2008; Li et al., 2012; Panzeri et al., 2014). This popular method concurrently estimates the state and the parameters by concatenating all unknown variables into one single state vector. In a reservoir history matching study, Wen and Chen (2006) argued that jointly updating the state and parameters may yield significant inconsistency, especially for strongly heterogeneous subsurface formations. To tackle this, they proposed a confirming step approach in which the updated parameters (permeabilities) are used after the analysis step to rerun the model and obtain a more reliable pressure solution. A similar strategy was proposed by Moradkhani et al. (2005), who used the so-called dual filtering approach to separately update the state and the parameters. The dual strategy employs two interactive and parallel filters (e.g., EnKFs). First, the –dynamically constant – parameters are updated and later used to perform another filtering step for the state. In the same context, Gu and Oliver (2007) presented a restart EnKF scheme for a multiphase fluid flow state-parameters estimation problem in which the flow simulations are repeated from the initial time every time the permeability gets updated. The goal was again to avoid

the physical inconsistencies that might result from the joint filtering approach.

Apart from the joint scheme, the aforementioned techniques are not consistent with the Bayesian filtering problem and they lack a proper mathematical rigour (Hendricks Franssen and Kinzelbach, 2008). In this work, we introduce a new EnKF-based state-parameters estimation approach, namely the One-Step-Ahead Joint-EnKF (called hereafter, Joint-EnKF_{OSA}). The scheme is derived following a one-step-ahead smoothing formulation of the Bayesian filtering problem. We show that this reverses the order of the measurement-update step that usually follows the model forecast step. The proposed Joint-EnKF_{OSA} algorithm involves at each assimilation cycle two Kalman-like update steps using the current observation; one for state smoothing and another for parameters updating. Unlike the Dual-EnKF, which updates the state ensemble after integrating the dynamical model, the proposed scheme uses available data for smoothing the state ensemble before integrating the model forward in time. Computationally, this means that the Joint-EnKF_{OSA} is roughly equivalent to the Dual-EnKF and is twice more expensive than the Joint-EnKF.

Introducing iterations to the filter update step was shown to be beneficial for estimation problems with highly nonlinear observational and dynamical models (Kivman, 2003; Sakov et al., 2012). This is more pronounced for the parameters estimation problem as these are usually indirectly related to the observations through a strongly nonlinear relation (Lorentzen and Nævdal, 2011). The ensemble randomized maximum likelihood (EnRML) introduced by Gu and Oliver (2007) and its variants (e.g., Li et al., 2009; Chen and Oliver, 2012; Sakov et al., 2012) are examples of iterative EnKFs (IEnKFs) that iterate on the parameters using quasi-Newton optimization techniques. Other formulations of the IEnKF iterate on the entire augmented state vector and have been successfully applied to different subsurface state-parameters estimation problems (e.g., Gaoming Li and Reynolds, 2009; Krymskaya et al., 2009; Lorentzen and Nævdal, 2011). The aforementioned techniques are, however, computationally intensive because they require integrating the forward model using the updated state and parameters at each iteration. In a recent study, Song et al. (2014)

developed an efficient iterative EnKF algorithm in which forward model integrations at each iteration are only applied to the mean of the ensemble, instead of all realizations. Here, we also propose, in the context of the Joint-EnKF_{OSA}, an efficient iterative procedure to further enhance the update step of the parameters. The iterations, which are applied only on the parameters, make use of a damping factor to limit the size of the correction term. Such a factor allows to avoid integrating the model for each ensemble member at every iteration, resulting in significant reduction in computational cost.

Numerical experiments are conducted to evaluate the performance of the proposed iterative Joint-EnKF_{OSA} scheme using a two-dimensional contaminant transport model. The model simulates the migration of a radioactive contaminant plume with a steady groundwater flow in a shallow aquifer system for 20 years. The contaminant state of the system and the hydraulic conductivity parameters are estimated at every cell of the domain based on observations of solute concentration. Different assimilation scenarios are tested by changing the ensemble size, the spatial and temporal density of the observations, and also the level of the measurement noise. The performance of the iterative Joint-EnKF_{OSA} scheme is further analyzed using different values of the damping factor.

The rest of the paper is organized as follows. Section 2 formulates the state-parameter estimation problem. The Joint-EnKF_{OSA} is derived and its relation with other state-parameter estimation techniques is also discussed in Section 3. Section 4 presents a conceptual contaminant model and outlines the experimental setup. Numerical results are presented and discussed in Section 5. Conclusions of the work are given in Section 6.

2. Problem formulation

We consider the state-parameter estimation problem of subsurface flow models following a discrete-time dynamical system formulation:

$$\begin{cases} \mathbf{x}_{n+1} &= \mathcal{M}_n(\mathbf{x}_n, \boldsymbol{\theta}) + \boldsymbol{\eta}_n \\ \mathbf{y}_n &= \mathbf{H}_n \mathbf{x}_n + \boldsymbol{\varepsilon}_n \end{cases}, \quad (1)$$

in which $\mathbf{x}_n \in \mathbb{R}^{N_x}$ and $\mathbf{y}_n \in \mathbb{R}^{N_y}$ denote the system state and the observation at time t_n , respectively, and $\boldsymbol{\theta} \in \mathbb{R}^{N_\theta}$ is the parameter vector. $\mathcal{M}_n$ is a nonlinear operator integrating the system from time t_n to t_{n+1} , and $\mathbf{H}_n$ denotes a linear observational operator at time t_n ¹. The model process noise, $\boldsymbol{\eta} = \{\boldsymbol{\eta}_n\}_{n \in \mathbb{N}}$, and the observation process noise, $\boldsymbol{\varepsilon} = \{\boldsymbol{\varepsilon}_n\}_{n \in \mathbb{N}}$, are assumed to be statistically independent, jointly independent and independent of $\mathbf{x}_0$ and $\boldsymbol{\theta}$, which, in turn, are assumed independent. Let also $\boldsymbol{\eta}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_n)$ and $\boldsymbol{\varepsilon}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_n)$. Throughout this paper, $\mathbf{y}_{0:n} \stackrel{\text{def}}{=} \{\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_n\}$, and $p(\mathbf{x}_n)$ and $p(\mathbf{x}_n | \mathbf{y}_{0:l})$ stand for the probability density function (pdf) of $\mathbf{x}_n$ and the pdf of $\mathbf{x}_n$ given $\mathbf{y}_{0:l}$.

We focus on the state-parameter *filtering* problem, i.e. the estimation, at each time t_n , of the state $\mathbf{x}_n$ and the parameters $\boldsymbol{\theta}$ from the historic of the observations $\mathbf{y}_{0:n}$. One solution to this problem is given by the *a posteriori* mean (AM) estimates:

$$\mathbb{E}_{p(\mathbf{x}_n | \mathbf{y}_{0:n})}[\mathbf{x}_n] = \int \mathbf{x}_n p(\mathbf{x}_n, \boldsymbol{\theta} | \mathbf{y}_{0:n}) d\mathbf{x}_n d\boldsymbol{\theta}, \quad (2)$$

$$\mathbb{E}_{p(\boldsymbol{\theta} | \mathbf{y}_{0:n})}[\boldsymbol{\theta}] = \int \boldsymbol{\theta} p(\mathbf{x}_n, \boldsymbol{\theta} | \mathbf{y}_{0:n}) d\mathbf{x}_n d\boldsymbol{\theta}, \quad (3)$$

which minimize the *a posteriori* mean squared error. Analytical computation of (2) and (3) is rarely feasible in practice because of the nonlinear character of the model operator, $\mathcal{M}_n(\mathbf{x}_n, \boldsymbol{\theta})$, in addition to the complex dynamics of subsurface flow models. The Joint-EnKF,

¹The proposed scheme can be easily extended to the more general nonlinear case. Only the term $\mathbf{H}_n \mathbf{x}_n^{f,(m)}$ is replaced by $\mathcal{H}_n(\mathbf{x}_n^{f,(m)})$ in (12).

and later the Dual-EnKF, were proposed as efficient algorithms providing good approximations of (2) and (3) at reasonable computational requirements.

The Joint-EnKF estimates the state and parameters simultaneously by directly computing the analysis pdf of the augmented state $\mathbf{z}_n = [\mathbf{x}_n^T \boldsymbol{\theta}^T]^T$. The Dual-EnKF, on the other hand, is based on a conditional estimation strategy that first computes the (marginal) parameters analysis pdf, before computing the state analysis distribution based on the estimated parameters pdf. Aside from this difference, the Joint-EnKF and Dual-EnKF still share the common framework of proceeding with a time-update step (without using the observations), yielding the forecast pdf, followed by a measurement-update step in which the forecast pdf is updated by the new observation, leading to the analysis pdf of interest. The reader may refer for instance to Moradkhani et al. (2005), Li et al. (2012), Panzeri et al. (2013) and Gharamti et al. (2014a) for detailed discussions on Joint- and Dual-EnKFs.

3. One-step-ahead smoothing-based Joint-EnKF (Joint-EnKF_{OSA})

The classical path that involves the forecast pdf $p(\mathbf{z}_n|\mathbf{y}_{0:n-1})$ when moving from the analysis pdf $p(\mathbf{z}_{n-1}|\mathbf{y}_{0:n-1})$ to the analysis pdf at the next time $p(\mathbf{z}_n|\mathbf{y}_{0:n})$ is not unique. We resort here to an alternative algorithm that involves a one-step-ahead smoothing pdf, $p(\mathbf{z}_{n-1}|\mathbf{y}_{0:n})$, between two successive analysis pdfs, $p(\mathbf{z}_{n-1}|\mathbf{y}_{0:n-1})$ and $p(\mathbf{z}_n|\mathbf{y}_{0:n})$. While the idea of using the one-step-ahead smoothing pdf in the filtering algorithm was already used to derive several Kalman and particle filters in low-dimensional state-space systems (see for example Desbouvries et al. (2011) and Lee and Farmer (2014) and references therein), it has, to the best of our knowledge, never been used in the context of high-dimensional state-parameter estimation problems.

3.1. The generic algorithm

The starting point is to build an augmented state-space system, equivalent to (1), for which the augmented vector, $\mathbf{z}_n \stackrel{\text{def}}{=} [\mathbf{x}_n^T \boldsymbol{\theta}^T]^T$, is the state of interest. One obtains,

$$\begin{cases} \mathbf{z}_{n+1} &= \widetilde{\mathcal{M}}_n(\mathbf{z}_n) + \widetilde{\boldsymbol{\eta}}_n, \\ \mathbf{y}_n &= \widetilde{\mathbf{H}}_n \mathbf{z}_n + \boldsymbol{\varepsilon}_n, \end{cases} \quad (4)$$

where, $\widetilde{\mathcal{M}}_n(\mathbf{z}_n) = \begin{bmatrix} \mathcal{M}_n(\mathbf{z}_n) \\ [\mathbf{0} \ \mathbb{I}] \mathbf{z}_n \end{bmatrix}$, $\widetilde{\boldsymbol{\eta}}_n = [\boldsymbol{\eta}_n^T \mathbf{0}]^T$, $\widetilde{\mathbf{H}}_n = [\mathbf{H}_n \ \mathbf{0}]$, and the matrices $\mathbf{0}$ and $\mathbb{I}$ respectively denote the zero and identity matrices with appropriate dimensions. Such a new system can be seen as a particular (continuous state) Hidden Markov Chain (HMC) with transition density,

$$p(\mathbf{z}_n | \mathbf{z}_{n-1}) = p(\mathbf{x}_n | \mathbf{x}_{n-1}, \boldsymbol{\theta}) = \mathcal{N}_{\mathbf{x}_n}(\mathcal{M}_{n-1}(\mathbf{x}_{n-1}, \boldsymbol{\theta}), \mathbf{Q}_{n-1}), \quad (5)$$

and likelihood,

$$p(\mathbf{y}_n | \mathbf{z}_n) = p(\mathbf{y}_n | \mathbf{x}_n) = \mathcal{N}_{\mathbf{y}_n}(\mathbf{H}_n \mathbf{x}_n, \mathbf{R}_n), \quad (6)$$

where $\mathcal{N}_{\mathbf{v}}(\mathbf{m}, \mathbf{C})$ denotes the Gaussian density of argument $\mathbf{v}$ and parameters $(\mathbf{m}, \mathbf{C})$.

Following this augmented state-space formulation, the analysis pdf, $p(\mathbf{z}_n | \mathbf{y}_{0:n})$, can be computed from $p(\mathbf{z}_{n-1} | \mathbf{y}_{0:n-1})$ following two steps. A measurement-update step (or smoothing step) in which $p(\mathbf{z}_{n-1} | \mathbf{y}_{0:n-1})$ is updated by the new observation $\mathbf{y}_n$, yielding the one-step-ahead smoothing pdf, $p(\mathbf{z}_{n-1} | \mathbf{y}_{0:n})$. This is followed by a time-update step (or analysis step) in which, given the same set of observations $\mathbf{y}_{0:n}$, the pdf of the joint state, $\mathbf{z}_n$, is computed from that of $\mathbf{z}_{n-1}$ at the previous time.

- Smoothing step: $p(\mathbf{x}_{n-1}, \boldsymbol{\theta} | \mathbf{y}_{0:n})$ is first computed using the likelihood $p(\mathbf{y}_n | \mathbf{x}_{n-1}, \boldsymbol{\theta})$ as:

$$p(\mathbf{x}_{n-1}, \boldsymbol{\theta} | \mathbf{y}_{0:n}) \propto p(\mathbf{y}_n | \mathbf{x}_{n-1}, \boldsymbol{\theta}) p(\mathbf{x}_{n-1}, \boldsymbol{\theta} | \mathbf{y}_{0:n-1}), \quad (7)$$

with

$$p(\mathbf{y}_n|\mathbf{x}_{n-1}, \boldsymbol{\theta}) = \int p(\mathbf{y}_n|\mathbf{x}_n)p(\mathbf{x}_n|\mathbf{x}_{n-1}, \boldsymbol{\theta}) d\mathbf{x}_n. \quad (8)$$

- Analysis step: Only the analysis pdf of $\mathbf{x}_n$, $p(\mathbf{x}_n|\mathbf{y}_{0:n})$, needs to be computed from $p(\mathbf{x}_{n-1}, \boldsymbol{\theta}|\mathbf{y}_{0:n})$ since the analysis pdf of $\boldsymbol{\theta}$ has already been computed in the smoothing step. Using the *a posteriori* transition pdf, $p(\mathbf{x}_n|\mathbf{x}_{n-1}, \boldsymbol{\theta}, \mathbf{y}_n)$, one obtains:

$$p(\mathbf{x}_n|\mathbf{y}_{0:n}) = \int p(\mathbf{x}_n|\mathbf{x}_{n-1}, \boldsymbol{\theta}, \mathbf{y}_n) p(\mathbf{x}_{n-1}, \boldsymbol{\theta}|\mathbf{y}_{0:n}) d\mathbf{x}_{n-1} d\boldsymbol{\theta}, \quad (9)$$

with

$$p(\mathbf{x}_n|\mathbf{x}_{n-1}, \boldsymbol{\theta}, \mathbf{y}_n) \propto p(\mathbf{y}_n|\mathbf{x}_n) p(\mathbf{x}_n|\mathbf{x}_{n-1}, \boldsymbol{\theta}). \quad (10)$$

Unlike the standard Joint-EnKF which applies one update to the state using the observations (originating from a Bayes' formula), the smoothing-based algorithm described above updates the state twice following the Bayes's formulas (7) and (10).

3.2. Ensemble implementation

Based on the Gaussian assumption of $p(\mathbf{x}_{n-1}, \boldsymbol{\theta}|\mathbf{y}_{0:n})$, $p(\mathbf{x}_n, \boldsymbol{\theta}|\mathbf{y}_{0:n})$ and $p(\mathbf{y}_n|\mathbf{y}_{0:n-1})$, for all n , we derive an EnKF-like approximation using the random sampling properties included in the Appendix. Hereafter, let for an ensemble $\{\mathbf{r}^{(m)}\}_{m=1}^{N_e}$, $\hat{\mathbf{r}}$ denotes its empirical mean, and $\mathbf{S}_{\mathbf{r}}$ denotes the ensemble analysis with the m^{th} column defined as $(\mathbf{r}^{(m)} - \hat{\mathbf{r}})$.

3.2.1. Smoothing step

In the following, $\boldsymbol{\theta}_{|n}^{(m)}$ denote the m^{th} member sampled from $p(\boldsymbol{\theta}|\mathbf{y}_{0:n})$ and the members $\mathbf{x}_n^{f,(m)}$, $\mathbf{x}_n^{a,(m)}$ and $\mathbf{x}_n^{s,(m)}$ correspond to the forecast, analysis and smoothing pdfs of $\mathbf{x}_n$, respectively. Starting at time t_{n-1} from an analysis ensemble of size N_e , $\{\mathbf{x}_{n-1}^{a,(m)}, \boldsymbol{\theta}_{|n-1}^{(m)}\}_{m=1}^{N_e}$, approximating $p(\mathbf{z}_{n-1}|\mathbf{y}_{0:n-1})$, one can use Property 1 of the Appendix in (8) to first obtain the state forecast ensemble $\{\mathbf{x}_n^{f,(m)}\}_{m=1}^{N_e}$, then the observation forecast ensemble $\{\mathbf{y}_n^{f,(m)}\}_{m=1}^{N_e}$,

194 approximating $p(\mathbf{x}_n|\mathbf{y}_{0:n-1})$ and $p(\mathbf{y}_n|\mathbf{y}_{0:n-1})$, respectively:

$$\mathbf{x}_n^{f,(m)} = \mathcal{M}_{n-1}(\mathbf{x}_{n-1}^{a,(m)}, \boldsymbol{\theta}_{|n-1}^{(m)}) + \mathbf{u}_{n-1}^{(m)}, \quad (11)$$

$$\mathbf{y}_n^{f,(m)} = \mathbf{H}_n \mathbf{x}_n^{f,(m)} + \mathbf{v}_n^{(m)}, \quad (12)$$

195 with $\mathbf{u}_{n-1}^{(m)} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_{n-1})$ and $\mathbf{v}_n^{(m)} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_n)$. Property 2 of the Appendix is then used
 196 in (7) to obtain the smoothing ensemble $\{\mathbf{x}_{n-1}^{s,(m)}, \boldsymbol{\theta}_{|n}^{(m)}\}_{m=1}^{N_e}$, that approximates $p(\mathbf{z}_{n-1}|\mathbf{y}_{0:n})$,
 197 as:

$$\mathbf{x}_{n-1}^{s,(m)} = \mathbf{x}_{n-1}^{a,(m)} + \mathbf{P}_{\mathbf{x}_{n-1}^a, \mathbf{y}_n^f} \underbrace{\mathbf{P}_{\mathbf{y}_n^f}^{-1} (\mathbf{y}_n - \mathbf{y}_n^{f,(m)})}_{\nu_n^{(m)}}, \quad (13)$$

$$\boldsymbol{\theta}_{|n}^{(m)} = \boldsymbol{\theta}_{|n-1}^{(m)} + \mathbf{P}_{\boldsymbol{\theta}_{|n-1}, \mathbf{y}_n^f} \times \nu_n^{(m)}. \quad (14)$$

198 with $\mathbf{P}_{\mathbf{x}_{n-1}^a, \mathbf{y}_n^f} = \frac{1}{N_e-1} \mathbf{S}_{\mathbf{x}_{n-1}^a} \mathbf{S}_{\mathbf{y}_n^f}^T$, $\mathbf{P}_{\mathbf{y}_n^f} = \frac{1}{N_e-1} \mathbf{S}_{\mathbf{y}_n^f} \mathbf{S}_{\mathbf{y}_n^f}^T$ and $\mathbf{P}_{\boldsymbol{\theta}_{|n-1}, \mathbf{y}_n^f} = \frac{1}{N_e-1} \mathbf{S}_{\boldsymbol{\theta}_{|n-1}} \mathbf{S}_{\mathbf{y}_n^f}^T$.

199 3.2.2. Analysis step

200 The analysis ensemble $\{\mathbf{x}_n^{a,(m)}\}_{m=1}^{N_e}$ is computed from $\{\mathbf{x}_{n-1}^{s,(m)}, \boldsymbol{\theta}_{|n}^{(m)}\}_{m=1}^{N_e}$ using Property
 201 1 in (9) after computing the *a posteriori* transition pdf $p(\mathbf{x}_n|\mathbf{x}_{n-1}, \boldsymbol{\theta}, \mathbf{y}_n)$ via (10). Eq. (10)
 202 leads to a Gaussian pdf:

$$p(\mathbf{x}_n|\mathbf{x}_{n-1}, \boldsymbol{\theta}, \mathbf{y}_n) = \mathcal{N}_{\mathbf{x}_n}(\mathcal{M}_{n-1}(\mathbf{x}_{n-1}, \boldsymbol{\theta}) + \tilde{\mathbf{K}}_n (\mathbf{y}_n - \mathbf{H}_n \mathcal{M}_{n-1}(\mathbf{x}_{n-1}, \boldsymbol{\theta})), \tilde{\mathbf{Q}}_{n-1}), \quad (15)$$

203 with $\tilde{\mathbf{K}}_n = \mathbf{Q}_{n-1} \mathbf{H}_n^T [\mathbf{H}_n \mathbf{Q}_{n-1} \mathbf{H}_n^T + \mathbf{R}_n]^{-1}$ and $\tilde{\mathbf{Q}}_{n-1} = \mathbf{Q}_{n-1} - \tilde{\mathbf{K}}_n \mathbf{H}_n \mathbf{Q}_{n-1}$. The computa-
 204 tion of $\tilde{\mathbf{K}}_n$ and $\tilde{\mathbf{Q}}_{n-1}$ (which may be an off-diagonal matrix even when $\mathbf{Q}_{n-1}$ is diagonal) can
 205 be prohibitive in large dimensions. One way to avoid (10) (and thus (15)), then (9), is to
 206 follow Smidl and Quinn (2006, 2008) and assume $\mathbf{x}_n$ and $\mathbf{y}_n$ independent conditionally on
 207 $(\mathbf{x}_{n-1}, \boldsymbol{\theta})$, i.e.

$$p(\mathbf{x}_n|\mathbf{x}_{n-1}, \boldsymbol{\theta}, \mathbf{y}_n) = p(\mathbf{x}_n|\mathbf{x}_{n-1}, \boldsymbol{\theta}). \quad (16)$$

This reasonable assumption, which does not break the dependence between $\mathbf{x}_n$ and $\mathbf{y}_n$, allows to eliminate the “gain” $\tilde{\mathbf{K}}_n$ from (15). In this case, one can see that the parameters of the Gaussian *a posteriori* transition pdf (15) reduce to those of the Gaussian (a priori) transition pdf $p(\mathbf{x}_n|\mathbf{x}_{n-1}, \boldsymbol{\theta})$ (5). Finally, under the assumption (16), one can use Property 1 in (9) to obtain the analysis ensemble, $\left\{\mathbf{x}_n^{a,(m)}\right\}_{m=1}^{N_e}$, from $\left\{\mathbf{x}_{n-1}^{s,(m)}, \boldsymbol{\theta}_{|n}^{(m)}\right\}_{m=1}^{N_e}$, as,

$$\mathbf{x}_n^{a,(m)} = \mathcal{M}_{n-1} \left(\mathbf{x}_{n-1}^{s,(m)}, \boldsymbol{\theta}_{|n}^{(m)} \right) + \mathbf{u}_{n-1}^{(m)}, \quad (17)$$

with $\mathbf{u}_{n-1}^{(m)} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_{n-1})$. Notice that the absence of a Kalman-like update in the analysis step (17) is due to the fact that the Bayes’ formula (10) has not been used but, in fact, replaced by the assumption (16). The algorithm is summarized in the following section.

3.2.3. The Joint-EnKF_{OSA} algorithm

Starting from an analysis ensemble, $\left\{\mathbf{x}_{n-1}^{a,(m)}, \boldsymbol{\theta}_{|n-1}^{(m)}\right\}_{m=1}^{N_e}$ at time t_{n-1} , the updated ensemble at time t_n is obtained in two steps:

- Smoothing step: The state forecast ensemble $\left\{\mathbf{x}_n^{f,(m)}\right\}_{m=1}^{N_e}$ is first computed as in (11), by integrating $\left\{\mathbf{x}_{n-1}^{a,(m)}, \boldsymbol{\theta}_{|n-1}^{(m)}\right\}_{m=1}^{N_e}$ with the model (1). The observation forecast ensemble $\left\{\mathbf{y}_n^{f,(m)}\right\}_{m=1}^{N_e}$ is then computed using (12), followed by the smoothing state ensemble $\left\{\mathbf{x}_{n-1}^{s,(m)}\right\}_{m=1}^{N_e}$, and parameters $\left\{\boldsymbol{\theta}_{|n}^{(m)}\right\}_{m=1}^{N_e}$, using (13) and (14), respectively.
- Analysis step: The analysis state ensemble $\left\{\mathbf{x}_n^{a,(m)}\right\}_{m=1}^{N_e}$ is obtained by integrating the members $\left\{\mathbf{x}_{n-1}^{s,(m)}, \boldsymbol{\theta}_{|n}^{(m)}\right\}_{m=1}^{N_e}$ with the model using (17).

The proposed Joint-EnKF_{OSA} computes the parameter members, $\boldsymbol{\theta}_{|n}^{(m)}$, following exactly the same formula as in the standard Joint- and Dual-EnKFs. The main difference between the Joint-EnKF and the Joint-EnKF_{OSA} in computing $\mathbf{x}_n^{a,(m)}$ lies in the fact that the first algorithm involves $\mathbf{x}_{n-1}^{a,(m)}$ and $\boldsymbol{\theta}_{|n-1}^{(m)}$, while the second algorithm uses $\mathbf{x}_{n-1}^{s,(m)}$ and $\boldsymbol{\theta}_{|n}^{(m)}$, which indeed are $\mathbf{x}_{n-1}^{a,(m)}$ and $\boldsymbol{\theta}_{|n-1}^{(m)}$, respectively, after the update with $\mathbf{y}_n$. On the other hand, to see how the Joint-EnKF_{OSA} differs from the Dual-EnKF in the calculation of the state members,

231 $\mathbf{x}_n^{a,(m)}$, let us focus on their update equations² (see for example (Samuel et al., 2014) and
 232 (Gharamti et al., 2014a) for (18), and Eqs. (13)-(17) for (19)):

$$\mathbf{x}_n^{a,(m)} \stackrel{\text{Dual-EnKF}}{=} \mathcal{M}_{n-1} \left(\mathbf{x}_{n-1}^{a,(m)}, \boldsymbol{\theta}_{|n}^{(m)} \right) + \overbrace{\mathbf{P}_{\mathbf{x}_n^f} \mathbf{H}_n^T \times \mu_n^{(m)}}^{\text{correction term}}, \quad (18)$$

$$\mathbf{x}_n^{a,(m)} \stackrel{\text{Joint-EnKF}_{\text{OSA}}}{=} \mathcal{M}_{n-1} \left(\underbrace{\mathbf{x}_{n-1}^{a,(m)} + \mathbf{P}_{\mathbf{x}_{n-1}^a, \mathbf{y}_n^f} \times \nu_n^{(m)}}_{\mathbf{x}_{n-1}^{s,(m)}}, \boldsymbol{\theta}_{|n}^{(m)} \right). \quad (19)$$

233 Given the parameter members $\boldsymbol{\theta}_{|n}^{(m)}$, both algorithms compute $\mathbf{x}_n^{a,(m)}$ from $\mathbf{x}_{n-1}^{a,(m)}$ using the
 234 model and a correction term based on $\mathbf{y}_n$. The Dual-EnKF integrates first the members
 235 $(\mathbf{x}_{n-1}^{a,(m)}, \boldsymbol{\theta}_{|n}^{(m)})$ with the model before updating them with the observation. In contrast, the
 236 Joint-EnKF_{OSA} updates the members $\mathbf{x}_{n-1}^{a,(m)}$ first which provides the smoothing samples,
 237 $\mathbf{x}_{n-1}^{s,(m)}$, before they are integrated with the model using the members $\boldsymbol{\theta}_{|n}^{(m)}$.

238 3.3. An iterative Joint-EnKF_{OSA} algorithm (Joint-IEnKF_{OSA})

239 Iterating on the parameters update-step was shown to improve the accuracy of the esti-
 240 mates (Lorentzen and Nævdal, 2011), particularly for strongly nonlinear estimation problems
 241 (Gu and Oliver, 2007; Sakov et al., 2012). In this section, we will exploit the separation of
 242 the state and parameters estimation process in the Joint-EnKF_{OSA} to derive an iterative
 243 version of the Joint-EnKF_{OSA} (called Joint-IEnKF_{OSA}) acting only on the parameters, and
 244 not on the state. This is motivated by the observation that the EnKF works well for the
 245 state estimation of subsurface flow and transport models, which usually evolve in quasi-linear
 246 regimes, especially when linear observation systems are considered (Hoteit et al., 2005). This
 247 is generally not true for parameters estimation, as these are often related to the observed
 248 state through a strongly nonlinear relation (Jardak et al., 2010). In our Joint-IEnKF_{OSA}
 249 scheme, the iterations are performed only in the smoothing step since the analysis step in-

²For simplicity, we ignore the term of the process noise, $\boldsymbol{\eta}_n$. Such a case is widely considered in geophysical applications.

250 volves only the state. The ensemble members $\mathbf{x}_n^{f,(m)}$, $\mathbf{y}_n^{f,(m)}$ and $\mathbf{x}_{n-1}^{s,(m)}$ are computed using
 251 Eqs. (11), (12) and (13), respectively. The iterative procedure of the analysis ensemble of
 252 the parameters can be summarized as follows:

- 253 • *Initialization*: Initial value for each member m can be computed using Eq. (14):

$$\boldsymbol{\theta}_{|n}^{0,(m)} = \boldsymbol{\theta}_{|n-1}^{(m)} + \mathbf{P}_{\boldsymbol{\theta}_{|n-1}, \mathbf{y}_n^f} \times \nu_n^{(m)}. \quad (20)$$

- 254 • *Iterations*: Starting from $\{\boldsymbol{\theta}_{|n}^{0,(m)}\}_{m=1}^{N_e}$, each iteration, ℓ , is performed as:

$$\boldsymbol{\theta}_{|n}^{\ell,(m)} = \boldsymbol{\theta}_{|n}^{\ell-1,(m)} + \omega \cdot \mathbf{P}_{\boldsymbol{\theta}_{|n}^{\ell-1}, \mathbf{y}_n^f} \times \nu_n^{(m)}, \quad \ell = 1, 2, \dots, N_i \quad (21)$$

255 where ω is a damping factor and N_i is the total number of iterations.

256 Equation (21) resembles the iterative procedure of Lorentzen and Nævdal (2011), except
 257 that we assume here that the change in the innovation term after each iteration is small,
 258 and thus no model integrations are longer required. This is generally the case when the
 259 measurements are collected at a high temporal frequency so the change in the parameters
 260 ensemble would be too small to impose large changes on the innovation. The forecasted
 261 observations would thus be very similar. In the case when the observations are not collected
 262 frequently enough, this assumption may become less valid, especially for highly nonlinear
 263 and rapidly changing dynamics. Subsurface contaminants usually involve slowly varying
 264 dynamics, and thus any changes in the parameters members during iterations are expected
 265 to have small impact on the innovation term, $\nu_n^{(m)}$. Under this assumption, the computational
 266 burden is significantly reduced. Moreover, in order to avoid overweighting the observations,
 267 we follow Hendricks Franssen and Kinzelbach (2008) and introduce a damping factor taking
 268 values between 0 and 1 ($0 < \omega < 1$) to limit the size of the correction term after each

iteration. We apply the following stopping criterion:

$$\frac{1}{N_x} \sum_{i=1}^{N_x} \underbrace{\sum_{m=1}^{N_e} |\boldsymbol{\theta}_{i,n}^{\ell,(m)} - \boldsymbol{\theta}_{i,n}^{\ell-1,(m)}|}_{\rho_i} < \kappa, \quad (22)$$

where i is the grid cell number and κ is a predefined tolerance value, based on which the parameters iterations were shown to converge after only few iterations in our experiments. The average of the residuals, ρ_i , measures the change of the ensemble members at each point in space through iterations.

3.4. Computational burden

The computational load of the discussed state-parameters EnKF estimation schemes can be split according to their time-update and measurement-update steps. The Joint-EnKF, for instance, requires N_e runs of the forecast model and N_e Kalman-type update steps for each joint state-parameter ensemble member. In most geoscience applications, the computational complexity for updating the ensemble members is generally small compared to integrating the forecast model (Gharamti et al., 2013). The dual algorithm is computationally more intensive, requiring $2N_e$ model runs for forecasting, and $2N_e$ Kalman corrections for updating both the forecast state and parameter ensembles. This is practically double the computational cost of the joint scheme. As for the proposed Joint-EnKF_{OSA} scheme, it requires $2N_e$ model runs for forecasting the state, N_e smoothing steps for the state and N_e updates to the parameter ensemble. Accordingly, the complexity of the Joint-EnKF_{OSA} is equivalent to that of the Dual-EnKF, except that the order of the update equations is reversed. For the iterative version of the proposed scheme, only few additional update steps are applied while iterating on the parameters. Since no additional model runs are required, iterations generally only impose a marginal increase in the computational cost. The approximate computational complexity and memory storage are summarized for each scheme in Table 1. The tabulated complexities are based on the assumption that the number of observations N_y is

much smaller than the dimension of the state N_x . Based on this assumption, which is always valid in subsurface hydrology problems, only the leading and most demanding algebraic operations are considered.

4. Numerical experiments with a contaminant transport subsurface system

4.1. Contaminant transport model

We consider a steady-state groundwater flow system inside a rectangular domain of total aquifer area 42000 m². North and south boundaries are assumed impermeable, whereas the west and east boundaries are assigned constant hydraulic heads equal to 18 and 12 m-water, respectively. The following two-dimensional (2D) saturated Darcy groundwater flow system is solved (Bear, 2012):

$$\frac{\partial}{\partial x} \left(K_x \frac{\partial h}{\partial x} \right) + \frac{\partial}{\partial y} \left(K_y \frac{\partial h}{\partial y} \right) = q, \quad (23)$$

where K_x and K_y are the hydraulic conductivity components in x and y directions [LT⁻¹], respectively. h is the hydraulic head [L] and q denotes the sources and sinks [T⁻¹]. The groundwater flow equation (23) is discretized in space into 60 × 20 grid cells of 7 m × 5 m using a cell-centered finite difference (CCFD) scheme. A sketch of the aquifer domain with boundary conditions and corresponding flow streamlines is shown in Figure 1. We ignore precipitation and infiltration sources, so that $q = 0$. Pure water conditions are assumed initially in the aquifer, except for an elongated ¹³⁷Cs plume of concentration 10 mg/l located near the west boundary. ¹³⁷Cs is a radioactive isotope of Caesium and a fission product with a half-life of 30.17 years. It is generally used for industrial purposes in gauges to measure liquid flows and thickness of materials. It possesses a unique combination of physical properties and historical notoriety, and is chemically reactive and highly soluble (Wallo et al., 1994). Its half-life is long enough that contaminated objects and regions may remain dangerous to humans for generations (Bunting, 1975). We simulate the migration of the radioactive plume across the domain towards the eastern boundary for a 20-years period using the following

316 reactive transport model (Bear, 2012; Gharamti et al., 2014b):

$$\mathcal{R} \left[\frac{\partial (\phi C)}{\partial t} + \lambda \phi C \right] = \phi \nabla \cdot \left[(\kappa_l |\mathbf{U}_x| + \kappa_t |\mathbf{U}_y| + \delta_m) \nabla C \right] - \nabla \cdot (\mathbf{U}C), \quad (24)$$

317 where C is the dissolved solute concentration [ML^{-3}], t is time [T], ϕ is the porosity of the
 318 subsurface media, and λ is first-order radioactive decay rate [T^{-1}]. κ_l and κ_t are, respectively,
 319 the longitudinal and transversal pore-scale dispersivity, both constants and equal to 0.5 and
 320 0.001 m. $\mathbf{U}_x$ and $\mathbf{U}_y$ are the Darcy velocity components in the xy plane and δ_m is the
 321 effective molecular diffusion, considered here equal to $5 \times 10^{-10} \text{ m}^2/\text{s}$. $\mathcal{R}$ is a retardation
 322 factor assuming linear sorption conditions in a medium textured sandy-silty soil.

323 The reference log-conductivity and porosity fields are generated independently using
 324 the sequential Gaussian simulation algorithm GCOSIM3D (Gómez-Hernández and Journel,
 325 1993). The variogram components of each random function are listed in Table 2. The pa-
 326 rameters of these aquifer properties are commonly observed field values and similar to those
 327 adopted in many other hydrological studies (e.g., Zhou et al., 2011; Li et al., 2012; Gharamti
 328 et al., 2014a). The correlation lengths, chosen for these parameters, in both longitudinal
 329 and transversal directions are small compared to the dimension of the domain. This can be
 330 clearly observed in the spatial maps of the two heterogeneous fields in Figure 2. Higher con-
 331 ductivity zones (red areas) located close to the western, top, and bottom boundaries suggest
 332 faster migration of the contaminant plume. This is also suggested by the orientation and
 333 the size of the streamlines in Figure 1. The transport is expected to occur at a slower pace
 334 in the vicinity of the small conductivity regions in the center of the domain. We discretize
 335 the time evolution of the contaminant using the classical fourth-order Runge-Kutta scheme
 336 with a time step of 15 days. We run the simulation and plot four different snapshots of the
 337 contaminant state in Figure 3. As seen from the maps, the ^{137}Cs plume tends to expand to-
 338 wards the eastern boundary, which is mostly due to dispersion and diffusion. The maximum
 339 concentration after 20 years is depleted to almost 0.80 mg/l because of radioactivity.

4.2. Experimental Setup

To imitate a realistic setting, we impose some perturbations on the reference model and set our goal to estimate the solute concentration and the hydraulic conductivity field using the perturbed forecast model and data extracted from the reference (true) run. Thus, in the forecast model we impose 15%, 20%, and 10% Gaussian noise on the sorption coefficient, diffusion rate, and radioactive decay, respectively. The longitudinal and transversal dispersivities, which could be major sources of uncertainty in realistic transport models, are also perturbed with 25% Gaussian noise. In addition, we further perturb the porosity of the aquifer medium using the same variogram parameters for ϕ (as shown in Table 2) but with mean and variance equal to 0.30 and 0.0032, respectively.

To demonstrate the efficiency of the proposed Joint-EnKF_{OSA} compared to the standard Joint- and Dual-EnKFs, we evaluate the filter's performance and robustness under different experimental scenarios. We conduct a number of sensitivity experiments while changing: (1) the ensemble size, (2) the temporal frequency of observations, (3) the number of observation wells in the domain, and (4) the measurement error. For the frequency of the observations, we consider 5 scenarios in which concentrations measurements are extracted from the reference run every 0.5, 5, 15, 30, 60, and 120 months. We consider four different observational scenarios assuming networks of 3, 10, 15, and 33 wells uniformly distributed throughout the aquifer domain (Figure 3). We further evaluate the performance of the filters under nine different scenarios in which observations are perturbed with Gaussian noise of zero mean and a standard deviation varying between 5% and 25%, with 5% increments of the total variability. We select these scenarios to study the performances of the different filters under challenging and imperfect modeling and observational conditions. This is typical in real-world hydrology applications, where wells data are sparse and subject to various sources of uncertainty, such as reading errors, instrumental errors, etc.

To initialize the filters, we sample the initial concentration ensemble from a long model run starting from a perturbed initial condition (different than the reference concentration

distribution in the top-panel of Figure 3) and using perturbed model parameters. Thus, we first perform a 50 years simulation run using the perturbed forecast model starting from the mean concentration state (i.e., time-average state) of the reference run snapshots. Conducting such a long spin-up period helps capturing most of the dynamic variations in the system. Using the output of this simulation, we save the concentration fields every 3 days. The total number of snapshots retained from this spin-up run is 6000. Then, we randomly select a set of N_e contaminant snapshots to form the initial state ensemble. By doing so, the concentration changes that may occur in the aquifer are well represented by the initial ensemble. The corresponding parameters' realizations are sampled with the geostatistical software, GCOSIM3D, using the same variogram parameters of the reference conductivity field but conditioned on three hard measurements as shown by black squares in Figure 2. A filter estimates are evaluated based on their average absolute errors (AAE) and their average ensemble spread (AES):

$$AAE = N_x^{-1} N_e^{-1} \sum_{j=1}^{N_e} \sum_{i=1}^{N_x} \left| \mathbf{x}_{j,i}^e - \mathbf{x}_i^t \right|, \quad (25)$$

$$AES = N_x^{-1} N_e^{-1} \sum_{j=1}^{N_e} \sum_{i=1}^{N_x} \left| \mathbf{x}_{j,i}^e - \hat{\mathbf{x}}_i^e \right|, \quad (26)$$

where $\mathbf{x}_i^t$ is the reference (true) value of the variable at cell i , $\mathbf{x}_{j,i}^e$ is the ensemble value of the variable, and $\hat{\mathbf{x}}_i^e$ is the ensemble mean at location i . AAE measures the estimate-truth misfit and AES measures the ensemble spread, or the confidence in the estimated values (Hendricks Franssen and Kinzelbach, 2008). Spatially, the estimated fields are assessed based on their correlation factors (γ) with the reference solution. Linear regression is also applied to statistically evaluate the conductivity estimates as:

$$\frac{1}{N_e} \sum_{i=1}^{N_e} \log(K_i^e) = \alpha + \beta \cdot \log(K^r), \quad (27)$$

where K_i^e is the i^{th} conductivity ensemble realization, K^r is the reference conductivity, and

α and β are regression coefficients.

5. Results and Discussion

This section analyzes and compares the performance of the Joint-EnKF_{OSA} against those of the Joint-EnKF and the Dual-EnKF. We further assess the benefit of the proposed iterative procedure for improving the update of the conductivity field.

5.1. Sensitivity to the ensemble size

We first study the sensitivity of the three filtering schemes to the ensemble size, N_e . In realistic and large-scale contaminant transport applications, one would be restricted by computational resources to using small ensembles. Computing accurate state and parameter estimates with small ensembles is thus desirable. We conduct the experiments using 6 ensemble sizes, $N_e = \{50, 100, 250, 500, 1000, 5000\}$. We consider 33 wells (Figure 3, bottom observation network) and set the standard deviation of the measurement error to 15%. Concentration observations are assumed available at every time step; i.e., every 15 days. We plot the bias and spread of the resulting conductivity and state estimates in Figure 4. As shown in the upper-left panel of Figure 4, the accuracy of the conductivity estimates resulting from the Joint-EnKF, Dual-EnKF and Joint-EnKF_{OSA} improves as the ensemble size increases, reaching a mean AAE of 0.5690, 0.5592, and 0.5321 m/s for $N_e = 5000$, respectively. Using 1000 members, the Dual-EnKF clearly outperforms the standard Joint-EnKF. Overall, the proposed Joint-EnKF_{OSA} provides the best estimates for ensemble sizes larger than 50. As the ensemble size increases beyond 500, the mean AAE resulting from the Joint-EnKF_{OSA} levels off at about 0.54 m/s. Regarding the average ensemble spread of the conductivity estimates (upper-right panel of Figure 4), the Joint-EnKF_{OSA} retains the largest spread. As for the contaminant estimates, the mean AAE values in the bottom-left panel of Figure 4 exhibit similar results to those of the conductivity estimates. The contaminant estimates improve with increasing ensemble size.

Overall, the Joint-EnKF_{OSA} is the most accurate, always outperforming the Joint- and Dual-EnKFs as the ensemble size increases from 50 to 5000. The associated average spread of the contaminant ensemble (bottom-right panel of Figure 4) is the largest in terms of AES.

Comparing the curves of the bias and the ensemble spread for both state and parameters, one may notice that the ensemble variance is underestimated for $N_e < 500$. This is an indication of the filter-inbreeding problem, which usually occurs when the filter is implemented with small ensemble sizes. As shown in Figure 4, however, for $N_e \geq 500$ the bias and spread of the contaminant estimates are of the same order of magnitude, with a ratio close to 1. For the contaminant concentration, one can see that ensemble sizes larger than 500 are needed to avoid filter-inbreeding. Hereafter, we use 500 members to conduct the sensitivity experiments, which seem to provide satisfactory results at reasonable computational cost.

Regarding the computational cost³, the Joint-EnKF is the least intensive, requiring 33.07 minutes to perform a 20-year assimilation run using 500 members. The Dual-EnKF and Joint-EnKF_{OSA}, on the other hand, require 65.40 and 64.01 minutes, respectively. This is expected because both include an additional propagation step of the ensemble members as discussed in Section 3.4. The small difference in the cost between the Joint-EnKF_{OSA} and the Dual-EnKF is due to the convergence of the transport model during the forecast step.

5.2. Sensitivity to the frequency of observations

In the second set of experiments, we test the sensitivity of the filters to the frequency of assimilating concentration observations. We set the ensemble size to 500 and assimilate data from 33 observation wells perturbed with a Gaussian noise set to 10% of the signal total variance.

Figure 5 plots the mean AAE of the contaminant concentration as it results from the three filtering schemes for six different observation frequencies, 0.5, 5, 15, 30, 60, and 120 months. Overall, the Dual- and Joint-EnKF schemes lead to similar results, although the

³Our assimilation results were obtained using a 2.66 GHz MAC workstation and 10 cores for parallel looping while integrating the ensemble members.

Dual-EnKF performs better when data are assimilated more frequently, i.e., every half and five months. The performance of the proposed Joint-EnKF_{OSA} is slightly better for small observation frequencies. The most pronounced improvements are obtained when assimilating data every 15 days. This however becomes less important as the time-frequency of available observations decreases. This could be related to the nature of the Joint-EnKF_{OSA} algorithm, which is based on applying a one-step-ahead smoothing step to the analyzed concentration ensemble members. Assimilating data every 120 months ($\equiv$ 10 years), the three schemes have almost similar results, as only two update steps are applied in this scenario.

Figure 6 plots the concentration ensemble mean as estimated by the three filters after 20 years of simulation for three different observation frequency scenarios; i.e, 0.5, 5, and 15 months. Spatially, the maps are close to the reference concentration (shown in the bottom panel of Figure 3). To illustrate, when assimilating data every 15 months, the best correlation with the true concentration field results from the Joint-EnKF_{OSA} with $\gamma = 0.90$. This reduces to 0.88 and 0.89 for the Joint- and the Dual-EnKFs, respectively. In general, all three filters are able to well capture the major contamination patterns concentrated close to the eastern boundary.

Regarding the estimates of the hydraulic conductivity, we compare the results of the three filters by analyzing the time series of AAE (Figure 7) for two observation frequency scenarios. When concentration observations are assimilated every five months (i.e., a total of 48 updates), the Joint-EnKF_{OSA} performs better than the other two schemes over the entire assimilation window. By the end of the 20th year, the conductivity estimates of the proposed scheme are respectively 27% and 13% more accurate than those of the standard Joint- and Dual-EnKFs. Most of the improvements resulting from the Joint-EnKF_{OSA} scheme take place after 10 years. In contrast, the AAE of the dual and joint EnKFs levels off after 10 years. Furthermore, one can clearly notice the better performance of the Dual-EnKF when compared to that of the Joint-EnKF. When assimilating data every 30 months, the proposed scheme only slightly outperforms the other two schemes, yielding more accurate

conductivity mapping by the end of the simulation window. These results demonstrate the effectiveness and the robustness of the Joint-EnKF_{OSA} with respect to the temporal sparsity of the observations, compared with the dual and joint filtering strategies.

To spatially examine the conductivity estimates, we plot in Figure 8 the recovered fields from the Joint-EnKF and the Joint-EnKF_{OSA} at three different times. We show results from the case where observations are assimilated every 5 months. We also report the change in the correlation coefficient values resulting from the three schemes in Figure 9. The maps of the Dual-EnKF are not shown as these were found very close to those of the Joint-EnKF. After 10 years, the major spatial patterns of the hydraulic conductivity are well estimated by the two algorithms. Yet, the proposed scheme leads to larger correlations with the reference field, especially in the low conductivity region located in the center of the domain. By the end of the assimilation window, the high conductivity region located near the bottom-east boundary (Figure 2) is very well recovered by the Joint-EnKF_{OSA}. The Joint-EnKF clearly underestimates the conductivity in this region. As can be seen from the time series plot of the correlation, the proposed scheme better delineates the conductivity patterns by the end of the assimilation period, resulting in a higher correlation factor $\gamma_{(20 \text{ years})} \approx 0.79$. In general, the benefits from the proposed smoothing-based scheme seem to be more pronounced for the estimation of the parameters.

5.3. Sensitivity to the number of observations and measurement errors

We further study the sensitivity of the Joint-EnKF_{OSA} to the number of observation wells as compared to the Joint- and Dual-EnKFs. We use three different observation networks in which concentration measurements are collected from 3, 15, and 33 wells. The observations are perturbed with Gaussian noise larger than the one used in the previous experiments and equal to 20% of signal total variance. We establish a linear relation between the reference log-conductivity field and the ensemble mean from each filter at the end of the assimilation window, to evaluate the conductivity estimates. We interpret the results for all scenarios using 2D scatter-plots as shown in Figure 10. When observations are assimilated from 3 wells

only, the scatter plots resulting from all three schemes exhibit poor correlations between the estimated and the reference fields. The orientation of the fitting line is far from the best fit. Ideally, a good fit should have α close to 0 and β close to 1. As more observations become available, the orientation of the fitting line improves for all schemes. For instance, assimilating concentration observations from 33 wells further reduces the misfit and provides a good distribution of the scattered conductivity points for all filters. The Joint-EnKF_{OSA} performs best, yielding regression parameters that are closest to the ideal ones ($\alpha = 0.54$ and $\beta = -4.80$). The fitting, as observed even with 33 data points, is not perfect and this is due to the nature of the conductivity field, which is dominated by small scale structures. Different experiments (not shown) were tested using heterogenous reference conductivity fields with large-scale patterns, and resulted in almost exact fitting. As in the previous experiments, the Joint and Dual-EnKFs provide close, but less accurate, results. The analysis of the estimates of the contaminant concentrations lead to similar conclusions.

The number of available observation wells would generally affect not only the accuracy of the estimates but also the uncertainty of the recovered fields. To elaborate on this, we plot in Figure 11 the ensemble variance maps of the 500 log-conductivity realizations as they result from the Joint-EnKF_{OSA}, at the end of the 20 years. When data are not assimilated (upper panel), a typical bull's eye shape is observed with zero variance at the sampled three locations and increasing variance away from these locations. The uncertainty of the estimated conductivity begins to decrease as more contaminant data are assimilated. To this extent, moving from the 3 to the 33 data points scenario, the maximum ensemble variance of the conductivity inside the aquifer is reduced from 0.1 to 0.004 (m/s)². The variance resulting from the other Joint- and Dual-EnKF schemes is quite comparable.

In the last set of sensitivity experiments, we fix the number of wells to 33, the observation frequency to 5 months, and we perturb the observations with different measurement errors. We plot the results for five different observation noise scenarios in Figure 12 and compare the bias of the conductivity as estimated by the Joint-EnKF, Dual-EnKF and the

Joint-EnKF_{OSA} with 500 members. In general, the performance of all three filters degrades as the noise level in the observations increases. The Joint- and the Dual-EnKFs perform poorly for conductivity estimates computed with very small observational errors. This, unexpected behavior, could be due to the large perturbations imposed on the log-conductivity field (Hendricks Franssen and Kinzelbach, 2008). This might also be a result of the nonlinear relation between the contaminant concentrations and the hydraulic conductivities and spurious numerical covariances (sampling errors). We note that the proposed scheme does not suffer from this problem. In contrast, its state and parameters estimates become more accurate as the observations noise decrease. The accuracy of the proposed scheme is always better than the other two filters, suggesting more robustness to observation noise. The improved accuracy of the Joint-EnKF_{OSA} for contaminant estimation is more pronounced than that of the parameters.

5.4. Effect of iterations on the Joint-EnKF_{OSA}

To assess the benefits of the proposed iterative scheme on the Joint-EnKF_{OSA} behavior, we first run sensitivity experiments while varying the value of the damping factor, ω , between 0.05 and 0.5. We select a challenging assimilation scenario in which data are available at only 3 wells with 20% noise. We set the ensemble size to 500, the tolerance criterion κ in equation (22) to 0.1, and considered different frequencies of available observations in time. Figure 13 plots the mean AAE of conductivity and contaminant concentration resulting from the Joint-IEnKF_{OSA} as compared to the estimates of the non-iterative scheme. As can be seen, the Joint-IEnKF_{OSA} leads to more accurate state and parameters estimates for all tested damping factor values, with the largest improvements obtained when assimilating concentration every 5 months. These improvements become less pronounced as the frequency of observations decreases, especially for the estimates of the conductivity. This behavior, i.e., the dependency of the results to the frequency of the collected data, is probably due to our assumption of constant innovation throughout iterations. Nevertheless, the Joint-IEnKF_{OSA} still outperforms the Joint-EnKF_{OSA} even in the scenario with sparse data. It is also worth

mentioning that the Joint-IEnKF_{OSA} incurs only a marginal increase in the computational cost compared to the non-iterative scheme as convergence always occurred after only 3 or 4 iterations in all tested assimilation steps. The effects of iterating on the parameters are also clear in the estimates of the state. With very sparse data (30 months frequency), the Joint-IEnKF_{OSA} leads to about 16% more accurate concentration estimates, with $\omega = 0.2$, compared to the estimates of the non-iterative scheme. This is due to the improved parameters estimates, which are used as inputs to the forecast transport model. For both state and parameters, the Joint-IEnKF_{OSA} exhibits smaller ensemble spreads, suggesting larger confidence in the estimated variables. Concerning the choice of the damping factor, we have tested values larger than 0.5 and the results were less accurate. In general, best performances of the iterative scheme were obtained using ω values close to 0.10.

In order to better understand the effects of iterations, we plot in Figure 14 the spatial maps of the conductivity and concentration at the end of the assimilation window as it results from the Joint-EnKF_{OSA} with and without iterations. We choose the dense observation scenario (5 months) and select the best case for the Joint-IEnKF_{OSA}, obtained using $\omega = 0.08$. The conductivity field estimated using the iterative scheme provides more details of the high and low conductivity regions. In contrast, no major hydrologic patterns are identified using the non-iterative scheme. Likewise, the concentration map resulting from the Joint-IEnKF_{OSA} correlates better with the reference solution in Figure 3. The concentration field obtained using the Joint-EnKF_{OSA} obviously underestimates the true concentration located close to the eastern boundary. On average, the iterative scheme yields around 10% more accurate estimates of the conductivity and concentration fields.

Finally, we tested the sensitivity of the two schemes to changing number of data points in the domain. Given an observations frequency of 5 months, we plot in Figure 15 the time series of AAE of conductivity using the Joint-EnKF_{OSA} and the Joint-IEnKF_{OSA} (for $\omega = 0.08$) assuming 3, 15, and 33 data points available for assimilation. From the plot, one can notice that as the observation networks become more dense, the performance of the

non-iterative scheme becomes as accurate as the iterative scheme. That is, as the number of observation wells increases, the effect of iterations on the parameters decreases. This is reasonable because with more data, the conductivity realizations are quickly adjusted and thus the expected improvements via iterations become smaller. A similar behavior was observed for the estimates of the concentration.

6. Conclusion

We presented a smoothing-based joint ensemble Kalman filter (Joint-EnKF_{OSA}) for state-parameter estimation of subsurface contaminant transport models. While sharing with the standard Joint-EnKF the idea of concatenating the state and parameters in one single vector, the new filter reverses the order of the time-update step (forecast by the model) and the measurement-update step (correction with incoming observations). This imposes one more update step to the state without violating the general Bayesian estimation framework, and was shown to provide improved estimates in our numerical experiments without any increase in the computational cost when compared to the traditional Dual-EnKF. We also proposed an iterative version of the Joint-EnKF_{OSA} scheme, a.k.a. Joint-IEnKF_{OSA}, for improve performances. The iterations are applied on the parameters only and make use of a damping factor to avoid the need of integrating the ensemble with the model during iterations, which significantly reduces the computational requirements of the iterative procedure.

We tested the proposed Joint-EnKF_{OSA} scheme on a conceptual contaminant transport state-parameter estimation problem in which we estimated the contaminant concentration and spatially variant conductivity parameters. We conducted a number of sensitivity experiments to evaluate the accuracy and the robustness of the proposed scheme and to compare its performance against the standard Joint- and Dual-EnKF. Our experimental results suggest that the Joint-EnKF_{OSA} is more robust with respect to different assimilation setups and was successfully able to estimate the contaminant concentration and the conductivity field. Sensitivity analyses suggest that the Joint-EnKF_{OSA} is robust and outperforms the standard

Joint- and Dual-EnKF schemes under various experimental conditions and scenarios. The iterative Joint-EnKF_{OSA} further enhanced the filter behavior, especially in the cases with sparse observation networks, at a marginal increase in computational cost. The damping factor, ω , introduced in the context of the iterative scheme has been arbitrarily chosen between 0 and 1 in our experiments. In general, selecting small values was found less risky than large values to avoid over-fitting and unrealistic parameters values. Optimizing the value of ω is a challenging problem. One possible approach is to examine the parameters' ensemble between two consecutive iterations and choose ω that maximizes the reduction in the estimated uncertainties; i.e., ensemble spread. This may require evaluating the Kullback-Leibler criterion for the parameters distribution during iterations (Gharamti et al., 2014b), which could be computationally very demanding.

The proposed Joint-EnKF_{OSA} scheme was proven efficient and robust in this preliminary study under different experimental settings and scenarios. Future work will consider applying this technique to realistic subsurface contaminant applications. This would require including various hyper parameters as part of the estimation problem (e.g., observational error variance), which we will also consider in our future work for efficient implementation of the state-parameters filtering problem.

7. Appendix

The ensemble implementation of the proposed one-step-ahead joint filtering approach was derived based on the following two random sampling properties:

Property 1. *Hierarchical sampling (Robert, 2007)*

Assuming that one can sample from $p(\mathbf{x}_1)$ and $p(\mathbf{x}_2|\mathbf{x}_1)$, then a sample, $\mathbf{x}_2^*$, from $p(\mathbf{x}_2)$ can be drawn as follows:

1. $\mathbf{x}_1^* \sim p(\mathbf{x}_1)$;
2. $\mathbf{x}_2^* \sim p(\mathbf{x}_2|\mathbf{x}_1^*)$.

Property 2. *Conditional sampling (Hoffman and Ribak, 1991)*

Consider a Gaussian pdf, $p(\mathbf{x}, \mathbf{y})$, with $\mathbf{P}_y$ and $\mathbf{P}_{xy}$ denoting the covariance of $\mathbf{y}$ and the cross-covariance of $\mathbf{x}$ and $\mathbf{y}$, respectively. Then a sample, $\mathbf{x}^*$, from $p(\mathbf{x}|\mathbf{y})$, can be drawn as follows:

1. $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \sim p(\mathbf{x}, \mathbf{y})$;
2. $\mathbf{x}^* = \tilde{\mathbf{x}} + \mathbf{P}_{xy}\mathbf{P}_y^{-1}[\mathbf{y} - \tilde{\mathbf{y}}]$.

Acknowledgements

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

References

- Abbaspour, K., Genuchten, M. v., Schulin, R., Schläppi, E., 1997. A sequential uncertainty domain inverse procedure for estimating subsurface flow and transport parameters. *Water Resources Research* 33 (8), 1879–1892.
- Altaf, M., El Gharamti, M., Heemink, A., Hoteit, I., 2013. A reduced adjoint approach to variational data assimilation. *Computer Methods in Applied Mechanics and Engineering* 254, 1–13.
- Bailey, R., Baù, D., 2011. Estimating spatially-variable first-order rate constants in groundwater reactive transport systems. *Journal of contaminant hydrology* 122 (1), 104–121.
- Bailey, R. T., Bau, D. A., Gates, T. K., 2012. Estimating spatially-variable rate constants of denitrification in irrigated agricultural groundwater systems using an ensemble smoother. *Journal of Hydrology* 468, 188–202.
- Bear, J., 2012. *Hydraulics of groundwater*. Courier Dover Publications.
- Billings, G. K., Billings, J. F., Jun. 22 1993. Subsurface contaminant remediation, biodegradation and extraction methods and apparatuses. US Patent 5,221,159.
- Bunting, R., 1975. Nuclear data sheets for $a=137$. *Nuclear Data Sheets* 15 (3), 335–369.
- Burris, D. R., Antworth, C. P., 1992. In situ modification of an aquifer material by a cationic surfactant to enhance retardation of organic contaminants. *Journal of Contaminant Hydrology* 10 (4), 325–337.
- Chen, Y., Oliver, D. S., 2012. Ensemble randomized maximum likelihood method as an iterative ensemble smoother. *Mathematical Geosciences* 44 (1), 1–26.
- Chen, Y., Zhang, D., 2006. Data assimilation for transient flow in geologic formations via ensemble Kalman filter. *Advances in Water Resources* 29 (8), 1107–1122.

- Dagan, G., Neuman, S. P., 2005. Subsurface flow and transport: a stochastic approach. Cambridge University Press.
- Desbouvries, F., Petetin, Y., Ait-El-Fquih, B., 2011. Direct, Prediction- and Smoothing-based Kalman and Particle Filter Algorithms. *Signal Processing* 91 (8), 2064–2077.
- Evensen, G., 2003. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics* 53 (4), 343–367.
- Gaoming Li, S., Reynolds, A. C., 2009. Iterative ensemble kalman filters for data assimilation. *SPE Journal* 109808, 496–505.
- Gharamti, M., Hoteit, I., 2014. Complex step-based low-rank extended Kalman filtering for state-parameter estimation in subsurface transport models. *Journal of Hydrology* 509, 588–600.
- Gharamti, M., Hoteit, I., Sun, S., 2011. Low-rank Kalman filtering for efficient state estimation of subsurface advective contaminant transport models. *Journal of Environmental Engineering* 138 (4), 446–457.
- Gharamti, M., Kadoura, A., Valstar, J., Sun, S., Hoteit, I., 2014a. Constraining a compositional flow model with flow-chemical data using an ensemble-based kalman filter. *Water Resources Research* 50 (3), 2444–2467.
- Gharamti, M., Valstar, J., Hoteit, I., 2014b. An adaptive hybrid EnKF-OI scheme for efficient state-parameter estimation of reactive contaminant transport models. *Advances in Water Resources* 71, 1–15.
- Gharamti, M. E., Hoteit, I., Valstar, J., 2013. Dual states estimation of a subsurface flow-transport coupled model using ensemble Kalman filtering. *Advances in Water Resources* 60, 75–88.

- Gómez-Hernández, J. J., Journel, A. G., 1993. Joint sequential simulation of multigaussian fields. In: *Geostatistics Troia92*. Springer, pp. 85–94.
- Gu, Y., Oliver, D. S., 2007. An iterative ensemble Kalman filter for multiphase fluid flow data assimilation. *Spe Journal* 12 (04), 438–446.
- Hendricks Franssen, H., Kinzelbach, W., 2008. Real-time groundwater flow modeling with the ensemble Kalman filter: Joint estimation of states and parameters and the filter inbreeding problem. *Water Resources Research* 44 (9).
- Hoffman, Y., Ribak, E., 1991. Constrained realizations of Gaussian fields - a simple algorithm. *The Astrophysical Journal* 380 (491), L5–L8.
- Hoteit, I., Korres, G., Triantafyllou, G., 2005. Comparison of extended and ensemble based kalman filters with low and high resolution primitive equation ocean models. *Nonlinear Processes in Geophysics* 12 (5), 755–765.
- Jardak, M., Navon, I., Zupanski, M., 2010. Comparison of sequential data assimilation methods for the kuramoto–sivashinsky equation. *International journal for numerical methods in fluids* 62 (4), 374–402.
- Kalman, R. E., 1960. A new approach to linear filtering and prediction problems. *Journal of basic Engineering* 82 (1), 35–45.
- Kivman, G., 2003. Sequential parameter estimation for stochastic systems. *Nonlinear Processes in Geophysics* 10 (3), 253–259.
- Krymskaya, M., Hanea, R., Verlaan, M., 2009. An iterative ensemble kalman filter for reservoir engineering applications. *Computational Geosciences* 13 (2), 235–244.
- Lee, W., Farmer, C., 2014. Data Assimilation by Conditioning of Driving Noise on Future Observations. *IEEE Transactions on Signal Processing* 62 (15), 3887–96.

- Li, G., Reynolds, A. C., et al., 2009. Iterative ensemble Kalman filters for data assimilation. *Spe Journal* 14 (03), 496–505.
- Li, L., Zhou, H., Gómez-Hernández, J. J., Hendricks Franssen, H.-J., 2012. Jointly mapping hydraulic conductivity and porosity by assimilating concentration data via ensemble Kalman filter. *Journal of Hydrology* 428, 152–169.
- Lorentzen, R. J., Nævdal, G., 2011. An iterative ensemble Kalman filter. *Automatic Control, IEEE Transactions on* 56 (8), 1990–1995.
- McLaughlin, D., 2002. An integrated approach to hydrologic data assimilation: interpolation, smoothing, and filtering. *Advances in Water Resources* 25 (8), 1275–1286.
- Moradkhani, H., Sorooshian, S., Gupta, H. V., Houser, P. R., 2005. Dual state–parameter estimation of hydrological models using ensemble Kalman filter. *Advances in Water Resources* 28 (2), 135–147.
- Panzeri, M., Riva, M., Guadagnini, A., Neuman, S., 2013. Data assimilation and parameter estimation via ensemble kalman filter coupled with stochastic moment equations of transient groundwater flow. *Water Resources Research* 49 (3), 1334–1344.
- Panzeri, M., Riva, M., Guadagnini, A., Neuman, S., 2014. Comparison of ensemble kalman filter groundwater-data assimilation methods based on stochastic moment equations and monte carlo simulation. *Advances in Water Resources* 66, 8–18.
- Reichle, R. H., McLaughlin, D. B., Entekhabi, D., 2002. Hydrologic data assimilation with the ensemble Kalman filter. *Monthly Weather Review* 130 (1), 103–114.
- Robert, C., 2007. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Science & Business Media, New York.
- Sakov, P., Oliver, D. S., Bertino, L., 2012. An iterative EnKF for strongly nonlinear systems. *Monthly Weather Review* 140 (6), 1988–2004.

- Samuel, J., Coulibaly, P., Dumedah, G., Moradkhani, H., 2014. Assessing model state and forecasts variation in hydrologic data assimilation. *Journal of Hydrology* 513, 127–141.
- Smidl, V., Quinn, A., 2006. *The Variational Bayes Method in Signal Processing*. Springer.
- Smidl, V., Quinn, A., 2008. Variational Bayesian Filtering. *IEEE Transactions on Signal Processing* 56 (10), 5020–5030.
- Song, X., Shi, L., Ye, M., Yang, J., Navon, I. M., 2014. Numerical comparison of iterative ensemble Kalman filters for unsaturated flow inverse modeling. *Vadose Zone Journal* 13 (2).
- Vrugt, J. A., Gupta, H. V., Nualláin, B., Bouten, W., 2006. Real-time data assimilation for operational ensemble streamflow forecasting. *Journal of Hydrometeorology* 7 (3), 548–565.
- Wallo, I., Moscovitch, M., Rodgers, J., Duffey, Soares, C., 1994. Investigations of natural variations of Cesium-137 concentrations in residual Soils. The Health Physics Society, 39th Annual Meeting.
- Wen, X.-H., Chen, W. H., 2006. Real-time reservoir model updating using ensemble kalman filter with confirming option. *Spe Journal* 11 (04), 431–442.
- West, C. C., Harwell, J. H., 1992. Surfactants and subsurface remediation. *Environmental Science & Technology* 26 (12), 2324–2330.
- Zhou, H., Gómez-Hernández, J. J., Hendricks Franssen, H.-J., Li, L., 2011. An approach to handling non-gaussianity of parameters and state variables in ensemble Kalman filtering. *Advances in Water Resources* 34 (7), 844–864.

Table Headings

Table 1: Approximate computational complexities of the Joint-EnKF, the Dual-EnKF, the Joint-EnKF_{OSA}, and the Joint-IEnKF_{OSA} algorithms. Notations are as follows. N_x : number of state variables, N_θ : number of parameter variables, N_y : number of observations, N_i : number of iterations, N_e : number of ensemble realizations, N : number of assimilation cycles, $\mathcal{C}_x$: state model cost ($= N_x^2$ is the linear KF), $\mathcal{C}_\theta$: parameter model cost (usually free $\equiv$ identity), $\mathcal{C}_y$: observation operator cost ($= N_y N_x$ in the linear KF), $\mathcal{S}_x$: storage volume for one state vector, $\mathcal{S}_\theta$: storage volume for one parameter vector.

Table 2: Parameters of the random functions for modeling the spatial distributions of $\log K$ and ϕ . τ denotes the rotation angle of one clockwise rotation around the positive y-axis.

Figure Captions

Figure 1: Plan view of the conceptual model for the 2D groundwater flow system. East and west boundaries are Dirichlet with a given prescribed hydraulic heads. North and south boundaries are impermeable (no flow boundaries). The black arrows are streamlines showing the major flow direction of the groundwater inside the aquifer domain.

Figure 2: The reference log-conductivity field obtained using the sequential Gaussian simulation code (Gómez-Hernández and Journel, 1993). We use a Gaussian variogram with a mean of $-11.01 \log(\text{m/s})$, variance of $1.04 \log(\text{m}^2/\text{s}^2)$, and range equal to 40 m and 20 m in the x and y directions, respectively. The black squares on the map correspond to hard data points used to condition the sampled conductivity realizations for assimilation. Shown in the bottom panel is the porosity field ($\times 100$) obtained using a Gaussian variogram model.

Figure 3: Reference contaminant maps at the initial time (top), after 5 years (second from top), after 10 years (second from bottom), and after 20 years of simulation (bottom). The well locations from which concentration data are extracted are shown by black circles. Four different observation networks are shown.

Figure 4: Average bias and ensemble spread of the log-conductivity and contaminant estimates as function of the ensemble size. Results are shown for a scenario in which assimilation of concentration data is obtained from 33 wells every half a month. The ensemble size varies between 50 and 5000 and the measurement noise standard deviation is set to 15% of the total contaminant variance. The x-axis is displayed in log-scale.

Figure 5: Mean average absolute errors (AAE) of contaminant concentration, C , obtained using three different filtering schemes. Results are shown for six different scenarios in

which assimilation of solute concentration data are obtained from 33 wells. The labels on the x-axis refer to a modeling step which is equal to 15 days in real time. All six experimental scenarios use 500 ensemble members and 10% as standard deviation for the Gaussian measurement errors.

Figure 6: Spatial maps of the recovered ensemble means of contaminant concentration (after 20 years) using the Joint-, Dual-, and Joint-EnKF_{OSA} schemes. Results are shown for three scenarios in which assimilation of concentration data is obtained from 33 wells every 0.5, 5, and 15 months. This experiment uses 500 ensemble members and 10% as standard deviation for the Gaussian measurement errors.

Figure 7: Time series of AAE of the log-conductivity using the Joint-, Dual- and Joint-EnKF_{OSA} schemes. Results are shown for two scenarios in which assimilation of concentration data are obtained from 33 wells (uniformly distributed throughout the aquifer domain) every 5 and 15 months. The two experimental scenarios use 500 ensemble members and 10% as standard deviation for the Gaussian measurement errors.

Figure 8: Spatial maps of the recovered ensemble means of log-conductivity after 1, 10, and 20 years using the Joint-EnKF and the proposed Joint-EnKF_{OSA}. Results are shown for an assimilation scenario in which concentration data is obtained from 33 wells every 5 months. This experiment uses 500 ensemble members and 10% as standard deviation for the Gaussian measurement errors.

Figure 9: Time series of the correlation coefficients of the estimated conductivity, obtained using the three filters, with the reference conductivity field. Results are shown for an assimilation scenario in which concentration data is obtained from 33 wells every 5 months. This experiment uses 500 ensemble members and 10% as standard deviation for the Gaussian measurement errors.

Figure 10: Scatter plots of hydraulic conductivity for different observation networks using the Joint-EnKF, Dual-EnKF and Joint-EnKF_{OSA} schemes. Red lines indicate the regression fitting lines and dashed blue lines indicate the best fit. The displayed hydraulic conductivity values are in log m/s. X-axis labels represent the reference field, whereas, the y-axis ones correspond to the estimated conductivity. Results are shown for three scenarios in which 3, 15, and 33 data points are assimilated into the transport model. The experiments use 500 ensemble members, 5 months of observation frequency, and 20% as standard deviation for the Gaussian measurement errors.

Figure 11: Ensemble variance maps of log-conductivity obtained using the Joint-EnKF_{OSA} scheme after 20 years-period. Four different scenarios for well data assimilation are displayed. Each scenario uses 500 ensemble members, 5 months of observation frequency, and 20% as standard deviation for the Gaussian measurement noise.

Figure 12: Average bias of the log-conductivity and concentration estimates as function of the measurement error. Assimilation of concentration data is obtained from 33 wells every 5 months. The ensemble size is set to 500 and the measurement noise level varies between 5% to 25% of the total contaminant variance.

Figure 13: Mean AAE using the Joint-IEnKF_{OSA} for different values of the damping factor, ω , and different observation frequencies. Also shown, with dashed lines, are the mean AAE of log-conductivity using the the non-iterative Joint-EnKF_{OSA} scheme. Results are displayed using 500 ensemble members, 3 data points, and 20% of observation noise.

Figure 14: Spatial maps of log-conductivity and contaminant concentration obtained using the Joint-EnKF_{OSA} and Joint-IEnKF_{OSA} schemes. The ensemble size used is 500, observation frequency is 5 months, the number of observation wells is 3, and the observation noise is 20%. The damping factor for the Joint-IEnKF_{OSA} is 0.08.

Figure 15: Time series of AAE of the log-conductivity using the Joint-EnKF_{OSA} and Joint-IEnKF_{OSA} schemes. Results are shown for three scenarios in which assimilation of concentration data are obtained from 3, 15, and 33 wells every 5 months. These experimental scenarios use 500 ensemble members and 20% as standard deviation for the Gaussian measurement errors. The damping factor for the Joint-IEnKF_{OSA} is 0.08.

Algorithm	Time-update	Measurement-update	Storage
<i>Joint-EnKF</i>	$NN_e(\mathcal{C}_x + \mathcal{C}_\theta)$	$NN_e(\mathcal{C}_y + N_y N_\theta) + NN_e^2(N_x + N_\theta)$	$2NN_e(\mathcal{S}_x + \mathcal{S}_\theta)$
<i>Dual-EnKF</i>	$NN_e(2\mathcal{C}_x + \mathcal{C}_\theta)$	$NN_e\mathcal{C}_y + NN_e^2(N_x + N_\theta)$	$2NN_e(\mathcal{S}_x + \mathcal{S}_\theta)$
<i>Joint-EnKF_{OSA}</i>	$NN_e(2\mathcal{C}_x + \mathcal{C}_\theta)$	$NN_e\mathcal{C}_y + NN_e^2(N_x + N_\theta)$	$2NN_e(\mathcal{S}_x + \mathcal{S}_\theta)$
<i>Joint-IEnKF_{OSA}</i>	$NN_e(2\mathcal{C}_x + \mathcal{C}_\theta)$	$NN_e\mathcal{C}_y + NN_e^2(N_x + N_i N_\theta)$	$2NN_e(\mathcal{S}_x + \mathcal{S}_\theta)$

	Mean	Variance	Variogram	Range $[x]$	Range $[y]$	Angle, τ
$\log K$	-11.01 log(m/s)	1.04 log(m/s) ²	Gaussian	40 m	20 m	45°
ϕ	0.20	0.0016	Gaussian	40 m	20 m	45°

Conductivity Estimates (Bias) **Conductivity Estimates (Spread)** **State Estimates (Bias)** **State Estimates (Spread)**

15 days

5 months

15 months

Initial Ensemble Mean **1 year** **10 years** **20 years**

Correlation with the True Conductivity

3 Data Points **15 Data Points** **33 Data Points**

No Data Points **3 Data Points** **15 Data Points** **33 Data Points**

Conductivity Estimates **Contaminant Estimates**

Joint-EnKF-OSA, Conductivity

Joint-EnKF-OSA, Contaminant

Joint-IEnKF-OSA, Conductivity

Joint-IEnKF-OSA, Contaminant

- We introduce a one-step-ahead smoothing joint ensemble Kalman filter (Joint-EnKF_{OSA}).
- We propose an efficient iterative analysis scheme for the parameters, Joint-IEnKF_{OSA}.
- We implement the new algorithm for a state-parameters estimation problem.
- Joint-EnKF_{OSA} provides more accurate estimates than standard Joint- and Dual-EnKFs.
- Joint-IEnKF_{OSA} improves performances at a marginal increase in computational cost.