Simultaneous identification of a non-point contaminant source with Gaussian spatially distributed release and heterogeneous hydraulic conductivity in an aquifer using the LES-MDA method

Wenjun Zhang[a,b], Teng Xu[a,b,*], Zi Chen[c], J. Jaime Gómez-Hernández[d], Chunhui Lu[a,b], Jie Yang[a,b], Yu Ye[a,b], Miao Jing[a,b]

[a] *The National Key Laboratory of Water Disaster Prevention, Hohai University, Nanjing, China*
[b] *Yangtze Institute for Conservation and Development, Hohai University, Nanjing, China*
[c] *Nanjing Center, China Geological Survey, Nanjing, China*
[d] *Institute of Water and Environmental Engineering, Universitat Politècnica de València, Valencia, Spain*

Abstract

Space-temporal distribution of the contaminant plumes and aquifer properties is critical for groundwater management. However, most previous studies have focused on point source identification, barely exploring the identification of non-point sources. Xu et al. (2022) proposed to identify non-point sources but did not consider uncertainties in aquifer properties and release mass loading. In this work, we have implemented an application of the localized ensemble smoother with multiple data assimilation (LES-MDA) for the simultaneous identification of Gaussian hydraulic conductivities and non-point source parameters including Gaussian release mass-loading by assimilating both piezometric head and concentration observations in a synthetic confined aquifer. The results prove that the LES-MDA is not only capable of providing accurate identification of the spatial architecture of non-point contaminant sources and related release parameters (such as initial release time, and release duration) but also spatially heterogeneous release mass-loading and hydraulic conductivities.

Keywords: Non-point contaminant source identification; Data assimilation; Ensemble

*Corresponding author

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

smoother with multiple data assimilation; Localization

1. Introduction

Accurate prediction of contaminant plumes in time is critical for groundwater contamination remediation and management. When contaminant sources and other hydrologic information are known, contaminant plumes can be predicted based on a solute transport equation calculation (Gómez-Hernández and Wen, 1994; Li et al., 2011a,b). However, in reality, due to scarce measurement data, technological limitations, and the nature of concealment and lag of contaminant transport (Russell and Shogren, 2012), it is a huge challenge to figure out contaminant source information (e.g., location, shape, release concentration, release duration) and aquifer properties.

In the past, source identification studies mainly focused on contaminant source parameters and rarely considered uncertainties in aquifer properties simultaneously due to computational burdens and technical limitations (Gorelick et al., 1983; Aral et al., 2001; Sun et al., 2006; Dokou and Pinder, 2009; Yeh et al., 2014; Xu and Gómez-Hernández, 2016; Cupola et al., 2015; Ayvaz, 2016). However, uncertainties in aquifer properties are widespread in reality and well-identified (Xu et al., 2013a,b; Xu and Gómez-Hernández, 2015; Zhan et al., 2022), and they should be taken into account in the identification of source information. Recently, with the development of computational techniques and inverse modeling approaches (Wen et al., 1999; Zhou et al., 2014), considerable research has sprung up on the topic of simultaneous identification of source and aquifer parameters. For example, Wagner (1992) uses nonlinear optimization to simultaneously estimate groundwater flow model parameters and single point source properties in a two-zone aquifer; Datta et al. (2009) developed an optimization algorithm for simultaneous pollution source identification and parameter estimation in groundwater systems; Koch and Nowak (2016) proposed an inverse Bayesian methodology to determine the permeability and the DNAPL contaminant architecture ensembles generated from a stochastic multiphase model in a 3D aquifer; Xu and Gómez-Hernández (2018) proposed a variant of ensemble Kalman filter (EnKF), restart-EnKF, to simultaneously estimate the source information and hydraulic conductivities in a synthetic aquifer, and later, Chen et al. (2018) and Chen et al. (2021) applied it for the joint identification of contaminant source, aquifer geometry and aquifer properties in sandbox experiment; Mo et al. (2019) proposed to use a deep neural network (DNN) coupled with a version of the ensemble smoother algorithm to estimate source information and high-dimensional conductivities. later, Zhang et al. (2020) developed a variant of the above method for the joint estimation of multi-component reactive parameters and contaminant transport information; Wang et al. (2022) constructed a kriging surrogate model algorithm to simultaneously identify source characteristics and sub-zone aquifer parameters; Dodangeh et al. (2022) combined artificial neural networks (ANN) with a variant of the EnKF for the identification of source properties with anisotropic conductivities in a 3D coastal aquifer. The reader is referred to a recent review paper by Gómez-Hernández and Xu (2022), which analyzed nearly 160 papers published since 1981 on contaminant source identification (Sonnenborg et al., 1996; Duffy and Brandes, 2001; Michalak and Kitanidis, 2002, 2003, 2004a,b).

Note that, in terms of the discharge scale of the contaminant source, the source can be classified into point and non-point. Point contaminant sources are small in scale and normally emit through a fixed pipeline, while non-point sources are relatively large in scale and have a random release (Ice, 2004). However, as mentioned above, most studies focus on point source information identification, while only a few studies have been done on non-point source identification. Even so, in these studies, the non-point sources are simply treated with a regular spatial architecture. For instance, both Jin et al. (2009) and Mahinthakumar and Sayeed (2005) estimated an areal source, which was assumed to be a rectangular prism with uniform concentration, using a genetic algorithm-local search algorithm; Mirghani et al. (2009) characterized a rectangular non-point source by identifying centroids, whose sizes were assumed to be known, using a parallel simulation-optimization approach; Ayvaz (2016) estimated a non-point source using a hybrid simulation-optimization approach, where the spatial architecture was randomly generated by the juxtaposition of a few aquifer discretization cells; Xu et al. (2022) simultaneously characterized a non-point source approximated by an ellipse and its relative release information using the ensemble smoother with multiple data assimilation (ES-MDA); Pan et al. (2021) simultaneously identified release intensities in three potential non-point sources and a hydraulic conductivity field containing four homogeneous zones using a deep regularization neural network-hybrid heuristic algorithm.

However, unlike the point source identification, the studies on the non-point source identification still only remain on the source location, homogeneous release intensities and spatial architecture (treated as homogenous or divided into several homogeneous subzones (Pan et al., 2021)). To the best of our knowledge, no study has considered the uncertainty in the spatial distribution of both the non-point source release mass-loading and the hydraulic conductivities. Delineating both parameters at high resolution provides valuable insights into the distribution and extent of contamination. It helps us understand the distribution of contamination and its extent, enabling us to allocate resources more efficiently and effectively and is crucial for further effective remediation planning and decision-making. Moreover, once the number of required updated unknown parameters is large, it leads to an increased computational cost, which can be mitigated by reducing the ensemble size. However, employing a smaller ensemble size in ensemble-based data assimilation algorithms brings about certain disadvantages and raises concerns (e.g., filter inbreeding and spurious correlation), which can be solved by the localization technique (Xu et al., 2013b). Therefore, in this work, we further demonstrate the applicability of the localized ensemble smoother with multiple data assimilation (LES-MDA) for the simultaneous identification of spatial architecture of an elliptical non-point source contaminant source and both spatially heterogeneous release mass-loading and hydraulic conductivities by assimilating piezometric heads and concentrations with a small ensemble size.

The remainder of the paper is organized as follows: Section 2 presents the groundwater flow and solute transport equations and the algorithmic description of the LES-MDA. The test and analysis of the method in a synthetic case are shown in Section 3 and Section 4, respectively. Finally, the paper concludes with the discussion presented in Section 5 and a comprehensive summary provided in Section 6.

2. Methodology

2.1. Groundwater flow and solute transport

In this work, we assume that inert contaminants spread under a transient groundwater flow, only attributed to advection and dispersion transport mechanisms. Hence, the governing equations for the state forecast include the three-dimensional transient groundwater flow and contaminant transport shown in Eq. (1) (Bear, 1972) and Eq. (2) (Zheng, 2010), respectively:

$$S_s \frac{\partial H}{\partial t} = \nabla \cdot (K \nabla H) + W, \tag{1}$$

where S_s is the specific storage [L^{-1}]; t is the simulation time [T]; K is the hydraulic conductivity [LT^{-1}]; $\nabla \cdot$ is the divergence operator; ∇ is the gradient operator; W is sources and sinks per unit volume [T^{-1}]; and H is the hydraulic head [L] generating the flow velocity vector through $v = (-K\nabla H)/\theta$ in time, and it is treated as an input to the solute transport equation:

$$\frac{\partial(\theta C)}{\partial t} = \nabla \cdot [\theta(D_m + \alpha v) \cdot \nabla C] - \nabla \cdot (\theta v C) - q_s C_s, \tag{2}$$

where C is the contaminant source concentration [ML^{-3}], regarded as the state variable together with H for subsequent assimilations in this study; t is the simulation time [T]; θ is the effective porosity [-]; D_m is the molecular diffusion coefficient [$\mathrm{L}^2\mathrm{T}^{-1}$]; α denotes the dispersivity tensor [L]; q_s denotes the volumetric flow rate per unit volume [T^{-1}]; and C_s

denotes the concentration of the sources or sinks [ML^{-3}].

In particular, the transient groundwater flow equation is solved numerically using the MODFLOW code with finite differences (McDonald and Harbaugh, 1988); and the contaminant transport equation is solved using the MT3DMS code (Zheng, 2010).

2.2. The localized ensemble smoother with multiple data assimilation

The ensemble smoother (ES) proposed by Van Leeuwen and Evensen (1996) is proven to be optimal to address linear state-transfer equations with Gaussian error statistics by assimilating all observations for all time steps at once, however, it is failed for non-linear problems (e.g., Evensen and Van Leeuwen, 2000; Crestani et al., 2013). To deal with this problem, the ES-MDA proposed by Emerick and Reynolds (2013) is developed by combining an iterative scheme with the ES. It also contains two main steps in nature to the ES algorithm: forecast and update. In the forecast step, the forecast equation is essentially the same as the ES, where the forecast state variables at the j^{th} assimilation iteration U_j^f are forecasted based on initial state variables U_0 and parameters obtained from the last iteration P_{j-1}^a by the state forecast equations $\psi(\cdot)$ involving groundwater flow equation and solute transport equation introduced above:

$$U_j^f = \psi(U_0, P_{j-1}^a), \tag{3}$$

.

In the update step, the updated parameters at the j^{th} assimilation iteration P_j^a are refined based on the parameters at the last assimilation iteration P_{j-1}^a and the discrepancy between forecasted state variables $U_j^{f,o}$ and observations at observation locations $U^o + \sqrt{a_j}\varepsilon_j$.

$$P_j^a = P_{j-1}^a + K_j(U^o + \sqrt{a_j}\varepsilon_j - U_j^{f,o}), \tag{4}$$

with

$$K_j = G_{PU,j}\left(G_{UU,j} + a_j R\right)^{-1}. \tag{5}$$

where K_j is the Kalman gain, a function of the cross-covariance between parameters and state variables $G_{PU,j}$ at the observation locations at all time steps $G_{PU,j}$, and the covariance between state variable observations at all time steps $G_{UU,j}$. ε_j denotes the observation error with observation error covariance R, being magnified by a sequence of inflation coefficient a_j due to the multiple data assimilation iterations. Note that the sum of one over the inflation coefficient should be equal to 1, and the inflation coefficients for observation error will be equal to the number of iterations, following the recommendations by Emerick and Reynolds (2013). They have shown that using decreasing inflation coefficients only leads to marginal improvements compared to using the inflation coefficients equal to the number of iterations.

$$\sum_{j=1}^{N_a} \frac{1}{a_j} = 1 \tag{6}$$

where N_a is the number of the iteration steps. As mentioned, the objective of this work is to simultaneously identify continuous heterogeneous hydraulic conductivities and non-point contaminant source parameters, including initial release time, release duration, source spatial architecture, and heterogeneous spatial distribution of release mass-loading by assimilating piezometric heads and concentrations, besides, the source spatial architecture is approximated by an ellipse. Therefore, the augmented state variable vector U is built containing both piezometric heads H and concentrations C; and the augmented parameter vector P is built containing the x and y coordinates of the ellipse's center point Xs [L] and Ys [L], the semi-major and semi-minor axes Ra [L] and Rb [L], the clockwise rotation angle B [°], the initial release time Ti [T], the release duration ΔT [T], the heterogenous log mass-loading rate lnM [MT^{-1}] and log-conductivities lnK [LT^{-1}]:

$$U = \begin{bmatrix} H & C \end{bmatrix}^\intercal. \tag{7}$$

$$P = \begin{bmatrix} Xs & Ys & Ra & Rb & B & Ti & \Delta T & lnM & lnK \end{bmatrix}^\intercal. \tag{8}$$

Xu et al. (2021, 2022) have demonstrated that the ES-MDA bears the ability to identify Gaussian distributed conductivities or simple non-point source information. However, since ES-MDA is an ensemble-based data assimilation algorithm, it suffers from the same drawback when the ensemble size is considerably smaller than the number of measurements to be assimilated, that is, the ensemble covariance emerges as an unreal correlation (Chen and Oliver, 2010). The spurious correlations enlarge the update region by using observations that would not be correlated with the updates, and although the analysis error decreases in the vicinity of the observations, the harm of the increased error across the whole domain is much greater than the weak benefit (Lorenc, 2003). To remove the spurious correlations, the localization is applied in the covariance derived from the Kalman gain, which controls the extent of correlations in the empirical cross-covariance between model parameters and state variables, or between state variables. Thus, Eq.5 can be replaced by:

$$K_j = \gamma_{PU,j} \circ G_{PU,j}(\gamma_{UU,j} \circ G_{UU,j} + a_j R)^{-1}, \tag{9}$$

with

$$\gamma_{PU}(e) = \gamma_{UU}(e) = \begin{cases} -\frac{1}{4}(\frac{e}{f})^5 + \frac{1}{2}(\frac{e}{f})^4 + \frac{5}{8}(\frac{e}{f})^3 - \frac{5}{3}(\frac{e}{f})^2 + 1 \quad \text{for} \quad 0 \leqslant e \leqslant f; \\ \frac{1}{12}(\frac{e}{f})^5 - \frac{1}{2}(\frac{e}{f})^4 + \frac{5}{8}(\frac{e}{f})^3 + \frac{5}{3}(\frac{e}{f})^2 - 5(\frac{e}{f}) + 4 - \frac{2}{3}(\frac{e}{f})^{-1}, \quad \text{for} \quad f < e \leqslant 2f; \\ 0 \quad \text{for} \quad e > 2f. \end{cases} \tag{10}$$

where $\gamma_{PU,j}$ and $\gamma_{UU,j}$ denote the localization functions; $\circ$ denotes the Schur product; e

denotes the Euclidean distance, and f denotes a distance parameter. In current applications of the localization, the fifth-order distance-dependent localization function of Gaspari and Cohn (1999) (see Eq.10) is widely used to remove spurious correlations with respect to the updates of continuity (e.g., Hamill et al., 2001; Houtekamer and Mitchell, 2001; Houtekamer et al., 2005).

2.3. Testing Criteria

As testing criteria, first, we evaluate the degree of uncertainty of the updated range of non-point sources using the probability of the source location, which is a fraction of the cumulative value of the indicator function. When the probability is getting close to one, this indicates that the uncertainty is getting vanishing, and vice versa.

$$P_i = \frac{1}{N_r} \sum_{j=1}^{N_r} I_{j,i}, \tag{11}$$

where P_i is the probability of source location at cell i; N_r is the number of the realizations; $I_{j,i}$ is the indicator function at cell i for the j^{th} realization, with a value equal to 1 if the source is present, and 0 otherwise.

Second, we use the average absolute bias (AAB) to measure the accuracy of the updated source parameters reproducing the reference one by calculating the average absolute misfit between the updated source parameters and the reference value, for each of the source parameters of interest except for lnM and lnK as:

$$AAB = \frac{1}{N_r} \sum_{j=1}^{N_r} |S_j - S_{ref}|\,, \tag{12}$$

where S_j is the source parameter value (except for lnM and lnK) for the j^{th} realization; S_{ref} is the corresponding reference source parameter value. Specifically, the calculation of the AAB for lnM and lnK can be written as:

$$AAB_i = \frac{1}{N_r} \sum_{j=1}^{N_r} \left| S_{j,i} - S_{ref,i} \right| , \tag{13}$$

where $S_{j,i}$ is the value of lnM and lnK at cell i for the j^{th} realization; $S_{ref,i}$ is the value of the reference lnM and lnK at cell i.

Third, we use the ensemble spread (ESp) to evaluate the degree of variability of the updated source parameters by calculating the square root of the variance of updated source parameters, for each of the source parameters of interest except for lnM and lnK as:

$$ESp = \sqrt{\sigma_S^2}. \tag{14}$$

where σ_S means the ensemble variance of the source parameters (also except for lnM and lnK). Specifically, the calculation of the (ESp) for lnM and lnK can be written as:

$$ESp_i = \sqrt{\sigma_{S_i}^2}. \tag{15}$$

where σ_{S_i} means the ensemble variance of lnM and lnK at cell i.

Notice that if the ratio ESp/AAB is close to 1, it indicates the performance of the method without filter inbreeding (Xu et al., 2013b, 2022).

3. Application

A two-dimensional synthetic confined aquifer is constructed on a grid of $80 \times 80 \times 1$ cells and the size of each cell is 10 [L] $\times$ 10 [L] $\times$ 80 [L]. A sequence multivariate multi-Gaussian simulation code —the GCOSIM3D program (Gómez-Hernández and Journel, 1993) is used to generate the reference lnK field (see Figure 1), following a multiGaussian distribution with the parameters given in Table 1.

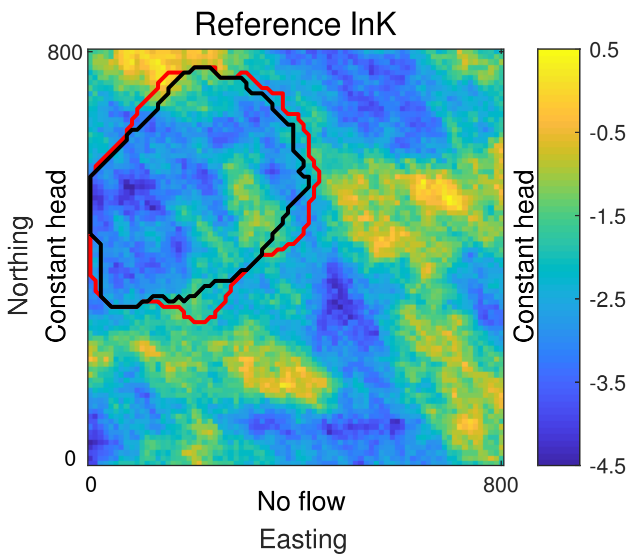

Figure 1: Reference lnK with boundary conditions and suspect contaminant area. The black line indicates the suspect contaminant area for S1 and S2. The red line indicates the suspect contaminant area for S3.

Table 1: Parameters of the random functions used to generate the lnK field.

	Mean	Std.dev.	Variogram	λ_{max}	λ_{min}	Angle
lnK	-2	1	Spherical	300	200	135

In the simulation of transient groundwater flow and solute transport, the east and west boundaries are set as prescribed heads with constant values of 80 [L] and 200 [L], respectively; and the north and south boundaries of the aquifer are impermeable. The initial piezometric head, excluding both the east and west boundaries, is set to 120 [L], and the initial concentration is 0 [MT^{-3}] throughout the domain. Additional parameters for the solute transport are set to be homogeneous: porosity of 0.3 [-], longitudinal dispersivity of 3.0 [L], and transverse dispersivity of 1.5 [L]. The shape of the reference non-point source (see Figure 2) is treated as an ellipse generated with the parameters shown in Table 2. We can also learn from this that the contaminants start to release at time 1381.5 [T] and the duration of the release is 3223.5 [T]. The release mass-loading rates in the source area follow a multiGaussian distribution and are also generated using the GCOSIM3D program with the parameters in Table 3. We deploy 30 observation wells to record the observations of both piezometric heads and concentrations and 2 verification wells for prediction verification (see Figure 2). The observational errors are set to zero mean and 0.01 variance. The total simulation time for both groundwater flow and contaminant transport is set to 15350 [T], and evenly discretized into 100 time steps. Notice that the observations of both piezometric head and concentration are only recorded at the first 50 time steps (at time 7675 [T]).

In this work, to evaluate how well LES-MDA performs for non-point source identification compared to the Localization-Free, we have designed three scenarios for the evaluation, as shown in Table 2. The ensemble size is the same for scenarios S1-S2, with a value of 130; the ensemble size for scenario S3, however, is 500 for comparison. Scenarios S1-S2 differ in that in scenario S1 a localization technique is employed to avoid the effect of spurious correlations induced by the small ensemble size (Xu et al., 2013b), and the distance parameter f is treated as 140 [L] for lnM and 470 [L] for lnK. Note that the localization technique is only used for the lnM and lnK update. Three different numbers of assimilation iterations (0, 1, and 7) for all scenarios are tested. Note that iteration 0 indicates ES without multiple data

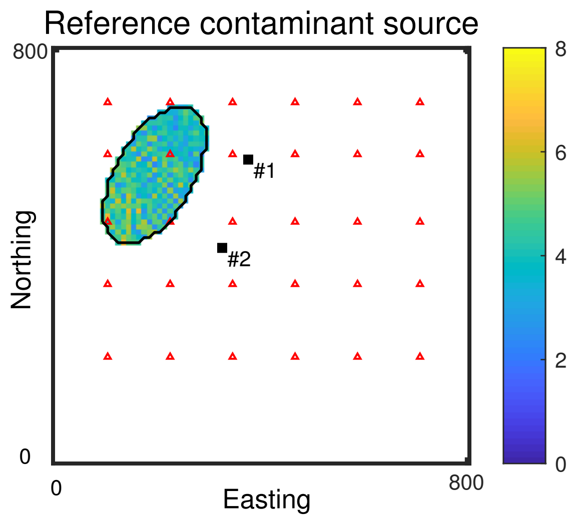

Figure 2: Reference lnM field of contaminant source and well locations. The observation wells correspond to red triangles and two verification wells correspond to black squares.

Table 2: Definition of scenarios

Scenario	S1	S2	S3
Number of realizations	130	130	500
Localization	$\surd$		
Number of assimilation iterations [l]	0,1,7	0,1,7	0,1,7
Contaminant source shape	Ellipse		
x-coordinate of center point of source[Xs]	200		
y-coordinate of center point of source[Ys]	560		
Semi-major axis of source[Ra]	150		
Semi-minor axis of source[Rb]	80		
Clockwise rotation angle[B]	30		
Initial release time[Ti]	1381.5		
Release duration time[ΔT]	3223.5		

Table 3: Parameters of the random functions used to generate the lnM field.

	Mean	Std.dev.	Variogram	λ_{max}	λ_{min}	Angle
lnM	4.605	1	Spherical	300	200	135

Table 4: Suspect range of contaminant source parameters

Parameters	Suspect Range
Xs	160-260
Ys	480-580
Ra	110-210
Rb	50-120
B	0-90
Ti	0-4451.5
ΔT	1688.5-9363.5

assimilation. The suspect parameters related to the non-point source are listed in Table 4.

The initial ensemble consists of 130 realizations in scenarios S1-S2 and 500 realizations in scenario S3, generated from a uniform distribution (see Table 4). Xs is randomly chosen from the uniform distribution $u[160, 260]$, Ys from $u[480, 580]$, Ra from $u[110, 210]$, Rb from $u[50, 120]$, B from $u[0, 90]$, Ti from $u[0, 4451.5]$, and ΔT from $u[1688.5, 9363.5]$. Note that the initial ensembles of parameter realizations for scenarios S1-S2 are the same. These initial geometric parameters generate an initial ensemble of the elliptical source area and the suspect contaminant source area shown in Figure 1. Note that after generating the initial ensemble of the elliptical source area, the initial ensemble of lnM is subsequently generated using the same procedure employed for the reference lnM. Additionally, the initial ensemble of lnK is generated using the same procedure employed for the reference lnK.

4. Results

Figure 3 shows the evolution of the probability of the source location and the underlying potential source area as the number of assimilation iterations increases. In all scenarios, the initial ensemble of probabilities exhibits significant uncertainty. However, the uncertainty decreases with increasing data assimilation and eventually vanishes almost completely by the seventh iteration, where the probabilities are equal to 1 for most of the potential source areas. In addition, we can notice that the potential source areas for scenarios S1 and S3 are closer to the reference source area than those for S2, indicating that the LES-MDA is more efficient and outperforms the ES-MDA when for a small ensemble size in the context of the source area identification.

Figures 4, 5, and 6 show the ensemble mean, AAB and ESp of lnM released from the source for all three scenarios, before and after assimilating the observations at iterations 0, 1 and 7, respectively. When comparing Figure 4 to Figure 2, it becomes apparent that the identification of lnM improves as the number of data assimilation iterations increases, especially for S1 and S3, and the updates are close to the reference lnM at iteration 7, although the lnM for S1 is more concentrated toward the southwest than that for S3. In contrast, the update of lnM for S2 is underestimated due to the numerical nature of the covariance calculation due to the small ensemble size. Figure 5 reveals that the updates in S1 more accurately reproduce the reference lnM compared to those in S2. This improvement in accuracy in S1 is attributed to the implementation of the localization technique, which effectively eliminates spurious correlations induced by the small ensemble size. Although the updates for S1 are not as good as those for S3, the computational burden is substantially reduced. Figure 6 demonstrates that the underestimation of the uncertainty in S2 is removed by the localization employed in S1. However, the uncertainties of the updates in S1 remain slightly larger than those in S3. This discrepancy arises from the application of the localization in the calculation of the cross-covariance.

Figure 7 shows the evolution of the AAB and ESp/AAB of the ensemble values of source parameters including the geometrical parameters (Xs, Ys, Ra, Rb, B) and the release temporal parameters (Ti, ΔT) for all scenarios. For all source parameters, we can see how, for S1 and S3, the AAB of the source parameters decreases as the number of data assimilation iterations increases, while for S2 the AAB of most of the source parameters becomes larger. In addition, the ratio ESp/AAB for the source parameters is too small for S2, indicating small filter inbreeding, while for S1 and S3 the ratio ESp/AAB is closer to 1 than for S2. It

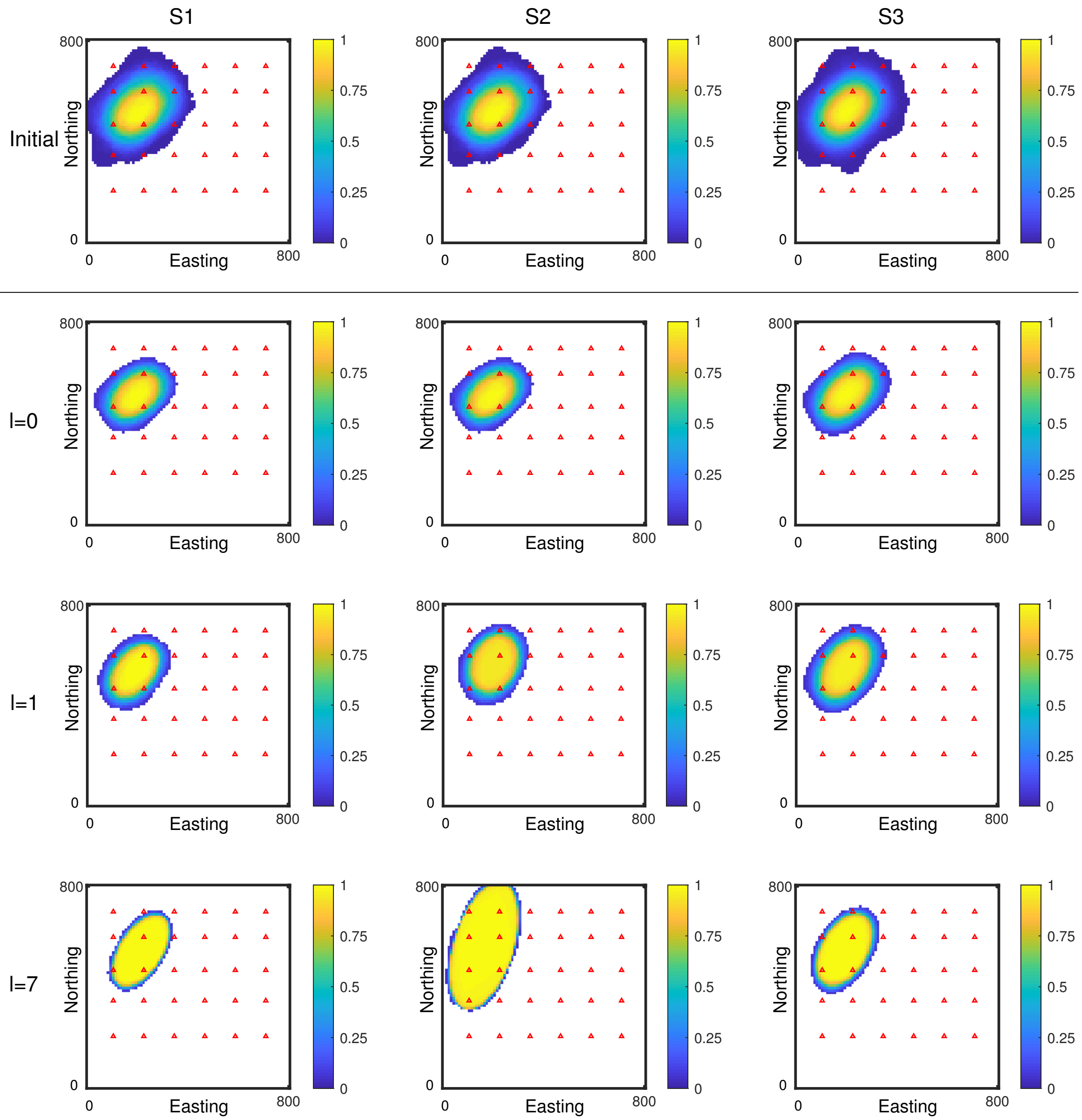

Figure 3: Scenarios S1-S3. Probability of source location as computed from the ellipses given by the parameters updated after the 0^{th}, 1^{st}, and 7^{th} assimilation iterations. Note that the initial ensembles of parameter realizations for scenarios S1-S2 are the same.

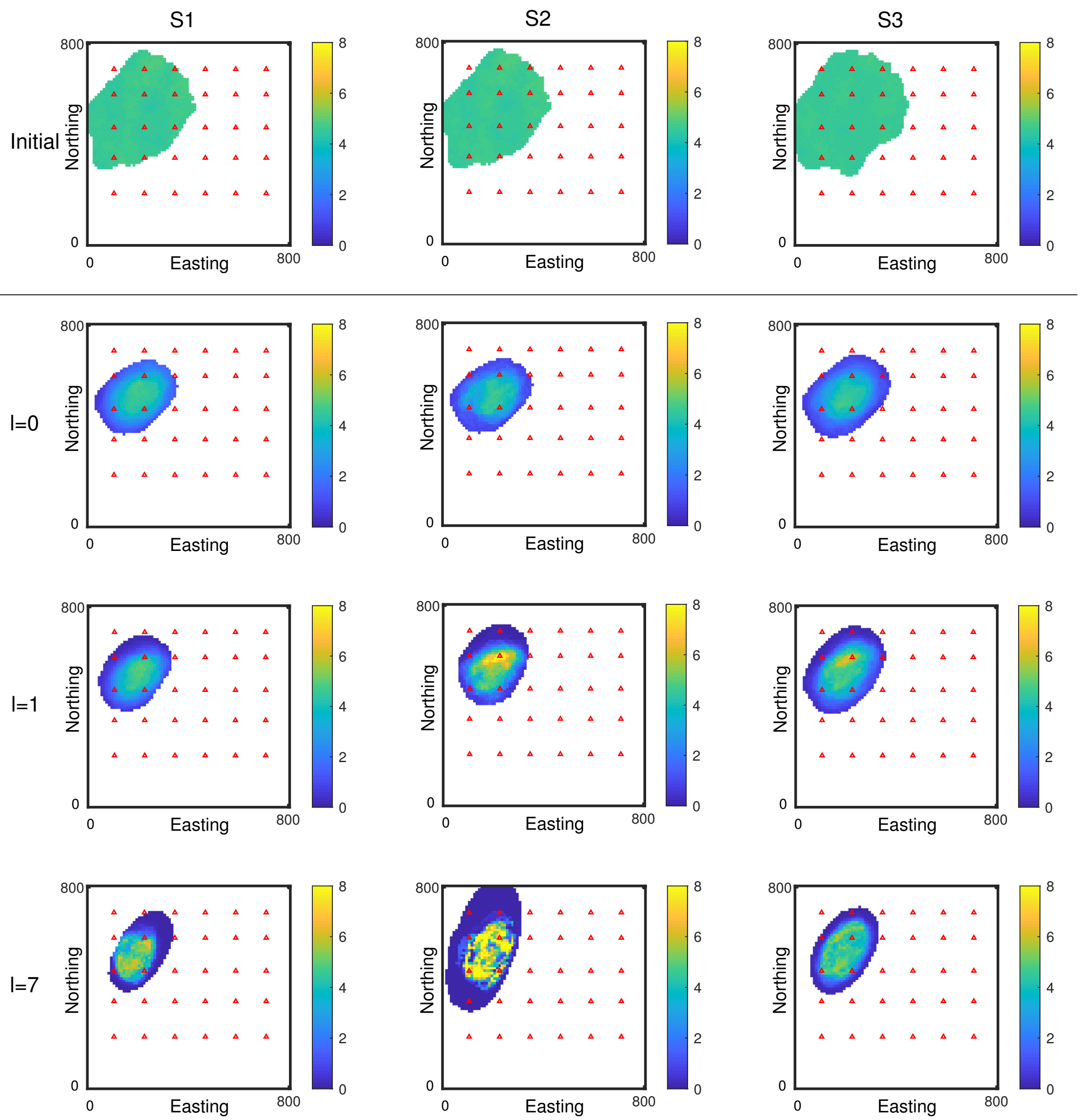

Figure 4: Scenarios S1-S3. Ensemble mean of lnM for the initial and updated ensemble of realizations after the 0^{th}, 1^{st}, and 7^{th} assimilation iterations.

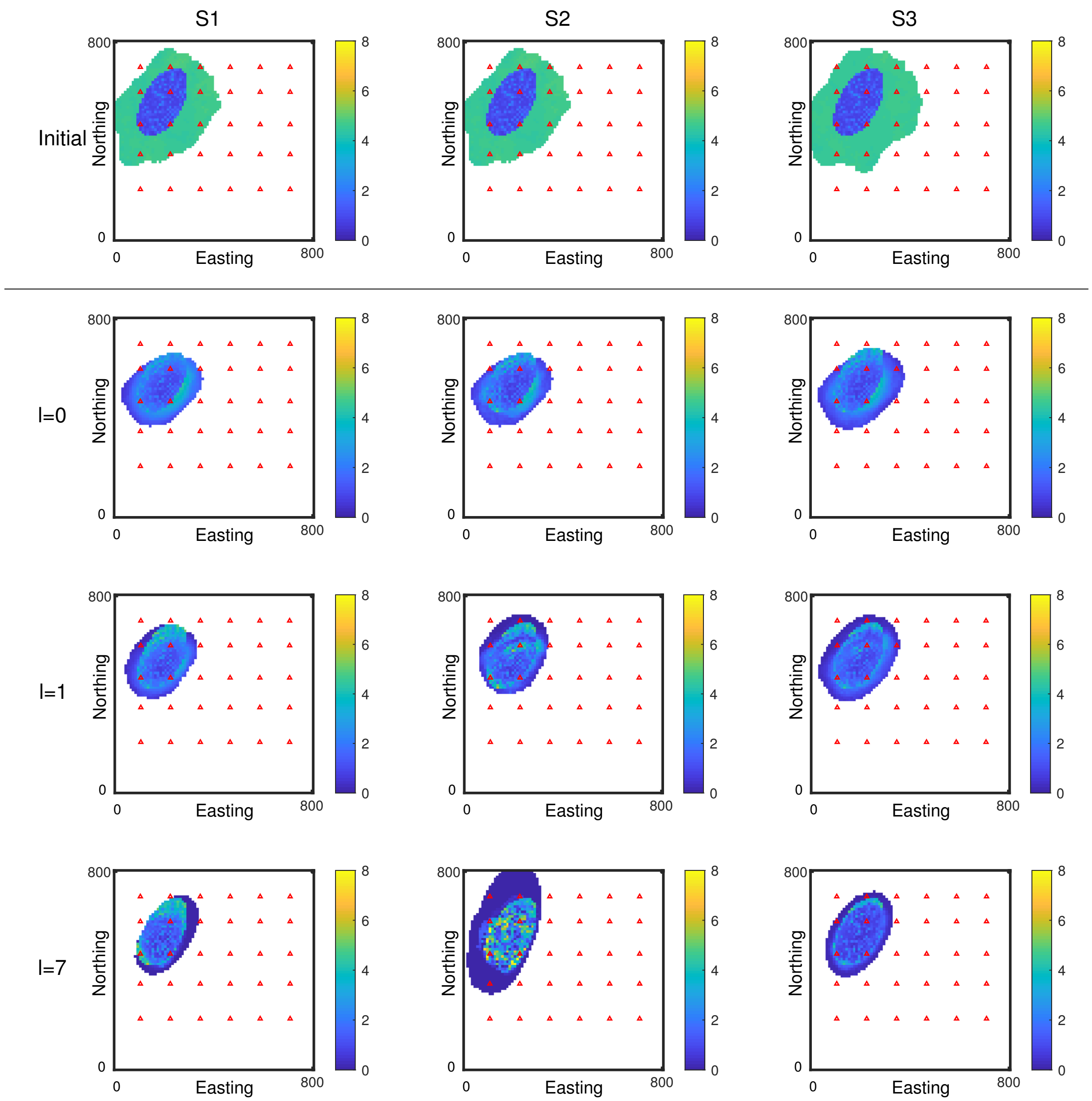

Figure 5: Scenarios S1-S3. AAB computed with the initial and updated ensemble of lnM realizations after the 0^{th}, 1^{st}, and 7^{th} data assimilation iterations.

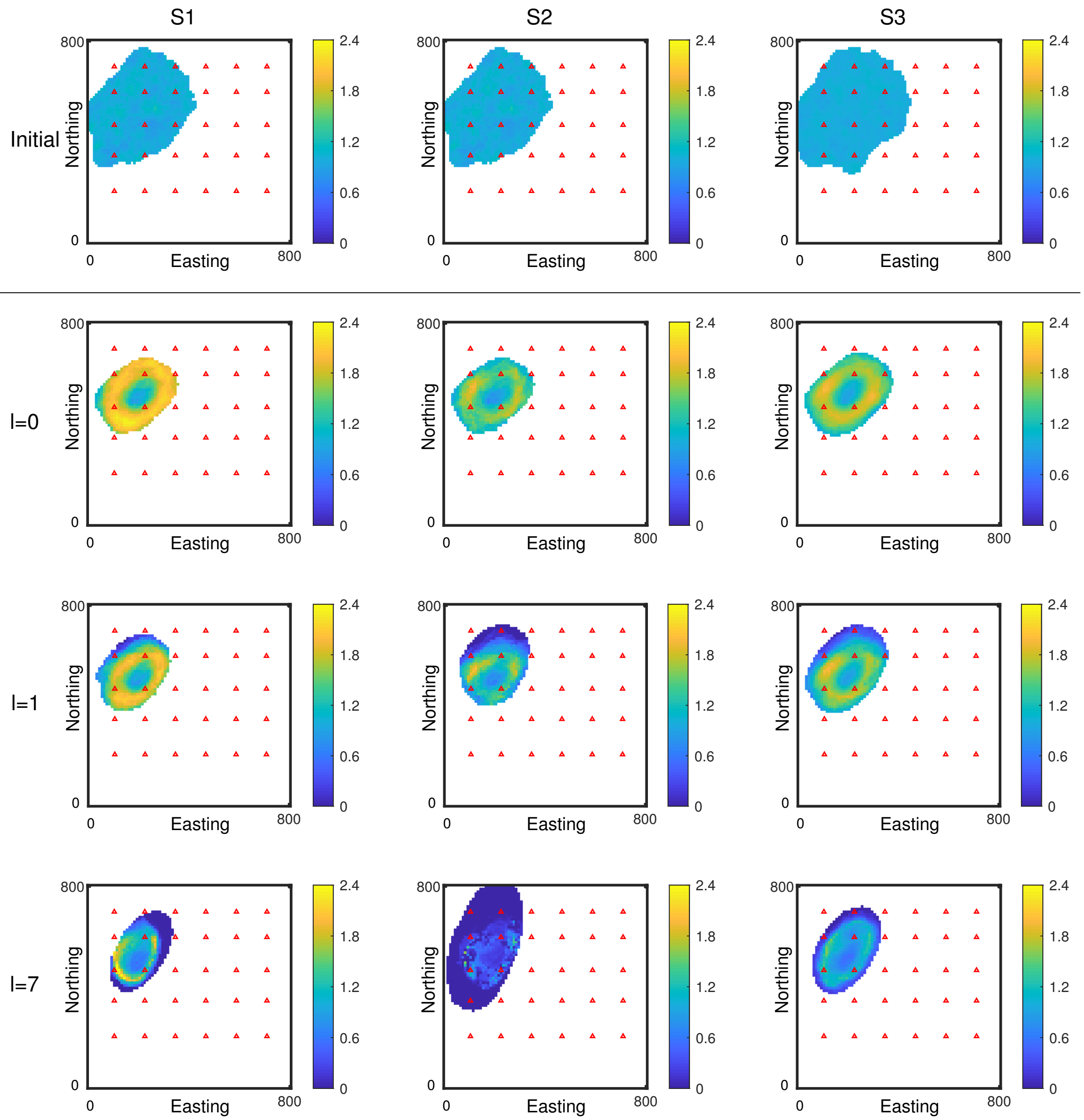

Figure 6: Scenarios S1-S3. ESp computed with the initial and updated ensemble of lnM realizations after the 0^{th}, 1^{st}, and 7^{th} data assimilation iterations.

is shown that LES-MDA can reduce filter inbreeding for small ensemble sizes.

Figures 8 shows the boxplots of the source parameters for all scenarios. We can see that the uncertainty is significant before the update and decreases with increasing iterations. In the final iteration, the ensemble median almost coincides with the true value for all parameters in both S1 and S3, whereas a clear misfit occurs in S2, which is induced by filter inbreeding. However, in scenario S1, the updates for Xs, Ra, and Rb are slightly overestimated, while the updates for Ys are slightly underestimated. Besides, At the cost of time consumption due to the large ensemble size, S3 performs the best, with all parameters close to the true values except for Ra and Rb, which are also slightly overestimated.

Figures 9, 10 and 11 show, sequentially from left to right columns, the ensemble mean, AAB and ESp of lnK computed with the initial and updated ensembles for all scenarios. We can find that the updates of lnK are able to retrieve the main features of the reference, and the AAB and ESp decrease significantly across the entire domain after iterative data assimilation for all three scenarios. When comparing the ensemble mean, AAB, and ESp among the three scenarios, we can see that both S1 and S3 perform more smoothly and accurately than S2. In addition, the ESp values for S2 are very close to zero across the entire domain when compared to those for S1, indicating an underestimation of the uncertainty. This underestimation has been effectively addressed through the use of the localization technique. These findings demonstrate the effectiveness of the localization in dealing with spurious correlations due to the small ensemble size.

To evaluate how well the flow and transport processes reproduced by the methods, we have shown the evolution of the predicted piezometric heads and concentrations in two validation wells (#1, #2), computed based on the initial and updated source parameters and lnK for all scenarios, in Figures 12 and 13, respectively. The uncertainties in the predicted piezometric heads and concentrations are large when computed from the initial source and lnK parameters, and decrease with increasing data assimilation. Specifically, after iteration

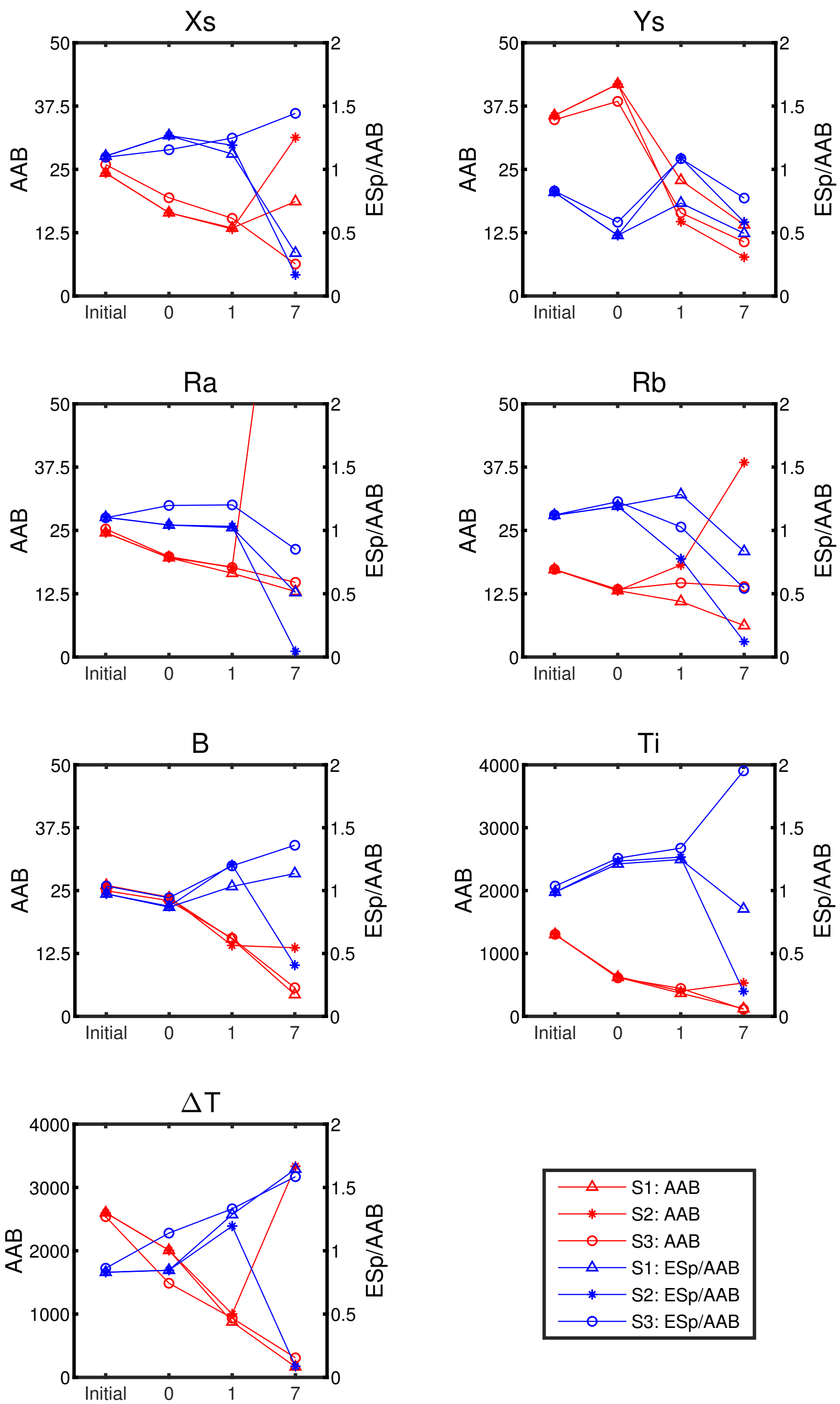

Figure 7: Scenarios S1-S3. AAB and ESp/AAB computed with the initial and updated ensemble of source information parameters including Xs, Ys, Ra, Rb, B, Ti, and ΔT after the 0^{th}, 1^{st}, and 7^{th} data assimilation iterations. The red line corresponds to AAB, and the blue line corresponds to ESp/AAB.

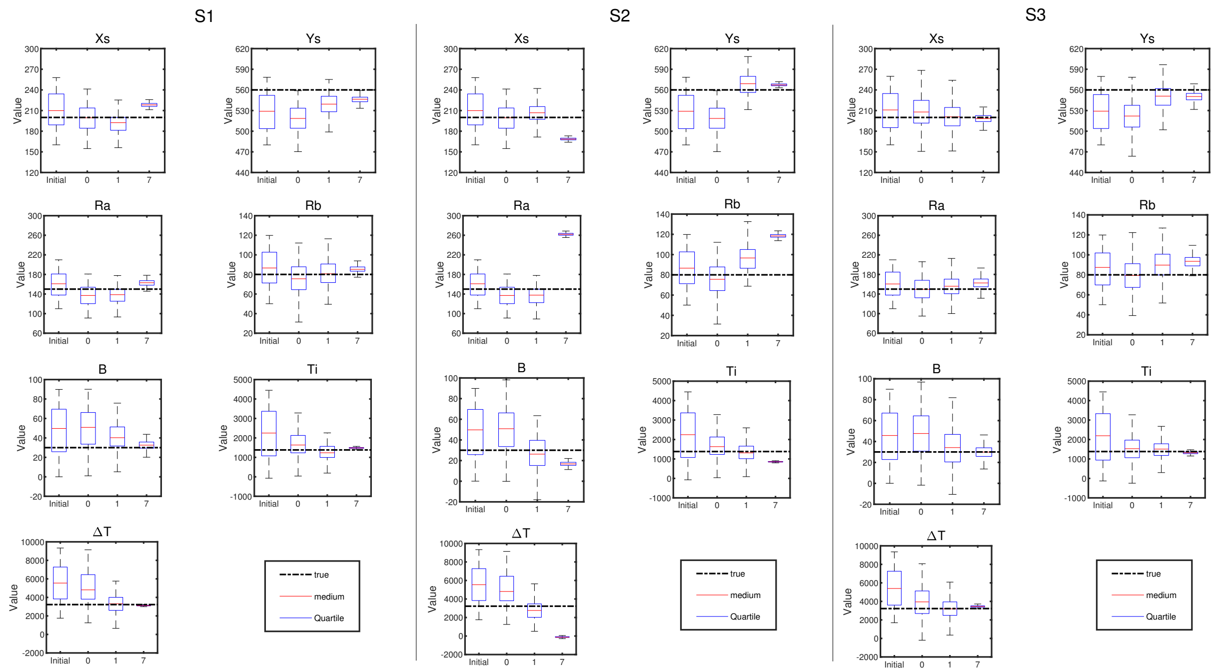

Figure 8: Scenarios S1, S2 and S3. Boxplots computed with the initial and updated ensemble of source information parameters, including Xs, Ys, Ra, Rb, B, Ti and ΔT after the 0^{th}, 1^{st}, and 7^{th} data assimilation. The dashed horizontal black line corresponds to the reference value.

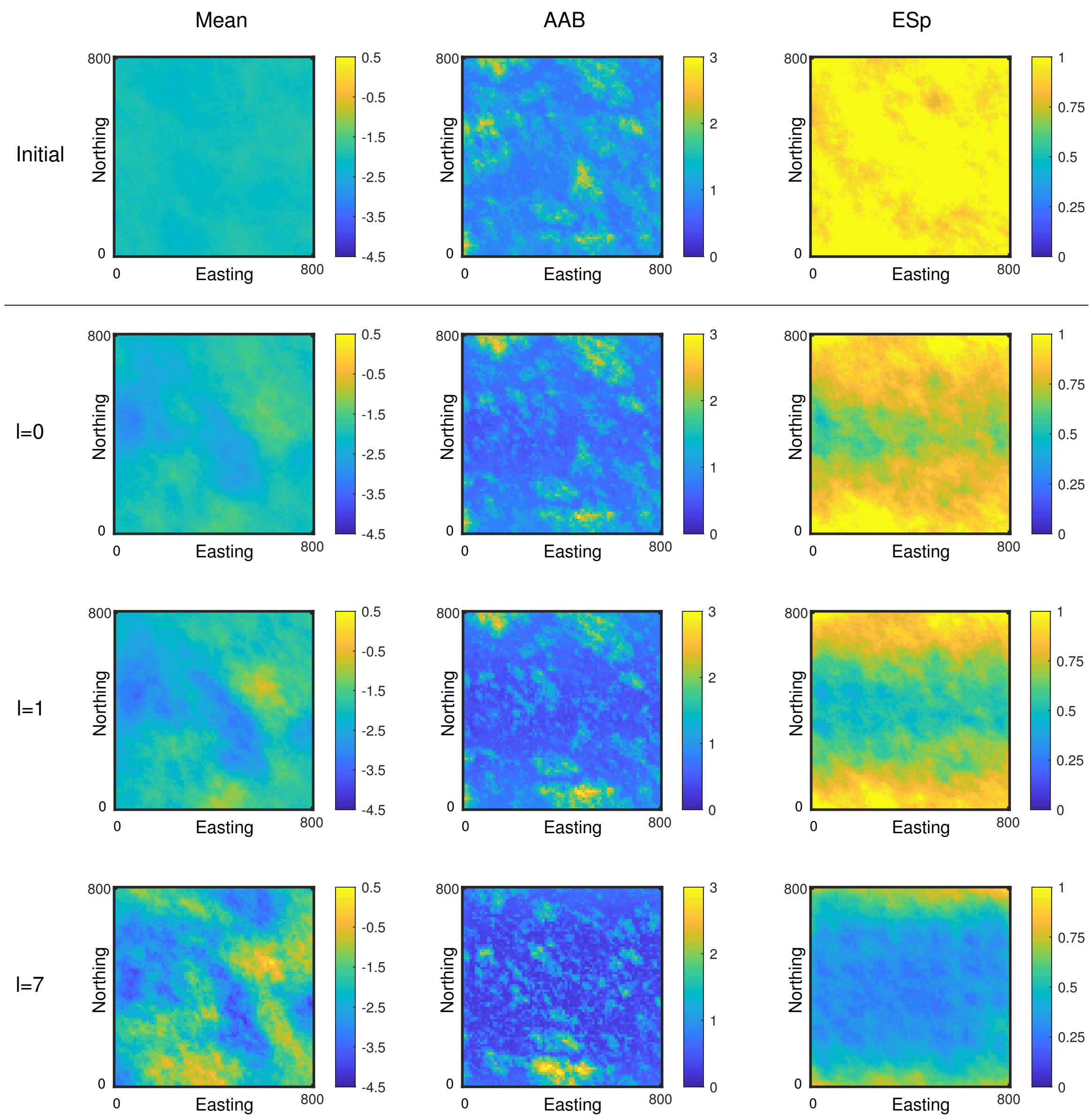

Figure 9: Scenario S1. Ensemble mean(left column), AAB (center column) and ESp (right column) computed with the initial and updated ensemble of lnK after the 0^{th}, 1^{st}, and 7^{th} data assimilation iterations.

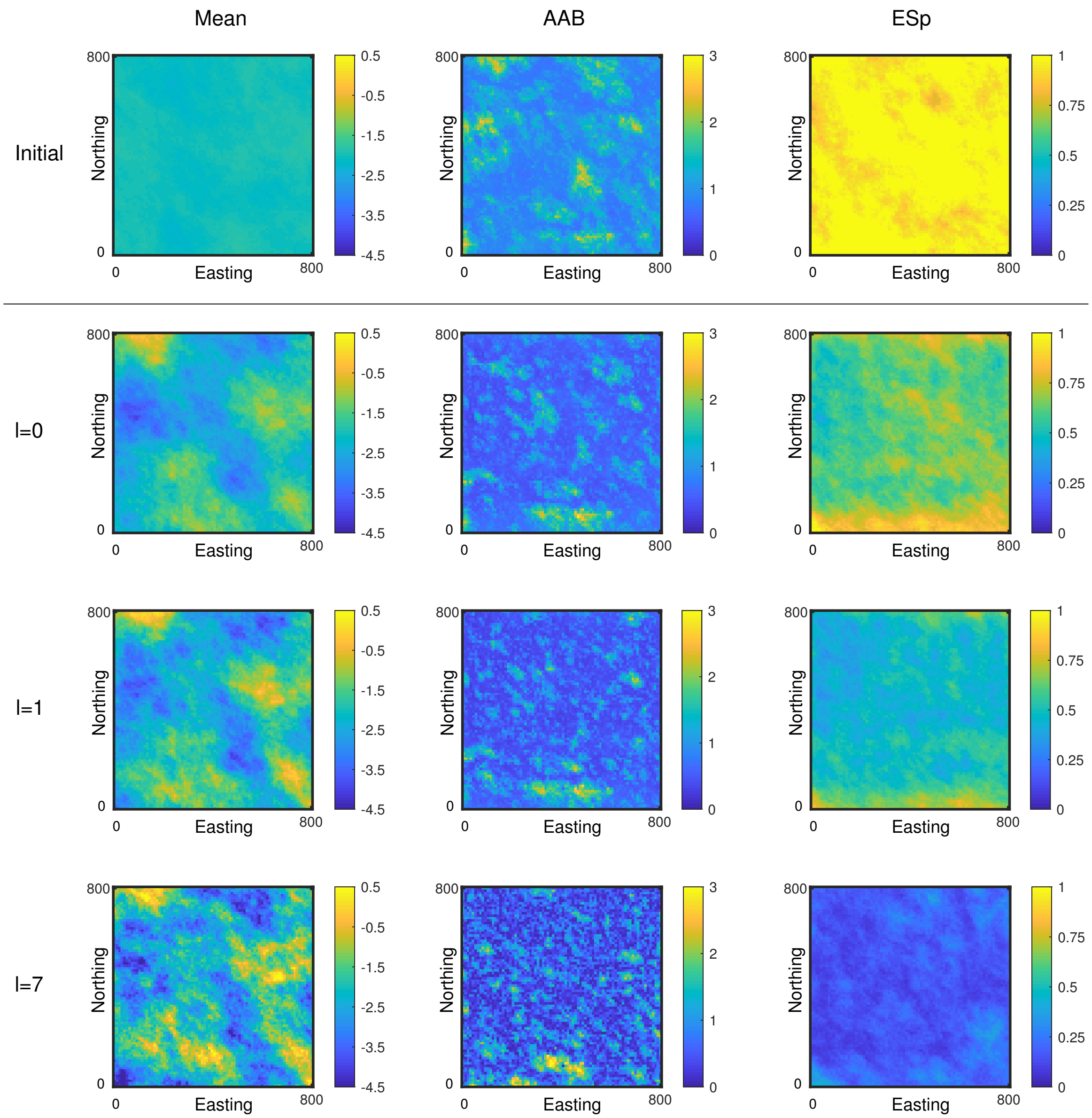

Figure 10: Scenario S2. Ensemble mean(left column), AAB (center column) and ESp (right column) computed with the initial and updated ensemble of lnK after the 0^{th}, 1^{st}, and 7^{th} data assimilation iterations.

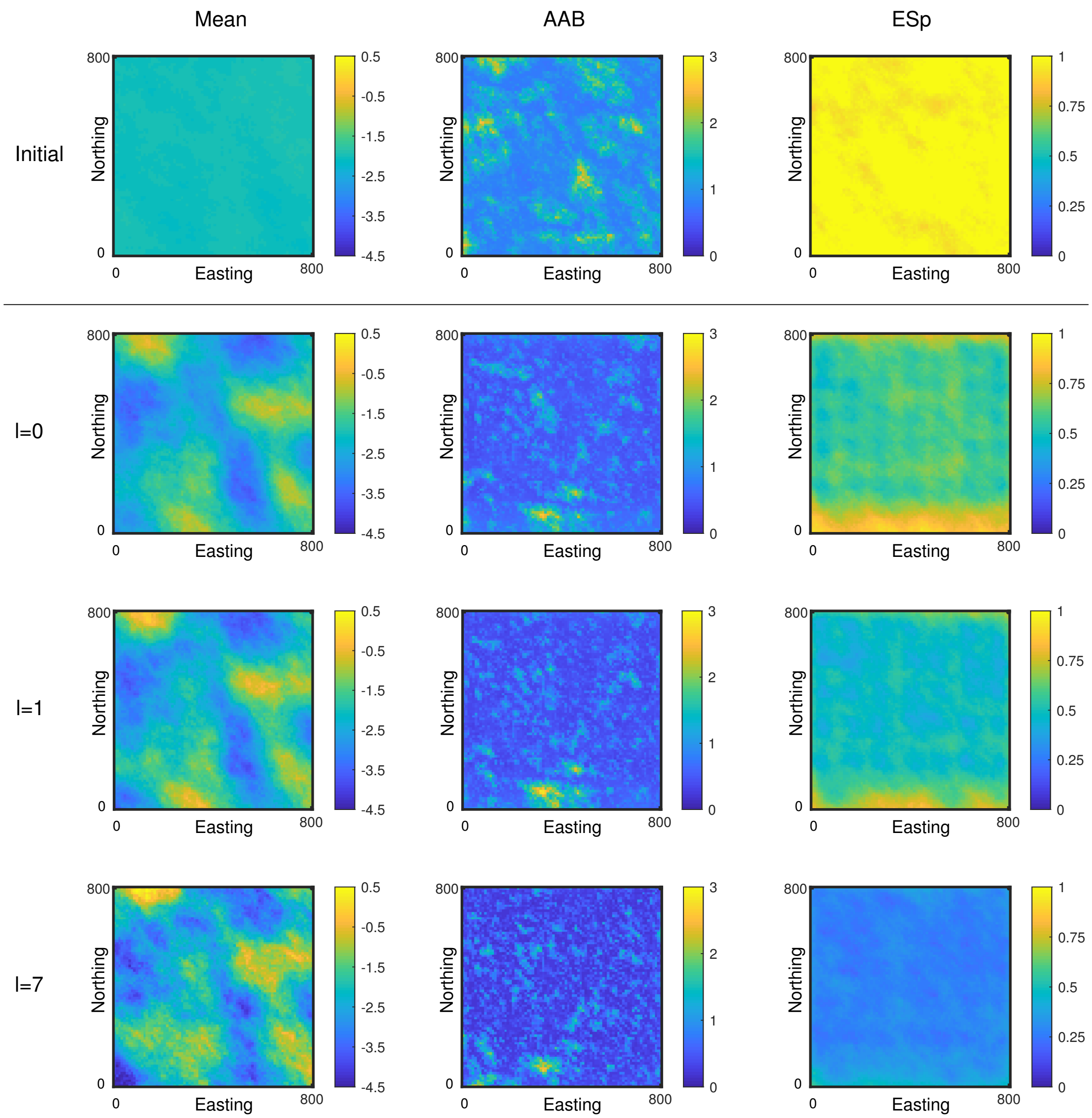

Figure 11: Scenario S3. Ensemble mean(left column), AAB (center column) and ESp (right column) computed with the initial and updated ensemble of lnK after the 0^{th}, 1^{st}, and 7^{th} data assimilation iterations.

7, the piezometric heads of S2 and S3 exhibit similar results with less uncertainty but lower accuracy than S1, where median values of S1 is reproduced almost perfectly and almost coincides with the reference. In contrast, the reference values for S2 and S3 are lower than those of the piezometric heads corresponding to the 5 percentiles of all realizations. Specifically, when comparing S1 with S2, we can find that with the help of the localization, the updates for S1 are not only closer to the true value but also have a smaller underestimation of the uncertainty. However, the reproduced concentration for S2 is significantly underestimated, which can be attributed to the poor estimation of the source parameters.

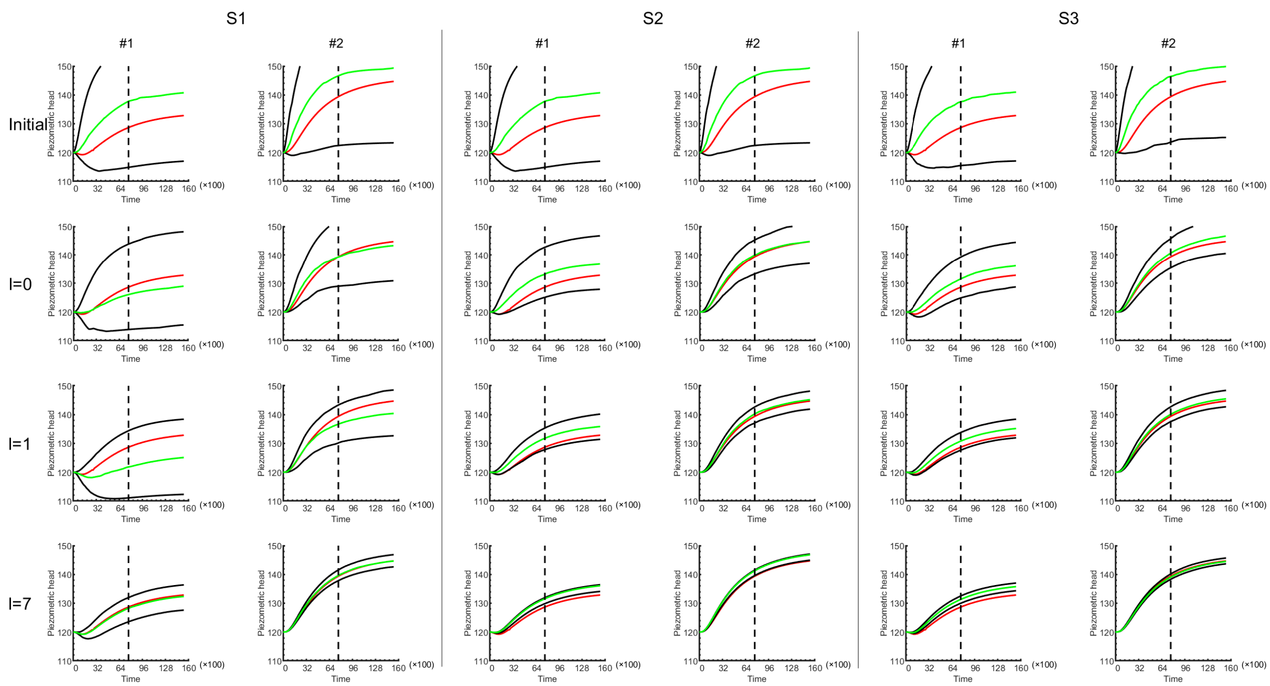

Figure 12: Scenarios S1, S2 and S3. Time evolution of the piezometric heads at the two verification wells #1 and #2 computed with the initial and updated ensembles of lnK after the 0^{th}, 1^{st}, 7^{th} data assimilation. The red line corresponds to the reference field. The black lines correspond to 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.

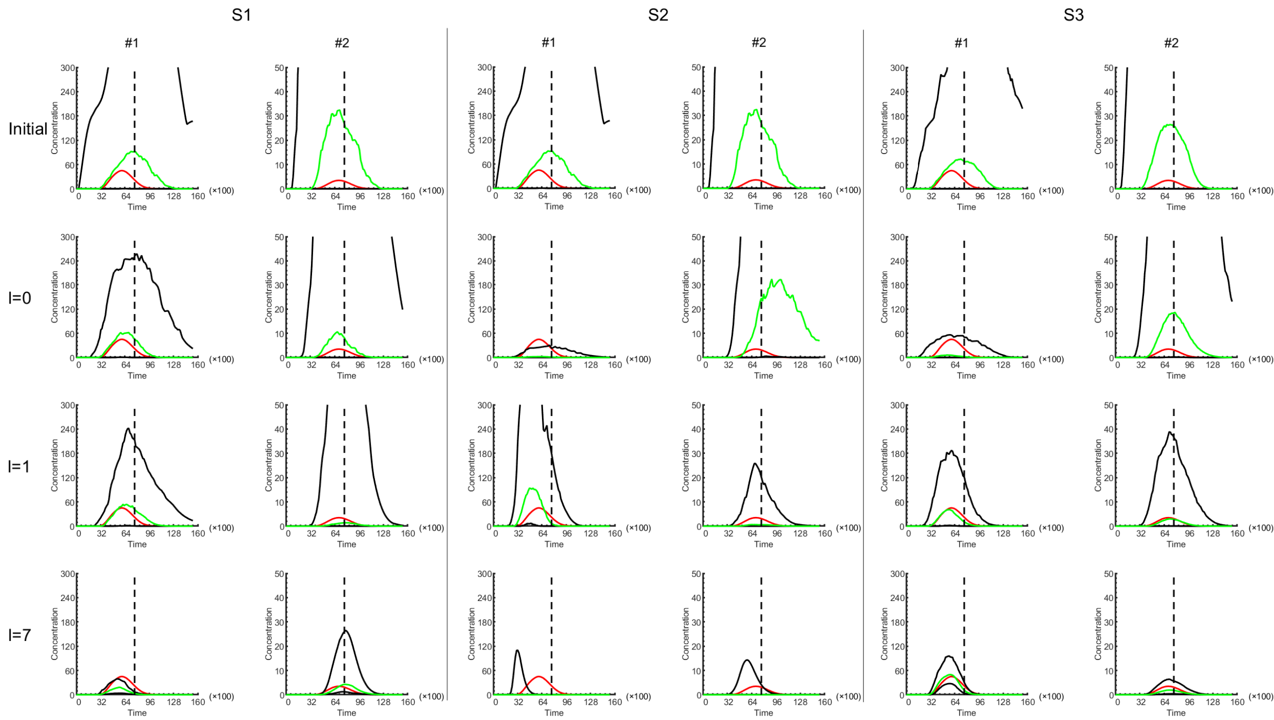

Figure 13: Scenarios S1, S2 and S3. Time evolution of the contaminant concentrations at the two verification wells #1 and #2 computed with the initial and updated ensembles of lnM and source parameters after the 0^{th}, 1^{st}, 7^{th} data assimilation. The red line corresponds to the reference field. 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.

5. Discussion

The aforementioned findings have demonstrated that the LES-MDA is capable of simultaneously characterizing the spatial configuration of an elliptical non-point contaminant source and both spatially variable release mass-loading and hydraulic conductivities within a synthetic confined aquifer. However, the current work is still in its early stages. For future applications in real-world settings, the following aspects will be considered:

(1) The identification of complex spatial architecture of non-point contaminant sources: This study employs the LES-MDA to identify the spatial architecture of non-point contaminant sources with an ellipse shape. However, it remains a challenge to accurately identify the complex spatial architecture of non-point contaminant sources. Our future work will propose a novel method suitable for the identification of complex spatial architecture of non-point contaminant sources.

(2) Performance comparison between homogeneous and heterogeneous release: In this study, the Gaussian release mass-loading may impose a potential computational burden relative to the homogeneous release. Consequently, our future research aims to evaluate the time consumption and efficiency between the homogeneous and heterogeneous release methods for non-point source identification.

(3) The optimization of observation well site layouts: Practical constraints, such as geological features and economic limitations, often dictate the arrangement of observation networks. In our future research, we aim to overcome these limitations by developing a multi-objective optimal well network algorithm, combined with an inverse simulation method, to solve complex non-point source estimation problems at minimal cost.

6. Summary

In this paper, we analyze the capability of the LES-MDA in the joint identification of a heterogeneous conductivity field and a non-point field with spatially heterogeneous mass loading. Our results demonstrate that the LES-MDA is capable of identifying Gaussian distributed hydraulic conductivity fields and elliptical source parameters including position, shape, initial release time, release duration, and Gaussian distributed mass loading. Based on those updated parameters, we are able to give accurate predictions of groundwater flow and contaminant transport.

We also demonstrate that the LES-MDA can effectively eliminate spurious correlations and reduce filter inbreeding when the ensemble size is small compared to the ES-MDA. Furthermore, the LES-MDA is able to give a proper identification of the source parameters with a small ensemble size, whereas the ES-MDA fails and requires a larger ensemble size to obtain proper identification.

Compared to the work by Xu et al. (2022), we further consider the uncertainties of the spatial distribution of the aquifer properties and mass-loading. This is much closer to the real environment. In the next step, we will further investigate how to develop and employ methods in a real-world setting and identify more complex and irregular non-point contaminant sources. Besides, it will be interesting and meaningful to further analyze the sensitivity of the parameters, the impact of different types of ellipse plumes and the effect of different well site layouts in our next work.

Acknowledgements Financial support to carry out this work was received from the financial support from the National Key Research and Development Project (2021YFC3200500), and the National Natural Science Foundation of China (42377046 and 523024911). J.J. Gómez-Hernández acknowledges the financial support funded by MCIN/AEI/10.13039/50110001103 through grant PID2019-109131RB-I00. C. Lu acknowledges the National Natural Science Foundation of China (51879088), Fundamental Research Funds for the Central Universities (B200204002), and the Natural Science Foundation of Jiangsu Province (BK 20190023).

References

Aral, M.M., Guan, J., Maslia, M.L., 2001. Identification of contaminant source location and release history in aquifers. Journal of hydrologic engineering 6, 225–234.

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

Bear, J., 1972. Dynamics of fluids in porous media. American Elsevier Pub. Co., New York, 764pp.

Chen, Y., Oliver, D., 2010. Cross-covariances and localization for enkf in multiphase flow data assimilation. Computational Geosciences 14, 579–601.

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.

Chen, Z., Xu, T., Gómez-Hernández, J.J., Zanini, A., 2021. Contaminant spill in a sandbox with non-gaussian conductivities: Simultaneous identification by the restart normal-score ensemble kalman filter. Mathematical Geosciences 53, 1587–1615.

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. Laboratory sandbox validation of pollutant source location methods. Stochastic Environmental Research and Risk Assessment 29, 169–182.

Datta, B., Chakrabarty, D., Dhar, A., 2009. Simultaneous identification of unknown groundwater pollution sources and estimation of aquifer parameters. Journal of Hydrology 376, 48–57.

Dodangeh, A., Rajabi, M.M., Carrera, J., Fahs, M., 2022. Joint identification of contaminant source characteristics and hydraulic conductivity in a tide-influenced coastal aquifer. Journal of Contaminant Hydrology 247, 103980.

Dokou, Z., Pinder, G.F., 2009. Optimal search strategy for the definition of a dnapl source. Journal of Hydrology 376, 542–556.

Duffy, C.J., Brandes, D., 2001. Dimension reduction and source identification for multispecies groundwater contamination. Journal of contaminant hydrology 48, 151–165.

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

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

Gaspari, G., Cohn, S.E., 1999. Construction of correlation functions in two and three dimensions. Quarterly Journal of the Royal Meteorological Society 125, 723–757.

Gómez-Hernández, J., Wen, X.H., 1994. Probabilistic assessment of travel times in groundwater modeling. Stochastic Hydrology and Hydraulics 8, 19–55.

Gómez-Hernández, J.J., Journel, A.G., 1993. Joint sequential simulation of multigaussian fields, in: Geostatistics Troia'92. Springer, pp. 85–94.

Gómez-Hernández, J.J., Xu, T., 2022. Contaminant source identification in aquifers: a critical view. Mathematical Geosciences 54, 437–458.

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

Hamill, T.M., Whitaker, J.S., Snyder, C., 2001. Distance-dependent filtering of background error covariance estimates in an ensemble kalman filter. Monthly Weather Review 129, 2776–2790.

Houtekamer, P., Mitchell, H., 2001. A sequential ensemble kalman filter for atmospheric data assimilation. Monthly Weather Review 129, 123–137.

Houtekamer, P.L., Mitchell, H.L., Pellerin, G., Buehner, M., Charron, M., Spacek, L., Hansen, B., 2005. Atmospheric data assimilation with an ensemble kalman filter: Results with real observations. Monthly weather review 133, 604–620.

Ice, G., 2004. History of innovative best management practice development and its role in addressing water quality limited waterbodies. Journal of Environmental Engineering 130, 684–689.

Jin, X., Mahinthakumar, G., Zechman, E.M., Ranjithan, R.S., 2009. A genetic algorithm-based procedure for 3d source identification at the borden emplacement site. Journal of Hydroinformatics 11, 51–64.

Koch, J., Nowak, W., 2016. Identification of contaminant source architectures—a statistical inversion that emulates multiphase physics in a computationally practicable manner. Water Resources Research 52, 1009–1025.

Li, L., Zhou, H., Gómez-Hernández, J.J., 2011a. A comparative study of three-dimensional hydraulic conductivity upscaling at the macro-dispersion experiment (made) site, columbus air force base, mississippi (usa). Journal of Hydrology 404, 278–293.

Li, L., Zhou, H., Gómez-Hernández, J.J., 2011b. Transport upscaling using multi-rate mass transfer in three-dimensional highly heterogeneous porous media. Advances in Water Resources 34, 478–489.

Lorenc, A.C., 2003. The potential of the ensemble kalman filter for nwp—a comparison with 4d-var. Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography 129, 3183–3203.

Mahinthakumar, G., Sayeed, M., 2005. Journal of water resources planning and management 131, 45–57.

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., 2002. Application of bayesian inference methods to inverse modelling for contaminants source identification at gloucester landfill, canada. Developments in Water Science 47, 1259–1266.

Michalak, A.M., Kitanidis, P.K., 2003. A method for enforcing parameter nonnegativity in bayesian inverse problems with an application to contaminant source identification. Water Resources Research 39.

Michalak, A.M., Kitanidis, P.K., 2004a. Application of geostatistical inverse modeling to contaminant source identification at dover afb, delaware. Journal of Hydraulic Research 42, 9–18.

Michalak, A.M., Kitanidis, P.K., 2004b. 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.

Mo, S., Zabaras, N., Shi, X., Wu, J., 2019. Deep autoregressive neural networks for high-dimensional inverse problems in groundwater contaminant source identification. Water Resources Research 55, 3856–3881.

Pan, Z., Lu, W., Chang, Z., et al., 2021. Simultaneous identification of groundwater pollution source spatial–temporal characteristics and hydraulic parameters based on deep regularization neural network-hybrid heuristic algorithm. Journal of Hydrology 600, 126586.

Russell, C.S., Shogren, J.F., 2012. Theory, modeling and experience in the management of nonpoint-source pollution. volume 1. Springer Science & Business Media.

Sonnenborg, T.O., Engesgaard, P., Rosbjerg, D., 1996. Contaminant transport at a waste residue deposit: 1. inverse flow and nonreactive transport modeling. Water Resources Research 32, 925–938.

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

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

Wagner, B.J., 1992. Simultaneous parameter estimation and contaminant source characterization for coupled groundwater flow and contaminant transport modelling. Journal of Hydrology 135, 275–303.

Wang, Z., Lu, W., Chang, Z., Wang, H., 2022. Simultaneous identification of groundwater contaminant source and simulation model parameters based on an ensemble kalman filter–adaptive step length ant colony optimization algorithm. Journal of Hydrology 605, 127352.

Wen, X.H., Capilla, J.E., Deutsch, C., Gómez-Hernández, J., Cullick, A., 1999. A program to create permeability fields that honor single-phase flow rate and pressure data. Computers & Geosciences 25, 217–230.

Xu, T., Gómez-Hernández, J.J., 2015. Inverse sequential simulation: Performance and implementation details. Advances in Water Resources 86, 311–326.

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, 6587–6595.

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., Chen, Z., Lu, C., 2021. A comparison between es-mda and restart enkf for the purpose of the simultaneous identification of a contaminant source and hydraulic conductivity. Journal of Hydrology 595, 125681.

Xu, T., Gómez-Hernández, J.J., Li, L., Zhou, H., 2013a. Parallelized ensemble kalman filter for hydraulic conductivity characterization. Computers & Geosciences 52, 42–49.

Xu, T., Gómez-Hernández, J.J., Zhou, H., Li, L., 2013b. 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.

Xu, T., Zhang, W., Gómez-Hernández, J.J., Xie, Y., Yang, J., Chen, Z., Lu, C., 2022. Non-point contaminant source identification in an aquifer using the ensemble smoother with multiple data assimilation. Journal of Hydrology , 127405.

Yeh, H.D., Lin, C.C., Yang, B.J., 2014. Applying hybrid heuristic approach to identify contaminant source information in transient groundwater flow systems. Mathematical Problems in Engineering 2014.

Zhan, C., Dai, Z., Soltanian, M.R., Zhang, X., 2022. Stage-wise stochastic deep learning inversion framework for subsurface sedimentary structure identification. Geophysical research letters 49, e2021GL095823.

Zhang, R., Zhou, N., Xia, X., Zhao, G., Jiang, S., 2020. Joint estimation of hydraulic and biochemical parameters for reactive transport modelling with a modified ilues algorithm. Water 12, 2161.

Zheng, C., 2010. Mt3dms v5. 3 supplemental user's guide. Department of Geological Sciences, University of Alabama, Tuscaloosa, Alabama , 1–56.

Zhou, H., Gómez-Hernández, J.J., Li, L., 2014. Inverse methods in hydrogeology: evolution and recent trends. Advances in Water Resources 63, 22–37.