A comparison between ES-MDA and restart EnKF for the purpose of the simultaneous identification of a contaminant source and hydraulic conductivity

Teng Xu[a,b,c], J. Jaime Gómez-Hernández[c,*], Zi Chen[c], Chunhui Lu[a,b,*]

[a] *State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing, China*
[b] *Yangtze Institute for Conservation and Development, Hohai University, Nanjing, China*
[c] *Institute of Water and Environmental Engineering, Universitat Politècnica de València, Valencia, Spain*

Abstract

Understanding a contaminant source may help in a better management and risk assessment of a polluted aquifer. However, contaminant source information may not be available when a pollutant is detected in a drinking well. The restart ensemble Kalman filter (restart EnKF, also named r-EnKF) has been demonstrated in synthetic and laboratory experiments as an efficient solution for the identification of a contaminant source. Recently, the ensemble smoother with multiple data assimilation (ES-MDA) has been proposed as an alternative to the r-EnKF as a more efficient solution given that the r-EnKF needs to restart the simulation of the state equation from time zero after each data assimilation step. An analysis, in a synthetic aquifer, of the accuracy of the ES-MDA for the simultaneous identification of a contaminant source and the spatial distribution of hydraulic conductivity by assimilating both piezometric head and concentration observations is carried out using the r-EnKF as a benchmark. The conclusion is that the ES-MDA can outperform the r-EnKF, but the expected speed advantage, associated with the possibility of assimilating all data at once, does not exist. For the ES-MDA to reach the same level of accuracy as the r-EnKF, the

*Corresponding author
Email addresses: `teng.xu@hhu.edu.cn` (Teng Xu), `jgomez@upv.es` (J. Jaime Gómez-Hernández), `dtpenguincz@gmail.com` (Zi Chen), `clu@hhu.edu.cn` (Chunhui Lu)

number of multiple data assimilations must be large, and final computing time is similar for both approaches. However, the ES-MDA can do much better than the r-EnKF if the number of iterations increases even further, with the consequent increase of computational cost.
Keywords: Contaminant source identification; Data assimilation; Ensemble smoother with multiple data assimilation; Restart ensemble Kalman filter

1. Introduction

When a contaminant is released into the subsurface, it will jeopardize not only human health but also damage the local ecosphere, especially if the contaminant is hazardous. When contamination happens inadvertently or is purposely hidden, it may be difficult to trace it back from concentration observations taken downstream from the source. Yet, knowledge of the contaminant source is vital for groundwater contamination management, contamination control, contamination risk assessment and remediation.

How to identify a contaminant source once contamination has been detected has attracted much attention in the last decades. It is an intricate problem that has been addressed using inverse modeling. According to their characteristics, the inverse modeling approaches for contaminant source identification could be classified into three categories: optimization approaches, probabilistic approaches, and deterministic approaches. The reader is referred to the reviews by Sun et al. (2006a); Atmadja and Bagtzoglou (2001b); Michalak and Kitanidis (2004); Bagtzoglou and Atmadja (2005) for further information.

In the optimization approaches, the objective is to minimize an objective function that measures the differences between simulated concentrations and measurement observations and that is written in terms of the parameters defining the contaminant source. Some of the approaches used are least-squares regression and linear programming (Gorelick et al., 1983), maximization of correlation coefficients (Sidauruk et al., 1998), constrained robust least squares (CRLS) (Sun et al., 2006a), CRLS estimation combined with a branch-andbound global optimization (Sun et al., 2006b), evolutionary search algorithms (Mirghani et al., 2009), or hybrid simulation-optimization (Ayvaz, 2016).

In the probabilistic approaches, the objective is, generally speaking, to maximize some posterior probability of the source parameters given the observations. Some approaches used for this purpose are minimum relative entropy (Woodbury and Ulrych, 1996; Woodbury et al., 1998; Cupola et al., 2015), the geostatistical approach (Sun, 2007; Gzyl et al., 2014; Butera et al., 2013), Markov chain Monte Carlo (Wang and Jin, 2013), or Bayesian approaches (Zeng et al., 2012; Zhang et al., 2015; Zanini and Woodbury, 2016).

In the deterministic approaches, the main objective is to solve the advection-dispersion equation backward in time. Some of the approaches employ the marching-jury backward beam equation method (Atmadja and Bagtzoglou, 2001a; Bagtzoglou and Atmadja, 2003), Tikhonov regularization (Skaggs and Kabala, 1994; Neupauer et al., 2000), or a quasi-reversibility method together with minimum relative entropy (e.g., Skaggs and Kabala, 1995; Bagtzoglou and Atmadja, 2003; Neupauer et al., 2000).

In addition to the approaches mentioned above, recently, the use of the restart ensemble Kalman filter (r-EnKF) (a probabilistic approach), was proposed by Xu and Gómez-Hernández (2016) to identify a contaminant source by assimilating concentration observations. The good results obtained by the r-EnKF in standard inversion problems (e.g., Hendricks Franssen and Kinzelbach, 2009; Xu et al., 2013; Xu and Gómez-Hernández, 2015b) prompted its use for source identification, where it proved to achieve good results, too. Later, Xu and Gómez-Hernández (2018) extended their work to jointly identify the source information and the underlying hydraulic conductivity field in a synthetic aquifer, and in a tank experiment (Chen et al., 2018). Their works have proven the capability of the r-EnKF for contaminant source identification.

The ensemble smoother (ES), also a probabilistic approach, first proposed by Van Leeuwen and Evensen (1996), is an alternative that could alleviate the computational burden of the

EnKF, because it assimilates all data for all time steps at once. This avoids the restart of the simulation at every time step and makes the ES faster and easier to implement than the EnKF (Emerick and Reynolds, 2013a). However, the performance of the ES for the case of non-linear state equations is not good (e.g., Evensen and Van Leeuwen, 2000; Crestani et al., 2013), the main reason being the lack of multiple updatings inherent to the EnKF (the ES does only one update).

A detailed explanation of why the EnKF outperforms the ES in dealing with non-linear problems can be found in the work by Evensen (2018). Here, a brief explanation is given. The updating step in both the EnKF and the ES are written in terms of covariances, which can only capture linear relationships. The EnKF recursively updates the parameters of interest by assimilating observation information in time and after each step the updates get closer to the reference solution. The ES makes a single update using all the data from all time steps. That is, the EnKF is equivalent to making many linear approximations to the state equation followed by incremental updates along the linear approximation, whereas the ES is equivalent to a single linear approximation to the state equation and a single large update along the linear approximation. Therefore, the EnKF is equivalent to a non-linear optimization based on local linear approximations, whereas the ES is a linear minimization, which may be very far from optimal if the state equation is highly nonlinear. Unless, iteration is also introduced into the ES. This is what Emerick and Reynolds (2013a) propose with their ensemble smoother with multiple data assimilation (ES-MDA). The basic idea is to assimilate all data from all time steps several times, progressively updating the parameters after each iteration.

Several successful applications of the ES-MDA are reported in the reservoir history-matching literature (e.g., Emerick et al., 2013; Emerick and Reynolds, 2013b; Le et al., 2015, 2016; Lee et al., 2013; Fokker et al., 2016). In these works, the reservoir state equations are nonlinear, and the ES-MDA results outperforms the EnKF for both synthetic and real field problems. Recently, a few applications have been reported in the hydrogeology literature (Li et al., 2018a,b) for the characterization of hydraulic conductivities by assimilating piezometric heads.

In this paper, the ES-MDA is used, for the first time, to the best of our knowledge, to jointly identify a heterogeneous hydraulic conductivity field and contaminant source information on a synthetic aquifer. As a benchmark, the accuracy of the ES-MDA will be compared with the r-EnKF. Note that the main aim of this work is to evaluate the capabilities of the ES-MDA and to benchmark it against the r-EnKF for the joint identification of conductivity field and contaminant source information.

The paper is organized as follows. First, we introduce the algorithmic description of the r-EnKF and the ES-MDA. Second, we test and compare the ES-MDA with the r-EnKF on a synthetic aquifer. And third, we discuss the results.

2. Methodology

2.1. Restart ensemble Kalman filter

The EnKF was developed based on the Kalman filter proposed by Kalman et al. (1960) to better tackle nonlinear state-transfer equations. The main difference between the EnKF and the Kalman filter is on how the covariance matrices are calculated. In the original filter, the covariances were propagated in time using a linear state-transfer function (or a linear approximation in case the function is non-linear), while in the EnKF, the covariances are calculated from the states obtained after solving the state-transfer function on an ensemble of realizations (e.g., Evensen, 2003, 2009; Chen and Zhang, 2006; Xu et al., 2013; Xu and Gómez-Hernández, 2015a). Like the Kalman filter, the EnKF consists of two steps: forecast and analysis. The first one is to forecast the state variables from the state variables and the best estimate of the model parameters from the last time step. And the second one is to update the state variables and model parameters at the current time step based on the deviations between forecasted and observed state variable values at selected observation points. However, as already discussed in Xu and Gómez-Hernández (2016), it is impossible to take into account the updated parameters in the forecast step when these parameters define the spatiotemporal position of a contaminant source, unless the forecast is restarted from time zero. This approach modifies the standard Kalman filter equations since there is no need to update the variable values at the analysis step: their values will be recomputed with the new estimates of the model parameters from times zero.

For any given realization of the ensemble, let V_t^f denote the forecasted state variables at time t, and S_t^a the best model parameter estimates after the analysis step at the same time. The forecast equation is

$$V_t^f = \psi(V_0, S_{t-1}^a). \tag{1}$$

where ψ represents the state-transfer function, and V_0 represents the state variables at time zero. The update step modifies the parameter values from the previous time step (S_{t-1}^a) as a function of the discrepancy between forecasted and observed state variables at observation locations

$$S_t^a = S_{t-1}^a + G_t^f(V_{o,t} + e_t - V_{o,t}^f) \tag{2}$$

with

$$G_t^f = D_{SV,t}^f(D_{VV,t}^f + R_t)^{-1}, \tag{3}$$

where $V_{o,t} + e_t$ is the vector of observed concentrations and piezometric heads (composed of the sum of the true head or concentration $V_{o,t}$ plus an observation error e_t of zero mean and covariance R_t), G_t^f is the Kalman gain, $D_{SV,t}^f$ is the cross-covariance between parameters and forecasted state variables at observation locations, and $D_{VV,t}^f$ is the auto-covariance between the forecasted state variables at the observation locations.

Consider that there are N_r realizations in the ensemble and each realization has been discretized into N_e elements. The state variable vector V contains piezometric heads H and concentrations C at all aquifer model cells

$$V = \begin{bmatrix} H \\ C \end{bmatrix}. \tag{4}$$

This vector contains N_r realizations of $2N_e$ variables.

The model parameter vector S contains hydraulic log-conductivity $\ln K$ in all aquifer model cells and the contaminant source parameters, which are source location, X for the x-coordinate, and Y for the y-coordinate, initial release time T, release duration ΔT, and mass-loading rate M

$$S = \begin{bmatrix} \ln K \\ X \\ Y \\ T \\ \Delta T \\ M \end{bmatrix}. \tag{5}$$

This vector contains N_r realizations of $(N_e + 5)$ variables.

Then, if we define $d_t = V_{o,t} + e_t - V_{o,t}^f$ and $P_{VV,t}^f = (D_{VV,t}^f + R_t)^{-1}$, and the covariances are split into the auto- and cross- covariances of each parameter, the updating equation (2), applicable to each realization independently, can be written as

$$S_t^a = \begin{pmatrix} \ln K \\ X \\ Y \\ T \\ \Delta T \\ M \end{pmatrix} + \begin{pmatrix} D_{(\ln K)C,t}^f & D_{(\ln K)H,t}^f \\ D_{XC,t}^f & D_{XH,t}^f \\ D_{YC,t}^f & D_{YH,t}^f \\ D_{TC,t}^f & D_{TH,t}^f \\ D_{(\Delta T)C,t}^f & D_{(\Delta T)H,t}^f \\ D_{MC,t}^f & D_{MH,t}^f \end{pmatrix} \begin{pmatrix} P_{CC,t}^f & P_{CC,t}^f \\ P_{HC,t}^f & P_{HC,t}^f \end{pmatrix} \begin{pmatrix} d_{C,t} \\ d_{H,t} \end{pmatrix} \tag{6}$$

2.2. *Ensemble smoother with multiple data assimilation*

The ES is, conceptually, the same as the r-EnKF but limited to one forecast step (for all the time steps for which observations are available) and a single update step (based on the discrepancies between observations and predictions at all time steps).

The equations that describe the ES are almost the same as those for the r-EnKF above, with some differences. The forecast step is given by

$$V^f = \psi(V_0, S_0). \tag{7}$$

where now V^f contains the state forecasted at all time steps —computed from the initial state V_0 and the initial ensemble of parameters S_0. And the update step is given by

$$S^a = S_0 + G^f(V_o + e - V_o^f), \tag{8}$$

with

$$G^f = D_{SV}^f(D_{VV}^f + R)^{-1}, \tag{9}$$

where $V_o + e$ are all of the observations at observation locations, e are the observation errors, and V_o^f are the forecasts at observation locations. The covariances appearing in Eq. (9), D_{SV}^f and D_{VV}^f are computed for all time steps; these covariance matrices include the cross-covariances between time steps, an aspect not accounted for in the r-EnKF that might render the ES superior to the r-EnKF. From a computational point of view, if there are N_o observations locations sampled N_t times, the sizes of the matrices involved in the r-EnKF are proportional to N_o, whereas in the ES they are proportional to the product $N_o \cdot N_t$. Hence, the sizes of the cross-covariances in the r-EnKF are $(N_e+5)\times 2N_o$ for $D_{SV,t}^f$, and $2N_o \times 2N_o$ for $D_{VV,t}^f$ and R_t; whereas the size of the cross-covariance for the ES are $(N_e+5)\times(2N_o \cdot N_t)$ for D_{SV}^f and $(2N_o \cdot N_t)\times(2N_o \cdot N_t)$ for D_{VV}^f and R.

As we stated before, the performance of the ES is not good when dealing with non-linear problems. The solution provided by Emerick and Reynolds (2013a) to improve the performance of the ES for non-linear state-transfer equations is to iterate, what is called multiple data assimilation (because the same data is assimilated multiple times) on the basis that each iteration of the ES is similar to a Gauss-Newton iteration (Reynolds et al., 2006; Gu and Oliver, 2007). Basically, Eq. (7) and Eq. (10) are iteratively applied using the latest updated parameters as the initial parameters for the next iteration. However, since all data are assimilated multiple times, there is a need to inflate the observation error for each assimilation step. For this purpose, a non-increasing sequence of error variance inflation coefficients $\{a_i, i = 1, ..., N_a\}$ is used in the updating equations, with N_a being the number of assimilation iterations, and satisfying that $\sum_{i=1}^{N_a} \frac{1}{a_i} = 1$.

The ES-MDA equations display the following differences. The forecast step is given by

$$V_i^f = \psi(V_0, S_i^a). \tag{10}$$

where i is the iteration counter, and for each iteration the forecast uses the last updated parameters from the previous iteration. And the update equation is given by

$$S_i^a = S_{i-1}^a + G_i^f (V_{o,i} + \sqrt{a_i} e_i - V_{o,i}^f) \tag{11}$$

with

$$G_i^f = D_{SV,i}^f (D_{VV,i}^f + a_i R_i)^{-1}, \tag{12}$$

In Eq. (11) and Eq. (12), we can see how the observation variance is amplified by a factor a_i and the observation error is amplified by $\sqrt{a_i}$.

If we define $d_i = V_{o,i} + \sqrt{a_i} e_i - V_{o,i}^f$ and $P_{VV,i}^f = (D_{VV,i}^f + a_i R_i)^{-1}$, and the covariances are split into the auto- and cross- covariances of each parameter, the updating equation Eq.(8)

can be written as

$$S_i^a = \begin{pmatrix} \ln K \\ X \\ Y \\ T \\ \Delta T \\ M \end{pmatrix} + \begin{pmatrix} D^f_{(\ln K)C,i} & D^f_{(\ln K)H,i} \\ D^f_{XC,i} & D^f_{XH,i} \\ D^f_{YC,i} & D^f_{YH,i} \\ D^f_{TC,i} & D^f_{TH,i} \\ D^f_{(\Delta T)C,i} & D^f_{(\Delta T)H,i} \\ D^f_{MC,i} & D^f_{MH,i} \end{pmatrix} \begin{pmatrix} P^f_{CC,i} & P^f_{CC,i} \\ P^f_{HC,i} & P^f_{HC,i} \end{pmatrix} \begin{pmatrix} d_{C,i} \\ d_{H,i} \end{pmatrix} \tag{13}$$

Please notice that, when $N_r < 2N_o$ in r-EnKF, or $N_r < 2N_o \times N_t$ in ES-MDA, the low rank of the matrices prevent their inversion; then, the subspace inversion introduced by Evensen (2004) is used to solve for $P^f_{VV,t}$ or $P^f_{VV,i}$. The detailed explanation can be found in the works by Evensen (2004); Emerick and Reynolds (2013a).

3. Application

A synthetic confined aquifer is designed and constructed on a 1000 [L] by 1000 [L] by 50 [L] prism discretized into 50 by 50 by 1 cells, where each cell is 20 [L] by 20 [L] by 50 [L]. (Please note that no specific units are used throughout, only their dimensional analysis is given. Any set of consistent units will yield the same results.) The reference log-conductivity field is drawn from a multivariate Gaussian random function defined by the parameters in Table 1 using the GCOSIM3D software —a sequential Gaussian simulation program (Gómez-Hernández and Journel, 1993). The resulting reference log-conductivity field is shown in Figure 1.

Table 1: Parameters of the random functions used to generate the ln$K$ realizations. Spherical variogram with anisotropic spatial correlation defined by λ_{max} and λ_{min}, which are the ranges in the maximum and minimum directions of continuity. The angle corresponds to the maximum continuity direction and it is measured clockwise from the North direction

	Mean	Std. dev.	Variogram	λ_{max}	λ_{min}	Angle
ln$K$	-1	1	Spherical	300	200	135

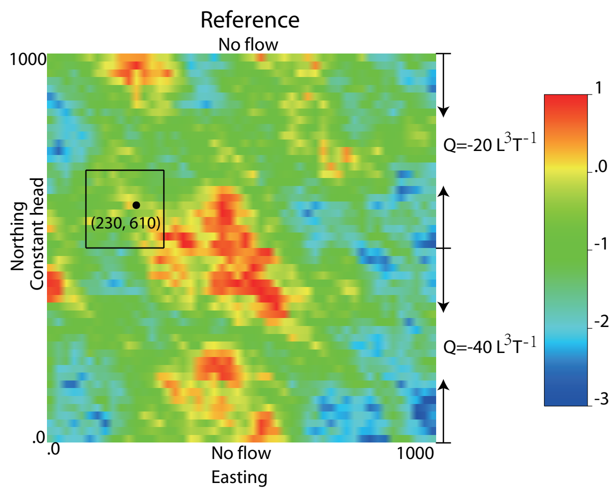

Figure 1: Reference ln$K$ and boundary conditions. The source location is marked with a dark dot. The inner square indicates the suspect contaminant source.

The model boundaries, as indicated in Figure 1, are set as follows: north and south boundaries are impermeable; west boundary is a prescribed head condition with a constant value of 50 [L]; east boundary is a prescribed flow boundary divided into two equal-length segments: the north segment with a total prescribed flow extraction rate of 20 [L^3T^{-1}] and the south segment with a total extraction prescribed flow rate of 40 [L^3T^{-1}]. Figure 2 shows the location of the 25 observation wells (red triangles) and the two verification wells (blue diamonds).

The initial concentration is zero [ML^{-3}] and the initial head for the whole domain is 58 [L], except at the west constant boundary. Other groundwater flow and contaminant transport

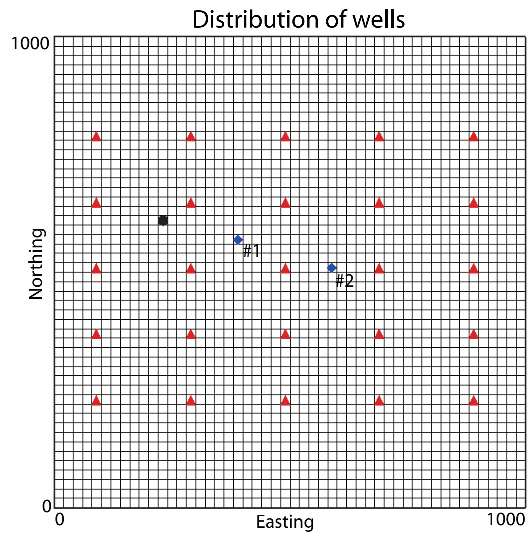

Figure 2: Location of wells. Red triangles mark observation wells; blue diamonds mark verification wells. The black circle is the contaminant source location.

parameters are assumed known and set as homogeneous: porosity of 0.3 [−], longitudinal dispersivity of 2 [L], transverse to longitudinal dispersivity ratio of 0.1.

We assume the contaminants are inert. Only advection and dispersion are considered as transport mechanisms. Both groundwater flow and contaminant transport are under transient conditions. The groundwater flow simulator MODFLOW (McDonald and Harbaugh, 1988) and the transport simulator MT3DMS (e.g., Zheng, 2010; Ma et al., 2012) are used as forward models to solve the groundwater flow and contaminant transport problems, respectively.

The total simulation time is 10000 [T] and is discretized into 100 time steps with increasing size following a geometric series with ratio 1.01 (The first time step is 58.66 [T]). The observations of both piezometric head and concentration from the first 60 time steps (around 4790 [T]) are assimilated for the purpose of parameter identification, so the total number of observations is $2 \times 25 \times 60$.

The contaminant is released at location $(X, Y) = (230, 610)$ [L] with a mass-loading rate of 1000 [MT^{-1}], starting at time 613 [T] (around the 10th time step) and ending at time

2867 [T] (around the 40th time step), with a release duration of 2254 [T].

Figure 3 shows three snapshots of piezometric head and solute concentration taken on the reference aquifer at the 10th simulation time step (beginning of contaminant injection), 40th time step (end of contaminant injection), and at 60th time step (end of assimilation period). This figure also shows the location where both piezometric heads and concentrations are sampled for the purpose of their assimilation in the different scenarios described next.

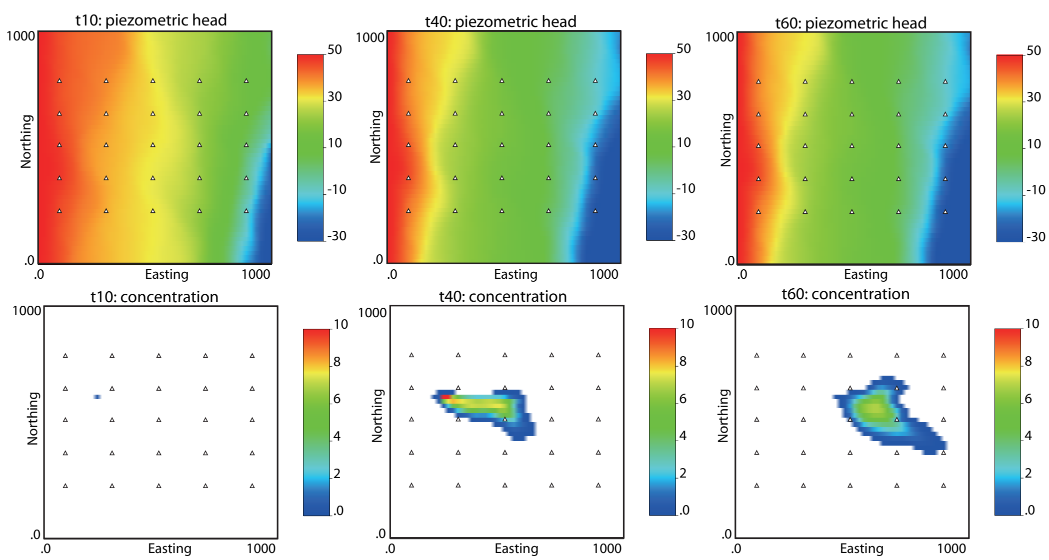

Figure 3: Reference. Piezometric head (top row) and contaminant plume (bottom row) at the 10th (beginning of solute injection), 40th (end of solute injection), and 60th (end of assimilation) time steps in the reference aquifer. White triangles mark the observation wells.

Seven scenarios will be evaluated. The first one, used as a benchmark to evaluate the efficiency of the ES-MDA, is the r-EnKF, which has already proven its ability for the identification of contaminant source parameters and hydraulic conductivity characterization; it will be referred to as S0. The second one is the ES in its original implementation, that is, without any iteration. Then, to evaluate the effect of the number of iterations, the ES-MDA is run for five different scenarios, the difference between them is the number of iterations (or data assimilations) performed; they will be labeled S2 to S6 with 2, 4, 6, 8 and 10 iterations, respectively. Notice that the observation error inflation coefficients a_i will, in all cases, be equal to the number of iterations, following the recommendations by Emerick and Reynolds (2013a), who show that the use of decreasing inflation coefficients leads to only small improvements with respect to using the inflation coefficients equal to the number of iterations.

As we stated before, the total number of cells of the log-conductivity field is 50×50, and the number of source parameters is 5, so the total number of parameters to be identified is 2505. An ensemble of 400 initial log-conductivity realizations is generated using the same random function model and parameters as for the reference log-conductivity field. The number of ensemble members was chosen after a previous analysis with ensemble sizes of 200, 400 and 800 members. The difference in results between the ensemble sizes of 400 and 800 were not large enough to grant the use of the largest ensemble. Notice that there are no conditioning log-conductivity data, thus the ensemble mean and ensemble variance of the initial log-conductivity realizations are homogeneous and equal to their marginal values. As already discussed by Xu et al. (2013) the use of the same random function parameters for the generation of the initial realizations as for the generation of the reference case is only a marginal advantage given that there are no conditioning conductivities. Indeed, Xu et al. (2013) demonstrate the effectiveness of the r-EnKF using a totally uninformative prior random function for the generation of the initial ensemble, with similar results as when the "true" random function is used. In addition, an ensemble of 400 5-tuplets for the source parameters is generated, each 5-tuplet contains five values drawn independently from the following uniform distributions: initial release time $T \in \mathcal{U}[550, 750]$, release duration $\Delta T \in \mathcal{U}[2100, 2300]$, mass-loading rate $M \in \mathcal{U}[900, 1100]$, and source location $(X, Y) \in (\mathcal{U}[100, 300] \times \mathcal{U}[500, 700])$.

4. Results

Before starting the analysis of the results, Table 2 shows the CPU consumption for all scenarios. Recall that in the r-EnKF (S0) there are 60 forecasting steps starting from time 0, and 60 assimilation steps to update the parameters 60 times based on the observations at 25 wells; whereas, in the ES-MDA the number of model runs for the whole simulation period is equal to the number of assimilation steps, but at each assimilation step, there are 1500 observations (25 observation locations times 60 time steps). For the current model and setup, the ES-MDA is cheaper to run than the r-EnKF up until data are assimilated four times. When ten iterations are performed, the ES-MDA costs two and half times that of the r-EnKF.

Table 2: Definition of scenarios and CPU time consumption. The number in parenthesis refers to the number of data assimilation steps used in the ES-MDA. (ES would be equivalent to ES-MDA(1))

Method	Scenario	CPU in s	CPU in % of S0
r-EnKF	S0	16366	100%
ES	S1	4981	30%
ES-MDA(2)	S2	9526	58%
ES-MDA(4)	S3	17937	110%
ES-MDA(6)	S4	27432	149%
ES-MDA(8)	S5	34936	210%
ES-MDA(10)	S6	42422	259%

The r-EnKF, the ES and the ES-MDA will be used to assimilate the piezometric head and concentration data at the 25 observation locations. This assimilation will result in an ensemble of updated parameters (for the spatial distribution of hydraulic conductivity and for the parameters defining the contaminant source) that are used to produce an ensemble of piezometric heads and concentrations past the assimilation period (60th time step) for 40 time steps more. The performance of the different scenarios will be evaluated by comparing the different final ensembles to their corresponding counterparts in the reference aquifer.

Figure 4 shows the ensemble mean and the ensemble variance of the updated log-conductivities for scenarios S0 to S3 and S6. (The corresponding maps for S4 and S5 for this and following figures are shown in the appendix.) The ensemble mean shows how the main patterns of variability of the reference are captured by the updated ensemble, and the ensemble variance shows the local variability of the updated log-conductivities. From a purely qualitative point of view it is clear that the r-EnKF does a good job in capturing the reference patterns with a small local uncertainty where the ensemble variance is close to zero, that the ES is able to extract patterns which are, overall, similar to the reference but still far from them and with a substantial local uncertainty, and that the ES-MDA gets better the more times data are assimilated, with scenario S6 —for which data are assimilated 10 times —giving the best results.

The above analysis can be quantified by computing the average absolute bias (AAB) and the ensemble spread (ESp). The AAB is used to measure the average absolute deviation between the updated values and the reference ones. The ESp measures the precision of the ensemble of updated realizations by calculating the root square of the ensemble variance. Their expressions are the following

$$\mathrm{AAB} = \frac{1}{N_e} \sum_{i=1}^{N_e} \frac{1}{N_r} \sum_{j=1}^{N_r} | \ln K_{i,j} - \ln K_{i,ref} |, \tag{14}$$

$$\mathrm{ESp} = \sqrt{\frac{1}{N_e} \sum_{i=1}^{N_e} \sigma_i^2}, \tag{15}$$

where N_e is the number of model elements, N_r is the number of realizations, $\ln K_{i,ref}$ is the reference log-conductivity value at node i, $\ln K_{i,j}$ is the log-conductivity at node i for

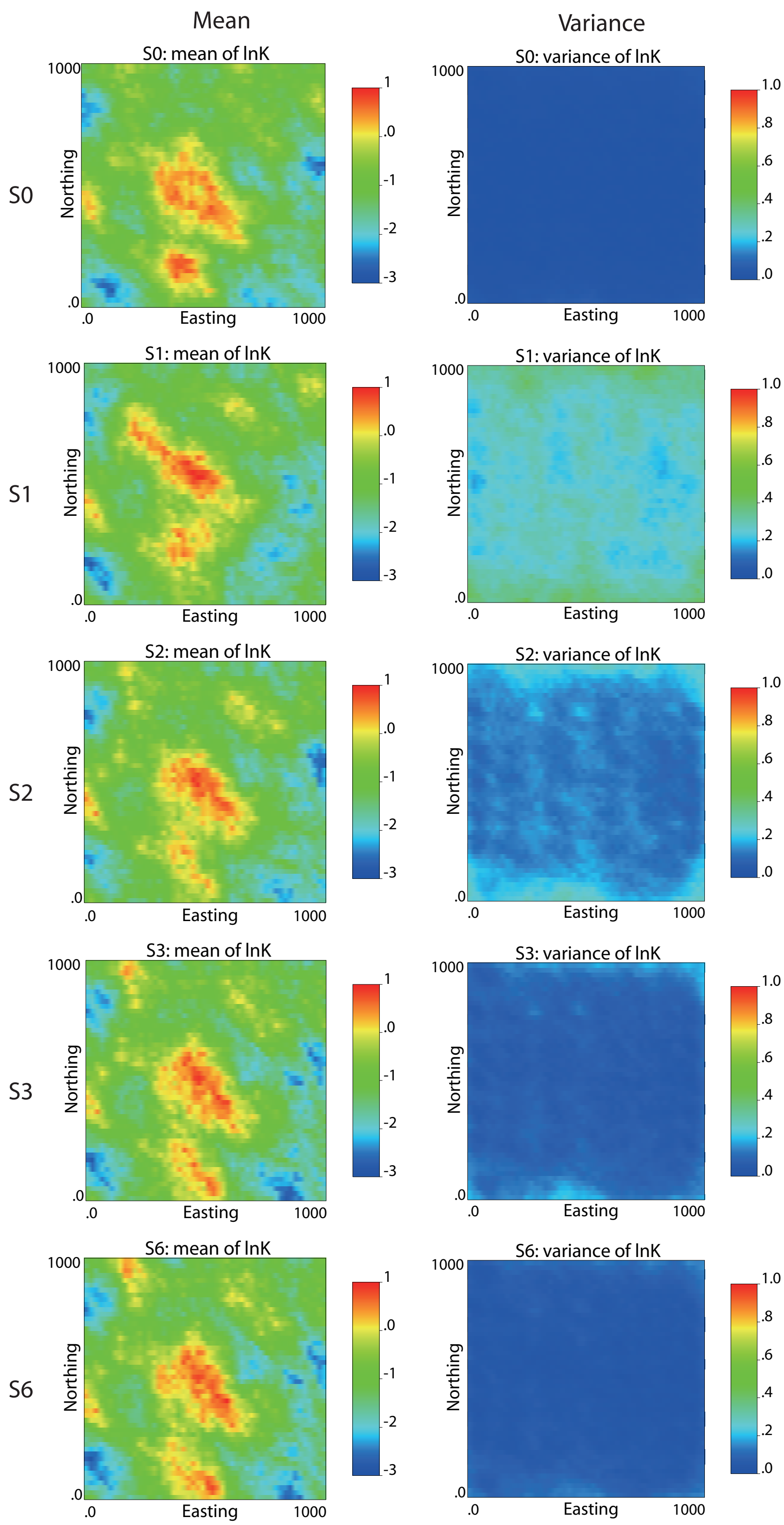

Figure 4: Scenarios S0-S3 and S6. Ensemble mean (left column) and ensemble variance (right column) of updated log-conductivity realizations.

realization j and σ_i is the log-conductivity ensemble variance at node i.

Figure 5 shows the AAB and ESp of lnK and of the parameters defining the contaminant source for all scenarios, computed before any data assimilation and after data have been assimilated over the first 60 time steps. The values, as expected, are the highest for the initial ensembles. They are drastically reduced for the r-EnKF except for ΔT and M. The smoother provides increasingly smaller values as the number of assimilation steps increases, with the best values for S6 after ten iterations. Specifically, the AAB and ESp of the updated lnK, and Y for scenarios S3-S6 is close to that of scenario S0, and the AAB and ESp of the updated T for scenario S6 is close to that of scenario S0; while, the AAB and ESp of the updated X, ΔT and M of scenarios S3-S6 is smaller than that of scenario S0. From these results we could conclude that, after four assimilation steps, the ES-MDA starts to perform better than the r-EnKF.

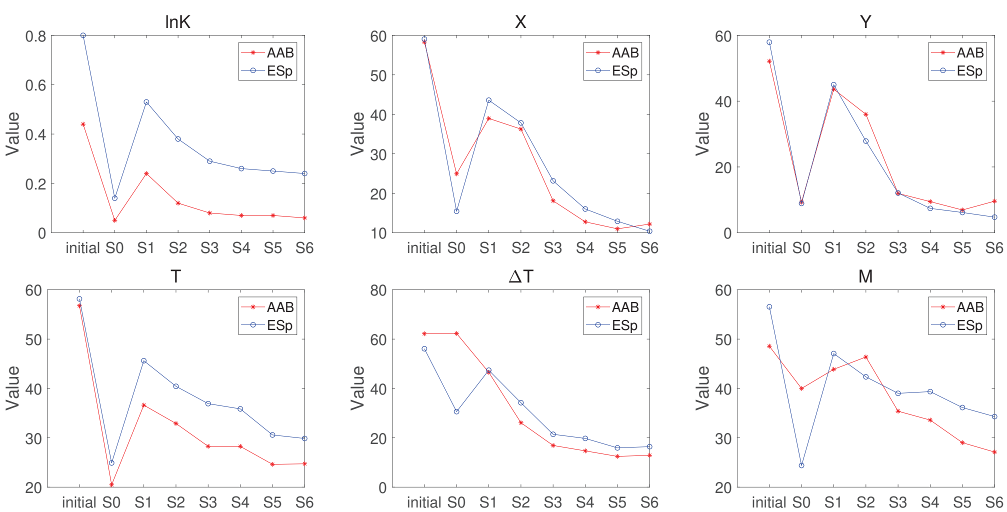

Figure 5: Scenarios S0-S6. Average absolute bias (AAB) and ensemble spread (ESp) of log-conductivity (lnK), source location (X and Y), initial release time (T), release duration (ΔT), and mass-loading rate (M) computed on the initial parameters and on the updated parameters for the different scenarios after 60 time steps.

Figure 6 shows the piezometric head distribution at the 60th time step computed with the final updated parameters for scenarios S0 to S3 and S6. The maps show, in the left column, the piezometric head distributions for an individual ensemble member (realization #300), in the center column, the ensemble mean obtained as the local mean of the piezometric head at each node through the 400 realizations, and in the right column the ensemble variance. Please, notice that the middle column with the ensemble mean piezometric heads is not the solution of the state equations in the ensemble log-conductivity average of Fig. 4. An analysis of these maps shows the robustness of the r-EnKF (S0) that produces an ensemble mean map quite close to the reference one (upper right corner in Fig. 3) and with little variability everywhere. The smoother performs well when comparing the average ensemble with the reference map, but the uncertainties associated are quite large, especially in scenarios S1 and S2; there is a need to assimilate the data at least four times (S3) to get a variance reduction that approximates that of the r-EnKF.

Figure 7 shows the concentration plume computed with the parameters updated using observations at 60 time steps. In the left column, the plume in realization #300, in the center, the ensemble mean of the 400 plumes computed in the 400 realizations with updated parameters, and in the right column the local concentration variance computed at each node through the ensemble of realizations. Please, notice that, as with piezometric heads, the middle column with the ensemble mean concentrations is not the solution of the state equations in the ensemble log-conductivity average of Fig. 4. An analysis of these maps reaches the same conclusions as for the piezometric heads, the r-EnKF is quite robust producing an ensemble mean plume quite close to the reference (lower right corner in Fig. 3) and with lower variability. The smoother performs well only when the number of iterations is large (S3 and S6); for the cases of one, and two iterations (S1 and S2, respectively), the ensemble mean plume is quite spread, the local variance is large, and the plume in the single selected realization shown in the left column of the figure can be quite far from the reference one.

Figure 8 shows the time evolution of piezometric heads and solute concentrations at the two verification wells (#1 and #2) computed using the initial ensembles of contaminant

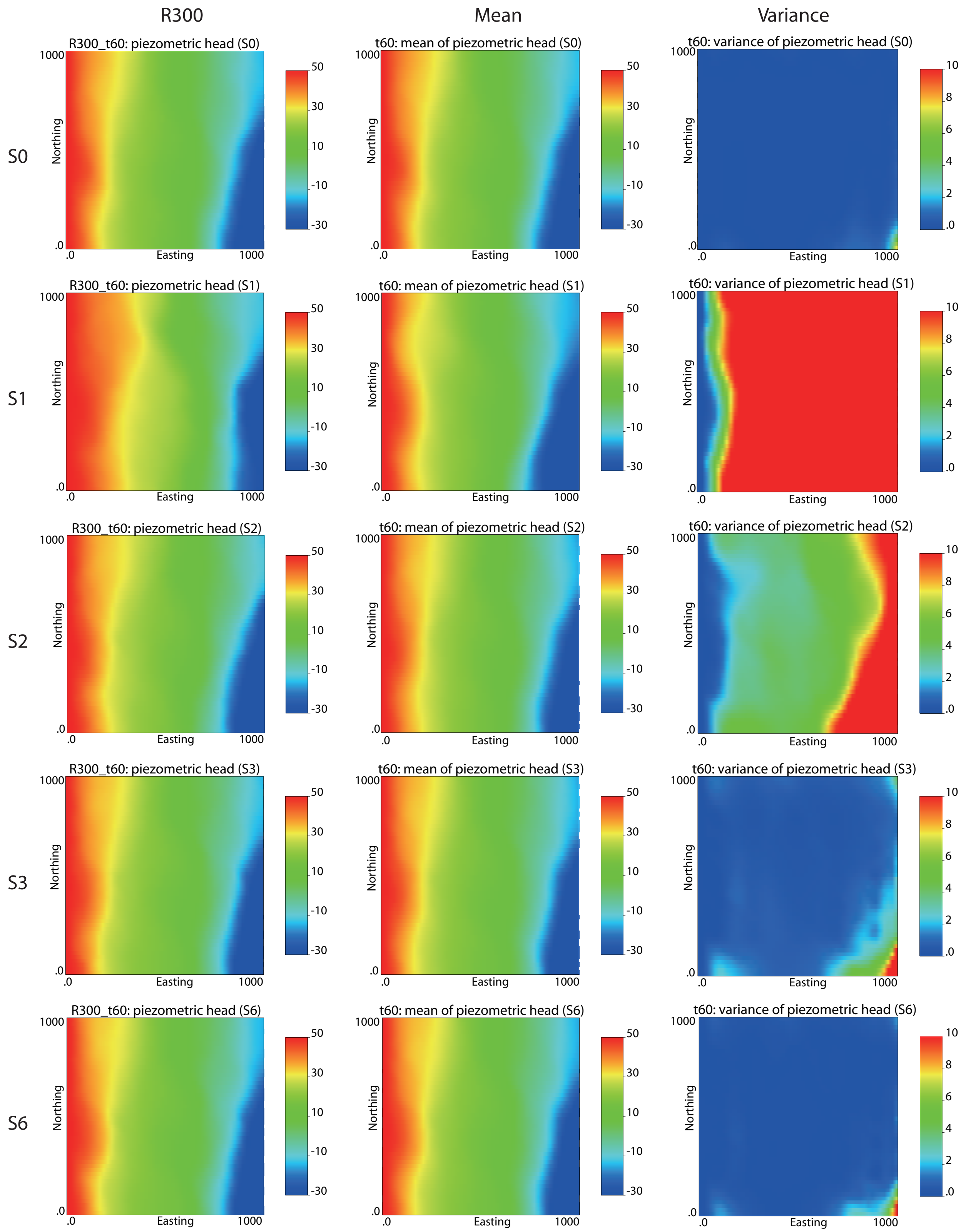

Figure 6: Scenarios S0-S3 and S6. Piezometric heads as computed with the updated parameters at the end of the 60th time step. From left to right, heads in realization #300; ensemble mean, and ensemble variance.

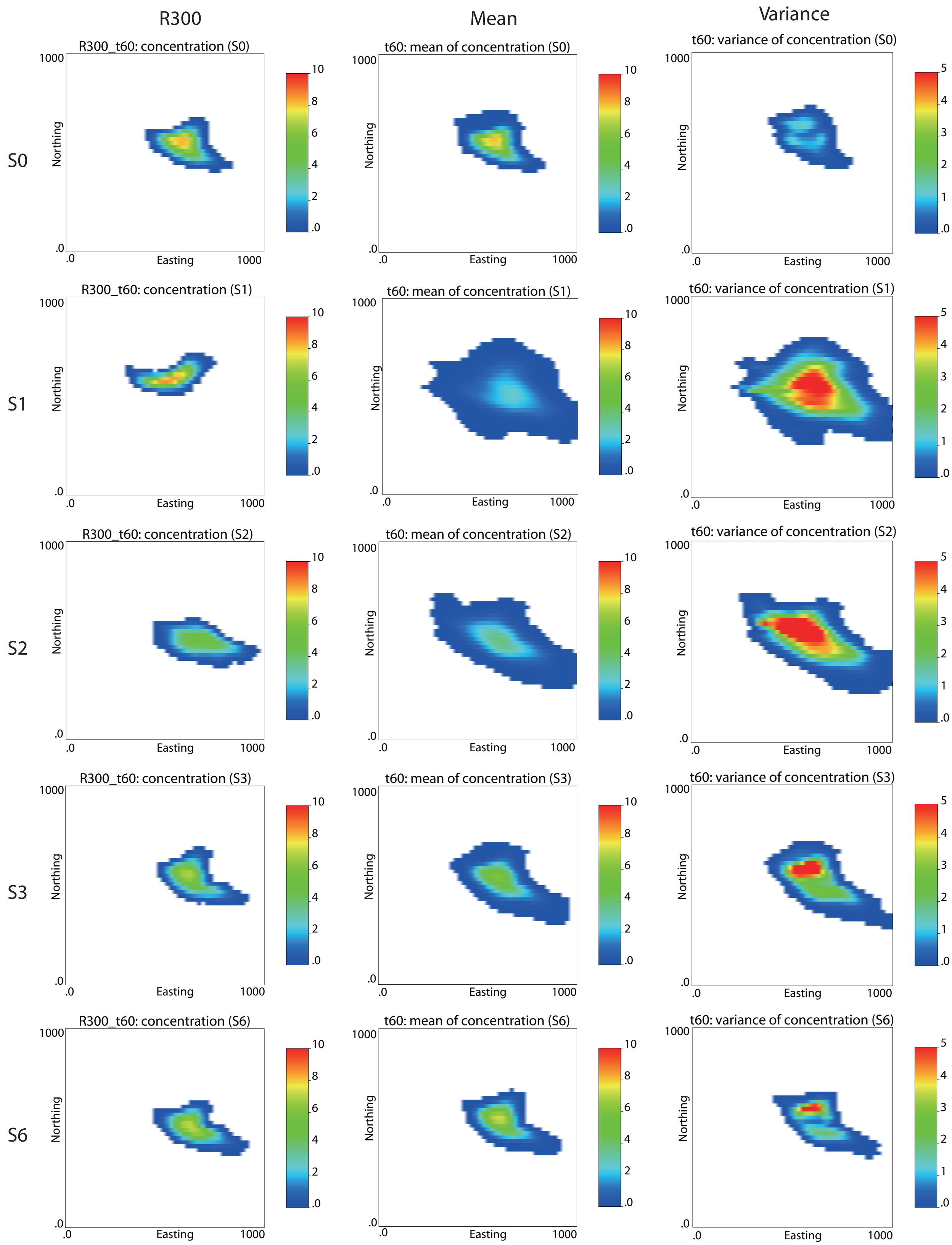

Figure 7: Scenarios S0-S3 and S6. Contaminant plume as computed with the updated parameters at the end of the 60th time step. From left to right, Contaminant plume in realization #300; ensemble mean of all contaminant plumes, and ensemble variance of all contaminant plumes.

source parameters and log-conductivities. The spread of predicted values is quite large since no observation has been assimilated yet. Figure 9 and 10 show the time evolution of piezometric heads and solute concentrations computed with the updated source parameters and log-conductivity fields after the assimilation of the observations during the first 60 time steps, respectively. The spread of the curves after the assimilation is considerably reduced, especially for scenarios S0, S3 and S6. Although these two wells were not used during the assimilation, the reproduction of piezometric heads, even after the assimilation period ends is very good both for the r-EnKF (S0) and for the ES-MDA with four and ten iterations (S3 and S6), with the former performing slightly better than the latter.

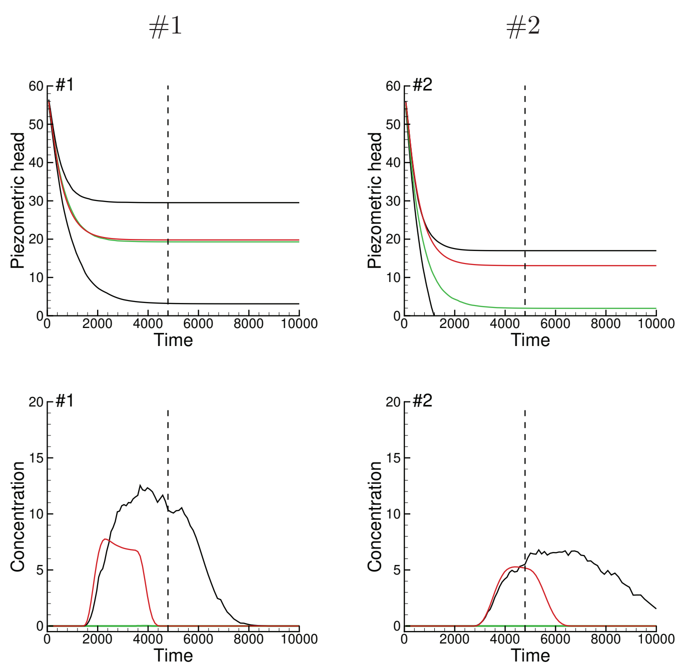

Figure 8: Time evolution of piezometric heads (top row) and solute concentrations (bottom row) at the two verification wells #1, and #2 computed on the initial ensemble of source information parameters and $\ln K$. The red line corresponds to the the reference field. The black lines correspond to the 5 and 95 percentiles of all realizations, and the green line corresponds to the median. The vertical dashed lines mark the end of the assimilation period.

Up to here, regarding the characterization of the log-conductivity field and the reproduction of the state variables, the r-EnKF seems to outperform the ES-MDA with four

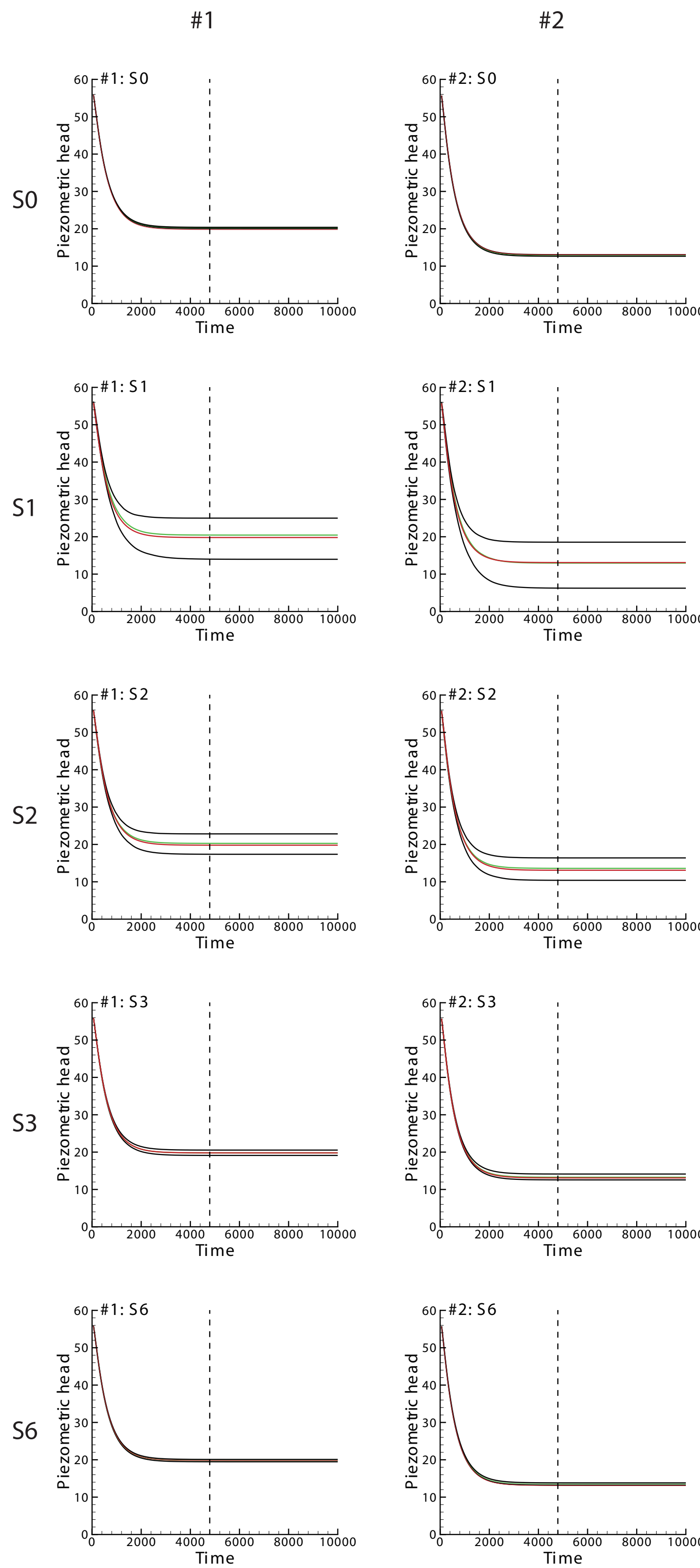

Figure 9: Scenarios S0-S3 and S6. Time evolution of the piezometric heads at the two verification wells #1, and #2 computed with the updated ensemble of source information parameters and $\ln K$ after the assimilation of the observations of the first 60 time steps. The red line is the evolution of the piezometric head in the reference. The black lines correspond to the 5 and 95 percentiles of all realizations, and the green line corresponds to the median. The vertical dashed lines mark the end of the assimilation period.

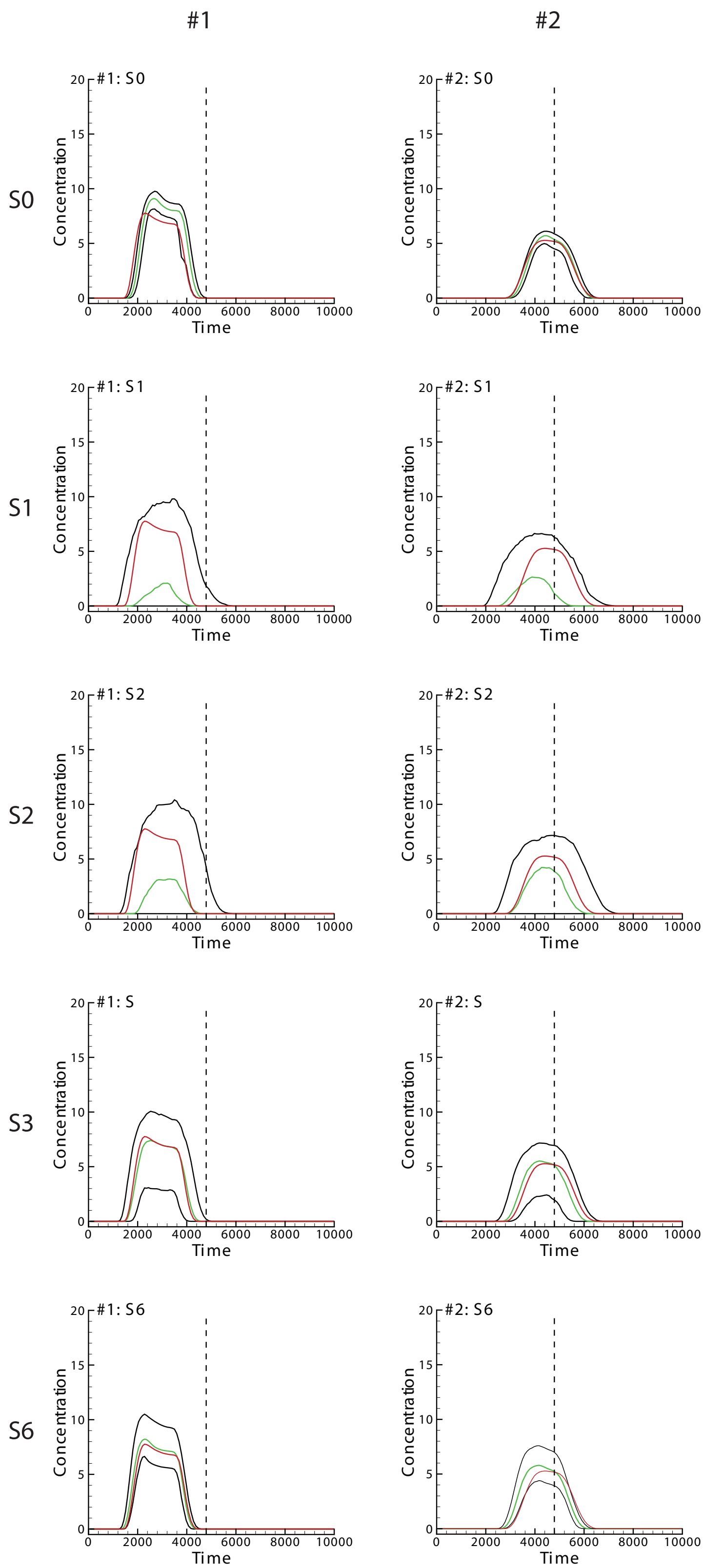

Figure 10: Scenarios S0-S3 and S6. Time evolution of the solute observations at the two verification wells #1, and #2 computed with the updated ensemble of source information parameters and $\ln K$ after the assimilation of the solute observations of the first 60 time steps. The red line is the evolution of the concentration in the reference. The black lines correspond to the 5 and 95 percentiles of all realizations, and the green line corresponds to the median. The vertical dashed lines mark the end of the assimilation period.

iterations. The AAB($\ln K$) and ESp($\ln K$) are the smallest for S0 (r-EnKF), and the piezometric head and concentration predictions are also the best for S0. Only the ES-MDA with ten assimilation steps (S6) gives comparable results, although at a CPU cost 2.6 times larger than the r-EnKF.

However, when we analyze the reproduction of the contaminant source parameters, we have already discussed Figure 5 showing that the ES-MDA is superior to the r-EnKF. This observation is complemented by the results shown in Figure 11, in which boxplots of the initial ensemble and the updated ensemble of the source parameters for the six scenarios are shown. Some observations that can be derived from this figure are: the r-EnKF (S0) produces good estimates for X, Y and T with a considerable reduction of uncertainty with respect to the initial ensemble, while the estimates for ΔT and M are somehow biased without a large reduction of uncertainty; the ES (S1) is not effective, the spreads of the ensemble is almost the same as for the initial ensemble prior to assimilation for all parameters; the ES-MDA starts to work well after four iterations, and gives the best results for ten iterations, outperforming the r-EnKF, particularly for parameters X, ΔT and M. The difficulty on estimating ΔT and M is due to the fact that several combinations of these two parameters can result in very similar sets of observations, making more difficult their identification with a reduction of uncertainty. The only way to solve this indetermination is the collection of additional observations. This is precisely the reason why the ES-MDA with ten iterations works better than the r-EnKF for these two parameters: the r-EnKF uses all observational data only once in a piecewise way, whereas the ES-MDA uses all observation data altogether ten times.

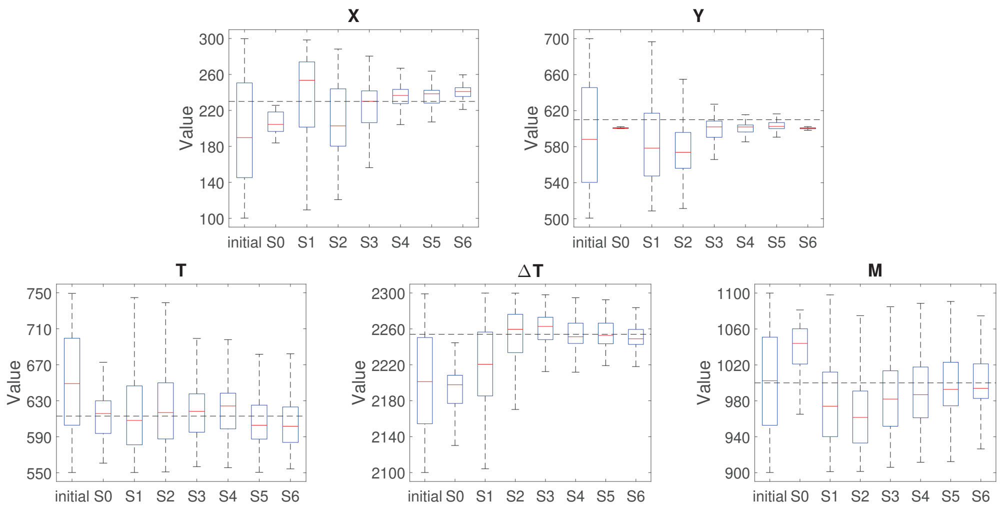

Figure 11: Scenarios S0-S6. Boxplots of the source location (X and Y), initial release time (T), release duration (ΔT), and mass-loading rate (M) computed with the initial parameters and with the updated parameters after 60 time steps. The dashed horizontal black line corresponds to the reference value.

5. Summary and Discussion

The purpose of this paper is to analyze the ability of the ES-MDA for the identification of contaminant source parameters together with a spatially heterogeneous hydraulic conductivity field in comparison with the r-EnKF. The results show that the ES-MDA has the ability to estimate hydraulic conductivity field and identify the contaminant source parameters — including source location, initial release time, release duration and mass-loading rate– with a proper number of iterations, besides, the results also indicate that these estimate parameters are good enough to provide good forecasts of solute concentrations and piezometric heads.

It is also worth pointing out that this is the first time that the ES-MDA is applied for contaminant source identification.

Furthermore, the comparison over all scenarios (including also the scenarios in the appendix) between the r-EnKF and the ES-MDA, shows that the ES-MDA performs better than the r-EnKF, especially for the identification of contaminant source parameters when using enough number of iterations. For the specific test done here, the ES-MDA starts to outperform the r-EnKF after four iterations, needing almost the same computer time as that for r-EnKF. The ES-MDA can perform even better using more iterations (and at a higher computational cost). Part of the reason of the better performance of ES-MDA than of r-EnKF is the fact that the number of observations is much larger for the ES-MDA, which is specially important for the proper identification of mass-loading and release duration. These two parameters are identified with large uncertainty by the r-EnKF.

It hovers over the whole paper whether an analysis on a single synthetic test case on seven scenarios is sufficient to draw general conclusions about the comparison between the r-EnKF and the ES-MDA. The answer is no, but drawing general conclusions was not the purpose of this paper, its purpose was to test the newcomer ES-MDA against the r-EnKF in a setting in which the r-EnKF had already proven to be quite effective. Given our extensive experience with the application of the r-EnKF, we can forecast that a sensitivity study to the number of observations will show that there is a threshold number below which it will be impossible to identify the source; or that reducing the number of members of the ensemble will require the use of localization and covariance inflation techniques to reach similar results, with a threshold number of realizations below which identification will be impossible; or that including a more uncertain prior distribution for the parameters describing the source will have little impact to effectively identify the source beyond increasing the number of assimilation steps.

While the just-mentioned sensitivity analyses are worth to carry out in a further study, there is an even more interesting issue that has not been addressed neither with the r-EnKF nor the ES-MDA, which is the analysis of more complex contamination events, such as non-punctual or multiple source ones. Addressing these events would require a thoughtful parameterization of the source.

Appendix A. Results of scenarios S4 and S5

Results for scenarios S4 and S5 are displayed in Figures A.12 to A.15. The details are as follows: Figure A.12 shows the ensemble mean and ensemble variance of the updated $\ln K$; Figure A.13 and A.14 show the 300th realization, ensemble mean and ensemble variance of piezometric heads and of the contaminant plume at the end of the 60th time step, respectively; Figure A.15 and A.16 show the time evolution of the piezometric heads and of solute concentrations at the two verification wells #1, and #2 computed with the updated source parameters and hydraulic conductivities.

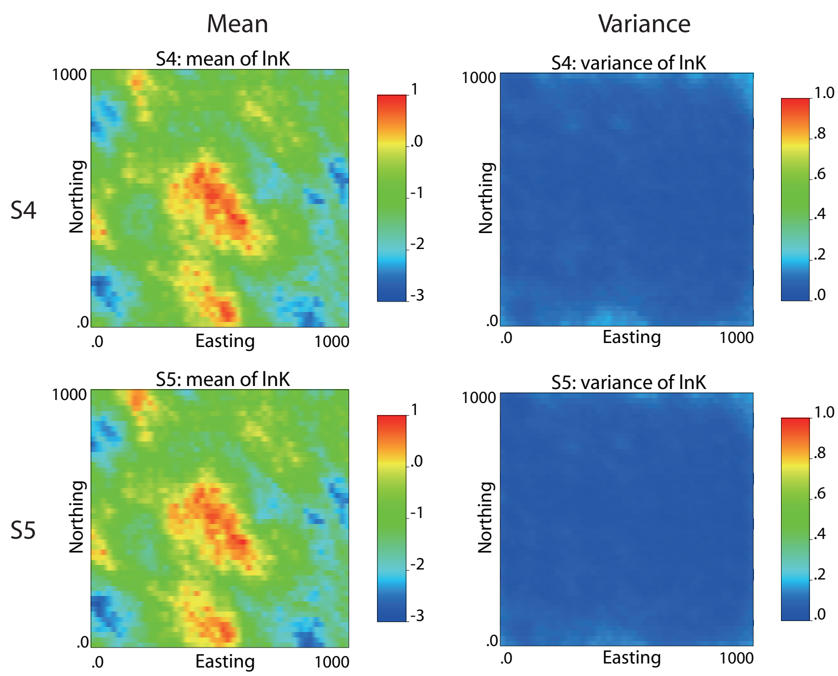

Figure A.12: Scenarios S3-S4. Ensemble mean (left column) and ensemble variance (right column) of updated log-conductivity realizations. (This figure complements Figure 4.)

Acknowledgements Financial support to carry out this work was received from the Spanish Ministry of Economy and Competitiveness through project CGL2014-59841-P, and from the Spanish Ministry of Education, Culture and Sports through a fellowship for the

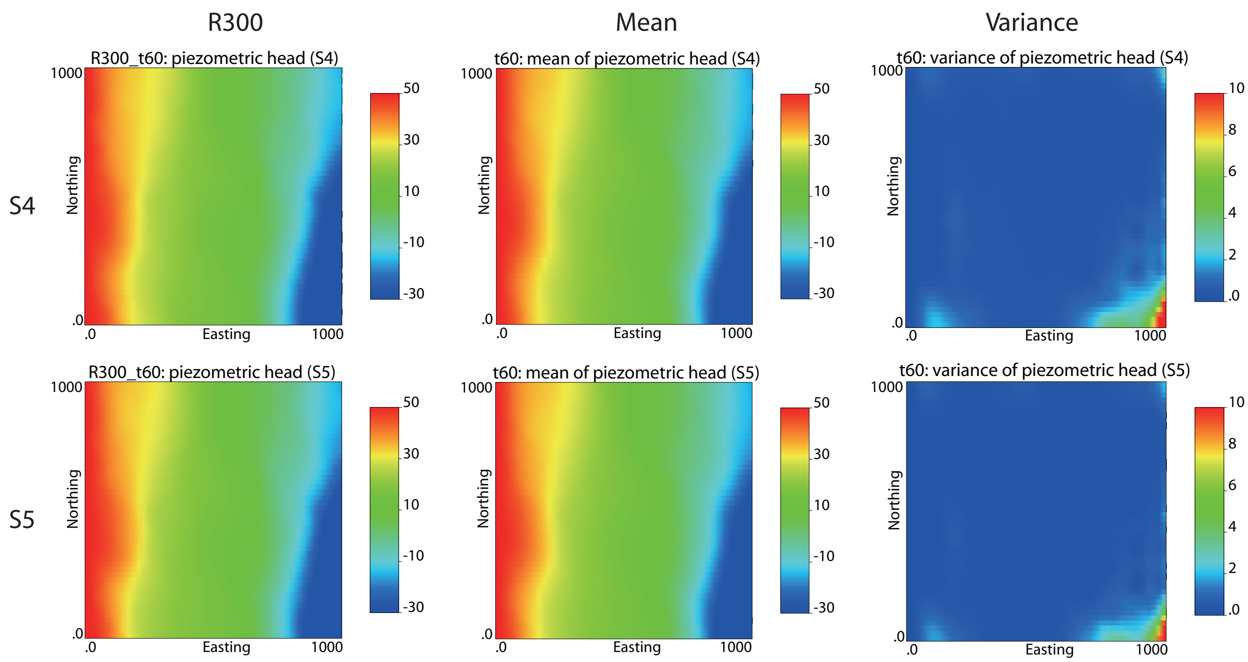

Figure A.13: Scenarios S4-S5. Piezometric heads computed with the updated parameters at the end of the 60th time step. From left to right, heads in realization #300; ensemble mean, and ensemble variance. (This figure complements Figure 6.)

mobility of professors in foreign research and higher education institutions of reference to the second author, reference PRX17/00150. Teng Xu also acknowledges the financial support from the Fundamental Research Funds for the Central Universities (B200201015) and Jiangsu Specially-Appointed Professor Program (B19052). Chunhui Lu acknowledges the financial support from the National Natural Science Foundation of China (51679067 and 51879088), and Fundamental Research Funds for the Central Universities (B200204002).

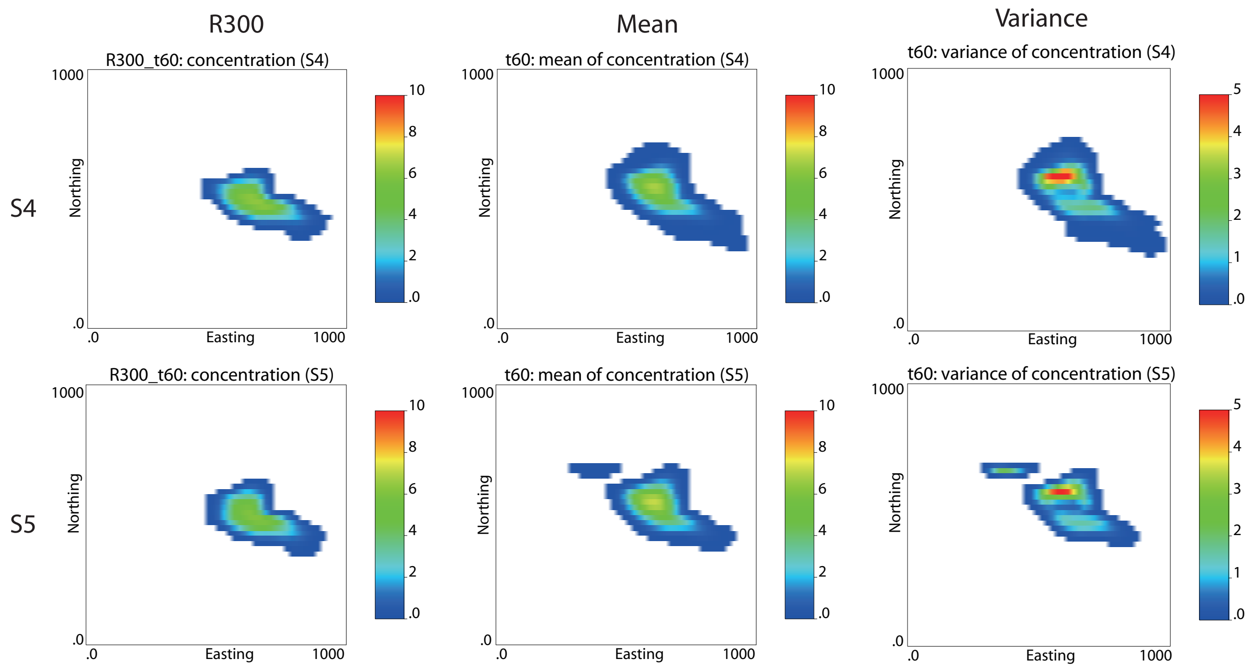

Figure A.14: Scenarios S4-S5. Contaminant plume computed with the updated parameters at the end of the 60th time step. From left to right, Contaminant plume in realization #300; ensemble mean, and ensemble variance. (This figure complements Figure 7.)

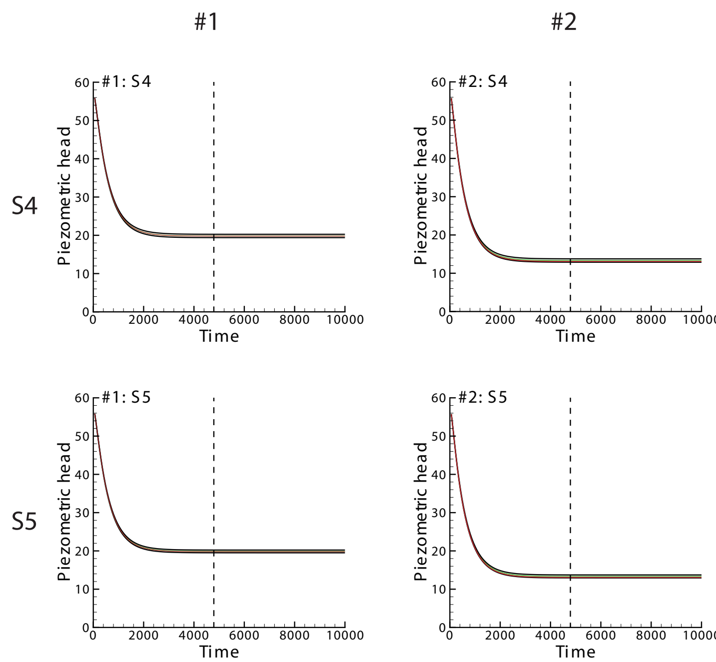

Figure A.15: Scenarios S4-S5. Time evolution of the piezometric heads at the two verification wells #1, and #2 computed with the updated ensemble of source information parameters at the end of the 60th time step. The red line is the evolution of the piezometric head in the reference. The black lines correspond to the 5 and 95 percentiles of all realizations, and the green line corresponds to the median. The vertical dashed lines mark the end of the assimilation period. (This figure complements Figure 9.)

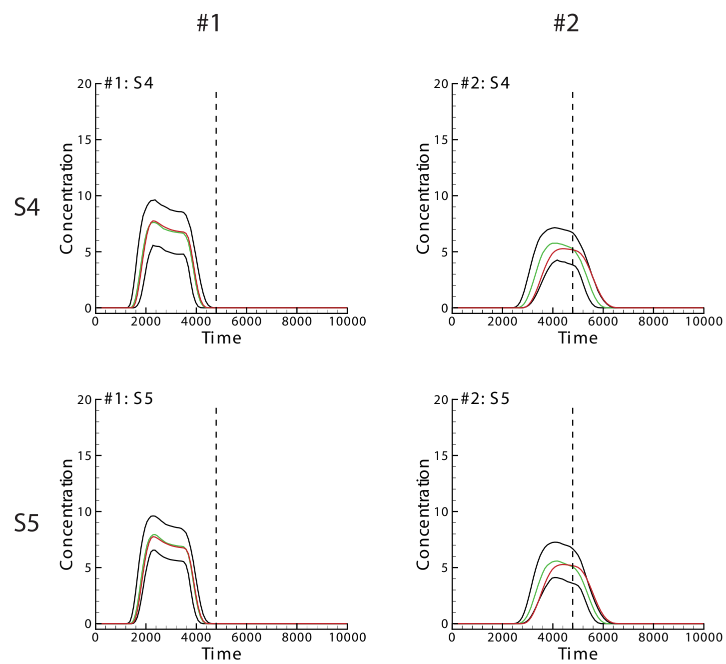

Figure A.16: Scenarios S4-S5. Time evolution of the solute concentrations at the two verification wells #1, and #2 computed with the updated ensemble of source information parameters at the end of the 60th time step. The red line is the evolution of the solute concentration in the reference. The black lines correspond to the 5 and 95 percentiles of all realizations, and the green line corresponds to the median. The vertical dashed lines mark the end of the assimilation period. (This figure complements Figure 10.)

References

Atmadja, J., Bagtzoglou, A.C., 2001a. Pollution source identification in heterogeneous porous media. Water Resources Research 37, 2113–2125.

Atmadja, J., Bagtzoglou, A.C., 2001b. State of the art report on mathematical methods for groundwater pollution source identification. Environmental Forensics 2, 205–214.

Ayvaz, M.T., 2016. A hybrid simulation–optimization approach for solving the areal groundwater pollution source identification problems. Journal of Hydrology 538, 161–176.

Bagtzoglou, A.C., Atmadja, J., 2003. Marching-jury backward beam equation and quasi-reversibility methods for hydrologic inversion: Application to contaminant plume spatial distribution recovery. Water Resources Research 39.

Bagtzoglou, A.C., Atmadja, J., 2005. Mathematical methods for hydrologic inversion: The case of pollution source identification, in: Water Pollution. Springer, pp. 65–96.

Butera, I., Tanda, M.G., Zanini, A., 2013. Simultaneous identification of the pollutant release history and the source location in groundwater by means of a geostatistical approach. Stochastic Environmental Research and Risk Assessment 27, 1269–1280.

Chen, Y., Zhang, D., 2006. Data assimilation for transient flow in geologic formations via ensemble Kalman filter. Advances in Water Resources 29, 1107–1122.

Chen, Z., Gómez-Hernández, J.J., Xu, T., Zanini, A., 2018. Joint identification of contaminant source and aquifer geometry in a sandbox experiment with the restart ensemble Kalman filter. Journal of Hydrology 564, 1074–1084.

Crestani, E., Camporese, M., Baú, D., Salandin, P., 2013. Ensemble Kalman filter versus ensemble smoother for assessing hydraulic conductivity via tracer test data assimilation. Hydrology and Earth System Sciences 17, 1517.

Cupola, F., Tanda, M.G., Zanini, A., 2015. Contaminant release history identification in 2-D heterogeneous aquifers through a minimum relative entropy approach. SpringerPlus 4, 656.

Emerick, A.A., Reynolds, A.C., 2013a. Ensemble smoother with multiple data assimilation. Computers & Geosciences 55, 3–15.

Emerick, A.A., Reynolds, A.C., 2013b. Investigation of the sampling performance of ensemble-based methods with a simple reservoir model. Computational Geosciences 17, 325.

Emerick, A.A., Reynolds, A.C., et al., 2013. History-matching production and seismic data in a real field case using the ensemble smoother with multiple data assimilation, in: SPE Reservoir Simulation Symposium, Society of Petroleum Engineers.

Evensen, G., 2003. The ensemble Kalman filter: Theoretical formulation and practical implementation. Ocean Dynamics 53, 343–367.

Evensen, G., 2004. Sampling strategies and square root analysis schemes for the EnKF. Ocean Dynamics 54, 539–560.

Evensen, G., 2009. Data assimilation: the ensemble Kalman filter. Springer Science & Business Media.

Evensen, G., 2018. Analysis of iterative ensemble smoothers for solving inverse problems. Computational Geosciences 22, 885–908.

Evensen, G., Van Leeuwen, P.J., 2000. An ensemble Kalman smoother for nonlinear dynamics. Monthly Weather Review 128, 1852–1867.

Fokker, P., Wassing, B., van Leijen, F., Hanssen, R., Nieuwland, D., 2016. Application of an ensemble smoother with multiple data assimilation to the Bergermeer gas field, using PS-InSAR. Geomechanics for Energy and the Environment 5, 16–28.

Gómez-Hernández, J.J., Journel, A.G., 1993. Joint sequential simulation of Multi-Gaussian fields, in: Soares, A. (Ed.), Geostatistics Tróia '92, Kluwer Academic Publishers, Dordrecht. pp. 85–94.

Gorelick, S.M., Evans, B., Remson, I., 1983. Identifying sources of groundwater pollution: an optimization approach. Water Resources Research 19, 779–790.

Gu, Y., Oliver, D., 2007. An iterative ensemble Kalman filter for multiphase fluid flow data assimilation. SPE Journal 12, 438–446.

Gzyl, G., Zanini, A., Fraczek, R., Kura, K., 2014. Contaminant source and release history identification in groundwater: A multi-step approach. Journal of contaminant hydrology 157, 59–72.

Hendricks Franssen, H.J., Kinzelbach, W., 2009. Ensemble Kalman filtering versus sequential self-calibration for inverse modelling of dynamic groundwater flow systems. Journal of Hydrology 365, 261–274.

Kalman, R., et al., 1960. A new approach to linear filtering and prediction problems. Journal of Basic Engineering 82, 35–45.

Le, D.H., Emerick, A.A., Reynolds, A.C., et al., 2016. An adaptive ensemble smoother with multiple data assimilation for assisted history matching. SPE Journal 21, 2–195.

Le, D.H., Younis, R., Reynolds, A.C., et al., 2015. A history matching procedure for non-Gaussian facies based on ES-MDA, in: SPE Reservoir Simulation Symposium, Society of Petroleum Engineers.

Lee, K., Jeong, H., Jung, S., Choe, J., 2013. Improvement of ensemble smoother with clustered covariance for channelized reservoirs. Energy Exploration & Exploitation 31, 713–726.

Li, L., Puzel, R., Davis, A., 2018a. Data assimilation in groundwater modelling: ensemble Kalman filter versus ensemble smoothers. Hydrological Processes 32, 2020–2029.

Li, L., Stetler, L., Cao, Z., Davis, A., 2018b. An iterative normal-score ensemble smoother for dealing with non-Gaussianity in data assimilation. Journal of Hydrology .

Ma, R., Zheng, C., Zachara, J.M., Tonkin, M., 2012. Utility of bromide and heat tracers for aquifer characterization affected by highly transient flow conditions. Water Resources Research 48.

McDonald, M.G., Harbaugh, A.W., 1988. A modular three-dimensional finite-difference ground-water flow model. US Geological Survey.

Michalak, A.M., Kitanidis, P.K., 2004. Estimation of historical groundwater contaminant distribution using the adjoint state method applied to geostatistical inverse modeling. Water Resources Research 40.

Mirghani, B.Y., Mahinthakumar, K.G., Tryby, M.E., Ranjithan, R.S., Zechman, E.M., 2009. A parallel evolutionary strategy based simulation–optimization approach for solving groundwater source identification problems. Advances in Water Resources 32, 1373–1385.

Neupauer, R.M., Borchers, B., Wilson, J.L., 2000. Comparison of inverse methods for reconstructing the release history of a groundwater contamination source. Water Resources Research 36, 2469–2475.

Reynolds, A.C., Zafari, M., Li, G., 2006. Iterative forms of the ensemble Kalman filter, in: ECMOR X-10th European Conference on the Mathematics of Oil Recovery.

Sidauruk, P., Cheng, A.D., Ouazar, D., 1998. Ground water contaminant source and transport parameter identification by correlation coefficient optimization. Ground Water 36, 208–214.

Skaggs, T.H., Kabala, Z., 1994. Recovering the release history of a groundwater contaminant. Water Resources Research 30, 71–79.

Skaggs, T.H., Kabala, Z., 1995. Recovering the history of a groundwater contaminant plume: Method of quasi-reversibility. Water Resources Research 31, 2669–2673.

Sun, A.Y., 2007. A robust geostatistical approach to contaminant source identification. Water Resources Research 43.

Sun, A.Y., Painter, S.L., Wittmeyer, G.W., 2006a. A constrained robust least squares approach for contaminant release history identification. Water Resources Research 42.

Sun, A.Y., Painter, S.L., Wittmeyer, G.W., 2006b. A robust approach for iterative contaminant source location and release history recovery. Journal of Contaminant Hydrology 88, 181–196.

Van Leeuwen, P.J., Evensen, G., 1996. Data assimilation and inverse methods in terms of a probabilistic formulation. Monthly Weather Review 124, 2898–2913.

Wang, H., Jin, X., 2013. Characterization of groundwater contaminant source using Bayesian method. Stochastic environmental research and risk assessment 27, 867–876.

Woodbury, A., Sudicky, E., Ulrych, T.J., Ludwig, R., 1998. Three-dimensional plume source reconstruction using minimum relative entropy inversion. Journal of Contaminant Hydrology 32, 131–158.

Woodbury, A.D., Ulrych, T.J., 1996. Minimum relative entropy inversion: Theory and application to recovering the release history of a groundwater contaminant. Water Resources Research 32, 2671–2681.

Xu, T., Gómez-Hernández, J.J., 2015a. Inverse sequential simulation: A new approach for the characterization of hydraulic conductivities demonstrated on a non-Gaussian field. Water Resources Research 51, 2227–2242.

Xu, T., Gómez-Hernández, J.J., 2015b. Probability fields revisited in the context of ensemble Kalman filtering. Journal of Hydrology 531, 40–52.

Xu, T., Gómez-Hernández, J.J., 2016. Joint identification of contaminant source location, initial release time and initial solute concentration in an aquifer via ensemble Kalman filtering. Water Resources Research 52.

Xu, T., Gómez-Hernández, J.J., 2018. Simultaneous identification of a contaminant source and hydraulic conductivity via the restart normal-score ensemble Kalman filter. Advances in Water Resources 112, 106–123.

Xu, T., Gómez-Hernández, J.J., Zhou, H., Li, L., 2013. The power of transient piezometric head data in inverse modeling: An application of the localized normal-score EnKF with covariance inflation in a heterogenous bimodal hydraulic conductivity field. Advances in Water Resources 54, 100–118.

Zanini, A., Woodbury, A.D., 2016. Contaminant source reconstruction by empirical Bayes and Akaike's Bayesian Information Criterion. Journal of contaminant hydrology 185, 74–86.

Zeng, L., Shi, L., Zhang, D., Wu, L., 2012. A sparse grid based Bayesian method for contaminant source identification. Advances in Water Resources 37, 1–9.

Zhang, J., Zeng, L., Chen, C., Chen, D., Wu, L., 2015. Efficient Bayesian experimental design for contaminant source identification. Water Resources Research 51, 576–598.

Zheng, C., 2010. Technical Report. Technical Report to the US Army Engineer Research and Development Center.