Non-point contaminant source identification in an aquifer using the ensemble smoother with multiple data assimilations

Teng Xu\textsuperscript{a,b,*}, Wenjun Zhang\textsuperscript{a,b}, J. Jaime Gómez-Hernández\textsuperscript{c}, Yifan Xie\textsuperscript{a,b}, Jie Yang\textsuperscript{a,b}, Zi Chen\textsuperscript{d}, Chunhui Lu\textsuperscript{a,b,*}

\textsuperscript{a}State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing, China
\textsuperscript{b}Yangtze Institute for Conservation and Development, Hohai University, Nanjing, China
\textsuperscript{c}Institute of Water and Environmental Engineering, Universitat Politècnica de València, Valencia, Spain
\textsuperscript{d}China Geological Survey, Nanjing, China

Abstract

Proper identification of groundwater contaminant sources is vital to assess groundwater contamination. However, the majority of previous studies focus on point source identification, only a few works have been conducted for non-point source parameter identification. Here, we employ the ensemble smoother with multiple data assimilation (ES-MDA) to simultaneously identify the spatial architecture of non-point contaminant sources and the related release information. Three different shapes of non-point contaminant sources are considered, an ellipse, a circle, and an irregular shape. We test the applicability of the ES-MDA for the simultaneous identification using three scenarios in a synthetic confined aquifer by assimilating concentration observations from all-time steps multiple times. The results demonstrate that the ES-MDA is capable to accurately identify both regular and irregular non-point contaminant source information; the accuracy of the identification can be improved by increasing the number of iterations.

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

*Corresponding author

Email addresses: teng.xu@hhu.edu.cn (Teng Xu), jgomez@upv.es (J. Jaime Gómez-Hernández), clu@hhu.edu.cn (Chunhui Lu)
1. Introduction

Groundwater is an important source of fresh water for drinking, and also for agricultural, domestic and industrial uses (Barzegar et al., 2017). The quality of groundwater may affect human health. Once it is contaminated, it is important to determine where and when contaminants were introduced into the aquifer. For this purpose, contaminant source identification techniques are used.

Groundwater contaminant sources can be broadly classified into two categories: point and non-point sources. Point sources are normally caused by landfills, gas stations, industry wastewater, and urban sewage, while non-point ones are normally caused by agricultural fertilizers, livestock, poultry farming manure disposal, and leakage from chemical plants.

It is a big challenge to identify a groundwater contaminant source from observations of the contaminants taken downgradient from the source (Ayvaz, 2007). In many cases, the pollution incident is random and accidental, and the discovery of its impact has a lagging nature, which makes difficult to determine the type, properties, source location, intensity and release history of the contaminants. This details are necessary for a proper site remediation design and risk assessment (Aral et al., 2001).

To date, many approaches have been proposed for the identification of groundwater contamination sources. A recent review paper by Gómez-Hernández and Xu (2021) analyzes close of 160 papers published since 1981. Most of the approaches fall in the realm of inverse modeling, which has been used successfully in hydrogeology for other purposes (e.g., Xu et al., 2013; Zhou et al., 2012; Franssen and Gómez-Hernández, 2002; Capilla et al., 1998, 1999; Wen et al., 1999; Li et al., 2012). Inverse modeling for contaminant source identification can be classified into three categories according to their characteristics: optimization, probabilistic, and backward-in-time simulation approaches. Optimization approaches have
been used since early on (Gorelick et al., 1983). They seek minimizing the differences between simulated concentrations and measurement observations (e.g., Gorelick et al., 1983; Sidauruk et al., 1998; Sun et al., 2006b,a; Mirghani et al., 2009); probabilistic approaches seek maximizing the posterior probabilities of the source parameters conditioned on observations (e.g., Woodbury and Ulrych, 1996; Woodbury et al., 1998; Cupola et al., 2015; Zeng et al., 2012; Zhang et al., 2015; Butera et al., 2013; Wang and Jin, 2013); and backward-in-time simulation approaches solve the advection-dispersion equation backwards in time to determine the locations and times with the highest probabilities for the source (e.g., Atmadja and Bagtzoglou, 2001; Bagtzoglou and Atmadja, 2003; Skaggs and Kabala, 1995; Bagtzoglou and Atmadja, 2003; Neupauer et al., 2000). The works published can also be classified in three categories as a function of how the source is treated during the identification process: identification of the release history with known source locations (e.g., Gorelick et al., 1983; Skaggs and Kabala, 1994; Atmadja and Bagtzoglou, 2001; Mahar and Datta, 2000), identification only of the source location(s) (e.g., Dimov et al., 1996; Neupauer and Wilson, 1999), simultaneous identification of both source location and release history (e.g., Aral et al., 2001; Mahinthakumar and Sayeed, 2005; Jamshidi et al., 2020; Xu and Gómez-Hernández, 2016, 2018).

Only a few papers have addressed the problem of identifying a non-point source. Most of them limit themselves to the identification of the corners of a rectangle or a prism (e.g., Mahinthakumar and Sayeed, 2005; Jin et al., 2009). Only the paper by Ayvaz (2016) addresses the problem of identifying an irregular areal source using a genetic algorithm, with the limitation that the final shape must be made up by the juxtaposition of some aquifer discretization cells.

We propose to employ the ensemble smoother with multiple data assimilations (ES-MDA) to solve non-point source identification problems. The work is based on previous works by Xu and Gómez-Hernández (2016, 2018); Chen et al. (2018, 2021); Xu et al. (2021), where
both the restart ensemble Kalman filter (r-EnKF) and the ES-MDA were shown as capable
to identify point contaminant sources in synthetic and laboratory cases. In this work, we
will explore the applicability of the ES-MDA for the identification of the spatial architecture
of irregular non-point contaminant sources and their release history. To the best of our
knowledge, it is the first time that the ES-MDA is used for the identification of non-point
contaminant source information.

This paper is organized as follows, first, we introduce the algorithmic description of the
ES-MDA, second, we test, analyze and discuss the ability of the ES-MDA for the identifica-
tion of regular and irregular non-point contaminant sources in a synthetic aquifer, and we
end with a summary and discussion.

2. Methodology

The ES-MDA developed by Emerick and Reynolds (2013) is an evolution of the en-
semble smoother (ES) porposed by Van Leeuwen and Evensen (1996) to account for the
non-linearities of the state equation. It blends the ES with and iteration technique (multiple
data assimilation). Unlike the EnKF, which updates parameters by assimilating observations
in time, the ES only makes a single update by assimilating all observations from all time
steps at once. Hence, the update is function of the covariances of all forecasted variables
from all time steps and of the misfit of all observations and corresponding forecasts from all
time steps. Since covariances only capture the linear relationship between two variates, the
ES is best suited for linear problems; it fails when the state transfer function is non-linear
(e.g., Crestani et al., 2013; Xu et al., 2021). However, the iterative application of the ES
with multiple data assimilation, as proposed in the ES-MDA, results in multiple progressive
parameter updates yielding excellent results for the non-linear cases.

In this work, the ES-MDA will be used to identify the parameters defining the spatial
architecture of the non-point source, and the corresponding release parameters including
initial release time $T_i$ [T], release duration $\Delta T$ [T], and mass-loading rate $Q$ [MT$^{-1}$] by assimilating observed concentrations $C$ [ML$^{-3}$] from all-time steps at a number of locations. We assume that the shape of the non-point source area can be approximated by an ellipse, hence, the parameters describing the spatial architecture of the non-point source include the $x$ and $y$ coordinates of the center point of the ellipse $X_s$ [L] and $Y_s$ [L], its semi-major and semi-minor axes of source $R_a$ [L] and $R_b$ [L], and the ellipse clockwise rotation angle $B$ [$^\circ$]. We build an augmented model parameter vector $S$ including all the above parameters:

$$S = \begin{bmatrix} X_s \\ Y_s \\ R_a \\ R_b \\ B \\ T_i \\ \Delta T \\ Q \end{bmatrix}.$$ 

(1)

Like the ES, The ES-MDA also consists of two steps: forecast and analysis. First, we set the total number of assimilation iterations to $Na$. Then, in the forecast step, at iteration $j$, concentrations for all-time steps $C_{f,j}^j$ are computed using the last update of the model parameters in vector $S_{j-1}^a$ and a solute transport model $\psi(\cdot)$. The forecast equation is

$$C_{f,j}^j = \psi(C_0, S_{j-1}^a).$$ 

(2)

In the analysis step, at iteration $j$, the augmented model parameter vector $S_j^a$ is updated accounting for the misfit ($C^o + \sqrt{\alpha_j} \varepsilon_j - C_{f,j}^{f,o}$) between forecasted $C_{f,j}^{f,o}$ and observed $C^o$.
concentrations for all-time steps. The update equation is

\[ S_j^a = S_{j-1}^a + K_j (C^o + \sqrt{a_j} \varepsilon_j - C_{f,o}^j), \]  

(3)

with

\[ K_j = D_{SC,j} (D_{CC,j} + a_j R_j)^{-1}, \]  

(4)

where \( \varepsilon_j \) is the observation error with mean zero and covariance \( R_j \), amplified by a non-increasing error variance inflation coefficient \( \sqrt{a_j} \), which should satisfy \( \sum_{i=1}^{N_a} \frac{1}{a_i} = 1 \) (Emerick and Reynolds, 2013)—in this work, we have chosen \( a_j = Na \) for all iterations—; \( K_j \) is the Kalman gain, a function of cross-covariances \( D_{SC,j} \) between parameters and forecasted concentrations at observation locations for all time steps, and auto-covariance \( D_{CC,j} \) between forecasted concentrations at the observation locations obtained for all time steps.

### 3. Application

A synthetic confined aquifer is constructed and discretized into 80 by 80 by 1 cells, each cell being 10 [L] by 10 [L] by 80 [L] (notice that all magnitudes will be unit-free; any set of consistent units with the given values will provide the results shown). A reference log-conductivity field (Figure 1) is generated using the GCOSIM3D code (Gómez-Hernández and Journel, 1993), a sequential multivariate multi-Gaussian simulation code, using the parameters in Table 1.

The boundary conditions shown in Figure 1 are set as follows: north and south boundaries are impermeable; west and east boundaries are prescribed heads with values of 300 [L] and 80 [L], respectively. The initial concentration is zero [ML\(^{-3}\)] throughout the domain.

**Table 1: Parameters used for the random function that models the spatial continuity of lnK**

<table>
<thead>
<tr>
<th>lnK</th>
<th>Mean</th>
<th>Std. dev.</th>
<th>Variogram</th>
<th>( \lambda_{\text{max}} )</th>
<th>( \lambda_{\text{min}} )</th>
<th>Anis. angle</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-2</td>
<td>1</td>
<td>Spherical</td>
<td>300</td>
<td>200</td>
<td>135</td>
</tr>
</tbody>
</table>
The rest of the parameters controlling transport simulation are homogeneous and take the following values: porosity, 0.3 [-], longitudinal dispersively, 3.0 [L], transverse to longitudinal dispersivity, 0.5. There are 30 observation wells and 2 verification wells within the domain (see Figure 2). We have analyzed three different contaminant events with sources of different shapes as shown in Figure 2: an ellipse, a circle and an elongated, wiggly shape. The values of the parameters describing the shapes are given in Table 2. We assume that groundwater flow is at steady-state. The total simulation time is 10950 [T]. The transport model is run for this time in 100 equally-sized time steps (the length of each time step is, therefore, 109.5 [T]). The contaminant enters the aquifer at time 985.5 [T] (around the 10th time step), and ends at time 3285.0 [T] (around at the 30th time step), with a constant mass-loading rate of 1000 [MT\(^{-1}\)]. The release duration is 2299.5 [T] (around 20 time steps). The concentrations have been recorded in the reference fields at observation wells at each time step until the 50th time step (around 5475 [T]) and are used as the observation data for the source identification problem. The numerical transport simulator MT3DMS (e.g., Zheng, 2010; Ma et al., 2012) is used to solve the transport equation. In this work, we only consider advection and dispersion as transport mechanisms. The transport equation is (Zheng, 2010)

Figure 1: Reference fields for lnK (left) and piezometric head (right). The red closed line in the reference lnK field marks the suspect source area.
\[
\frac{\partial (\theta C)}{\partial t} = \nabla \cdot [\theta (D_m + \alpha v) \cdot \nabla C] - \nabla \cdot (\theta v C) - q_s C_s,
\]  
(5)

where \( C \) is the contaminant concentration \([\text{ML}^{-3}]\); \( t \) is the simulation time \([\text{T}]\); \( \nabla \cdot \) is the divergence operator; \( \theta \) is the effective porosity \([-]\); \( D_m \) is the molecular diffusion coefficient \([\text{L}^2\text{T}^{-1}]\); \( \alpha \) is the dispersivity tensor \([\text{L}]\); \( \nabla \) is the gradient operator; \( q_s \) is the volumetric flow rate per unit volume of the aquifer representing fluid sources or sinks \([\text{T}^{-1}]\); \( C_s \) is the concentration of the source or sink flux \([\text{ML}^{-3}]\); \( v \) is the flow velocity vector \([\text{LT}^{-1}]\), related to the hydraulic head \( H \) through \( v = (\frac{-K\nabla H}{\theta}) \), where \( H \) can be calculated by solving the groundwater steady-state equation:

\[
\nabla \cdot (K \nabla H) + W = 0,
\]  
(6)

where \( W \) denotes sources and sinks per unit volume \([\text{LT}^{-1}]\). The numerical groundwater flow simulator MODFLOW (McDonald and Harbaugh, 1988) is used to solve this equation.

The objective of the work is to test the capacity of the ES-MDA in the identification of non-point contaminant sources. For the three scenarios, an ellipse is used as the best shape approximating the true source. In all three scenarios, the initial release time, the release duration and the mass-loading to be identified are the same (see Table 2). Note that to test the need of multiple assimilation of the observation, we show the results after the 1\textsuperscript{st}, 2\textsuperscript{nd}, 4\textsuperscript{th}, and 6\textsuperscript{th} iterations. The location of the three true contaminant source areas is shown in Figure 2, and the corresponding contaminant plumes in the references at the 10\textsuperscript{th}, 30\textsuperscript{th}, and 50\textsuperscript{th} time steps are shown in Figure 3. It is important to note that the differences among the three plumes are not too large, which will make difficult for the ES-MDA to correctly identify each source shape. An ensemble of 500 8-tuplets for the source parameters
is generated, each 8-tuplet stores 8 values drawn independently from the following uniform
distributions with a wide range around the true values: $x$-coordinate of center point of ellipse $X_s \in \mathcal{U}[110, 210]$, $y$-coordinate of center point of ellipse $Y_s \in \mathcal{U}[460, 560]$, semi-major axis of ellipse $Ra \in \mathcal{U}[40, 140]$, semi-minor axis of ellipse $Rb \in \mathcal{U}[10, 80]$, clockwise rotation angle of ellipse major axis $B \in \mathcal{U}[0, 90]$, initial release time $Ti \in \mathcal{U}[0, 3175.5]$, release duration $\Delta T \in \mathcal{U}[1204.5, 6679.5]$, and mass-loading rate $Q \in \mathcal{U}[950, 1200]$. Table 3 summarizes these uniform distributions.

![Figure 2: Well locations and the three sources used in the analysis (yellow areas). The red triangles correspond to observation wells; the black squares mark verification wells.](image-url)
Table 2: Definition of scenarios

<table>
<thead>
<tr>
<th>Scenario</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of assimilation iterations $[Na]$</td>
<td>1,2,4,6</td>
<td>1,2,4,6</td>
<td>1,2,4,6</td>
</tr>
<tr>
<td>Contaminant source shape</td>
<td>Ellipse</td>
<td>Circle</td>
<td>Irregular</td>
</tr>
<tr>
<td>$x$-coordinate of center point of ellipse $[X_s]$</td>
<td>150</td>
<td>150</td>
<td>/</td>
</tr>
<tr>
<td>$y$-coordinate of center point of ellipse $[Y_s]$</td>
<td>540</td>
<td>540</td>
<td>/</td>
</tr>
<tr>
<td>Semi-major axis of ellipse $[Ra]$</td>
<td>80</td>
<td>60</td>
<td>/</td>
</tr>
<tr>
<td>Semi-minor axis of ellipse $[Rb]$</td>
<td>40</td>
<td>60</td>
<td>/</td>
</tr>
<tr>
<td>Clockwise rotation angle $[B]$</td>
<td>30</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Initial release time $[Ti]$</td>
<td>985.5</td>
<td>985.5</td>
<td>985.5</td>
</tr>
<tr>
<td>Release duration $[\Delta T]$</td>
<td>2299.5</td>
<td>2299.5</td>
<td>2299.5</td>
</tr>
<tr>
<td>Mass-loading rate $[Q]$</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>

Table 3: Suspect range of source parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Suspect Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_s$</td>
<td>110 – 210</td>
</tr>
<tr>
<td>$Y_s$</td>
<td>460 – 560</td>
</tr>
<tr>
<td>$Ra$</td>
<td>40 – 140</td>
</tr>
<tr>
<td>$Rb$</td>
<td>10 – 80</td>
</tr>
<tr>
<td>$B$</td>
<td>0 – 90</td>
</tr>
<tr>
<td>$Ti$</td>
<td>0 – 3175.5</td>
</tr>
<tr>
<td>$\Delta T$</td>
<td>1204.5 – 6679.5</td>
</tr>
<tr>
<td>$Q$</td>
<td>950 – 1200</td>
</tr>
</tbody>
</table>
Figure 3: Reference contaminant plumes. Contaminant plumes in the reference \( \ln K \) field induced by the sources shown in Figure 2. Ellipse (top row), circle (middle row), and irregular (bottom row) after time steps 10 (beginning of contaminant injection, left column), 30 (end of contaminant injection, middle column), and 50 (end of assimilation, right column) in the reference aquifer. Red triangles mark the observation wells.
4. Results

Figures 4, 5 and 6 show boxplots for all the 8 contaminant source parameters for the three scenarios, before any updating and after updating at the 1st, 2nd, 4th, and 6th assimilation iteration. The reference parameter values are omitted for B in S2 (circle source) and for Xs, Ys, Ra, Rb and B in S3 (irregular source). We can see the large uncertainties of the initial source parameter values, and how these uncertainties reduce as the number of assimilation iterations increases, with the median of the updated ensembles almost matching the target value after 4 iterations. The only parameter that is not almost exactly reproduced by the ensemble median of the updated parameters is the mass-loading rate M for scenario S3, for which, the final update underestimates the reference value. This underestimation of M for scenario S3 is because the area of the assimilated ellipse is larger than the area of the irregular source in S3 as will be discussed below (recall that an ellipse is used to approximate all sources, including the irregular source area). Consequently, the total mass introduced in the aquifer is well estimated for S3.

Figures 7, 8 and 9 show the average absolute bias ($AAB$) and the ensemble spread ($ESp$) of the ensemble values of the source parameters for all three scenarios. Here, the $AAB$ is used to evaluate the accuracy of the updated source parameters by calculating the average absolute discrepancy between the final updated ensemble values and the true values, while the $ESp$ is to measure the precision of the updated source parameters by calculating the root square of the ensemble variance. The expressions of the $AAB$ and the $ESp$ are

$$AAB = \frac{1}{N_r} \sum_{j=1}^{N_r} |S_j - S_{ref}|, \quad (7)$$

$$ESp = \sqrt{\sigma_S^2}, \quad (8)$$
Figure 4: Scenario S1. Boxplots computed with the initial and updated ensemble of source information parameters including $X_s$, $Y_s$, $R_a$, $R_b$, $B$, $T_i$, $\Delta T$ and $M$ after the 1st, 2nd, 4th, and 6th data assimilation iterations. The dashed horizontal black line corresponds to the reference value.
Figure 5: Scenarios S2. Boxplots computed with the initial and updated ensemble of source information parameters including $X_s$, $Y_s$, $Ra$, $Rb$, $B$, $Ti$, $\Delta T$ and $M$ after the 1$^{st}$, 2$^{nd}$, 4$^{th}$, and 6$^{th}$ data assimilation iterations. The dashed horizontal black line corresponds to the reference value.
Figure 6: Scenarios S3. Boxplots computed with the initial and updated ensemble of source information parameters including $X_s$, $Y_s$, $Ra$, $Rb$, $B$, $Ti$, $\Delta T$ and $M$ after the $1^{st}$, $2^{nd}$, $4^{th}$, and $6^{th}$ data assimilation iterations. The dashed horizontal black line corresponds to the reference value.
where $N_r$ is the number of realizations, $S_{ref}$ is the reference source parameter value, $S_j$ is the source parameter value for the $j^{th}$ realization in the ensemble, and $\sigma_S^2$ is the ensemble variance of the source parameters. The AAB has not been computed for those parameters for which there is reference value is undefined. An analysis of the figures shows how both parameters decrease as the number of assimilation iterations increase reaching a value close to zero at iteration 6, indicating that the ES-MDA has retrieved successfully the source parameters with great accuracy and precision. Notice also that the ratio $ESp/AAB$ is close to 1 for almost all parameters and all iterations, an indication that the filter is performing well without any filter inbreeding.

Figures 10 and 11 display a statistic about the shape of the source. It measures the probability that the source is at a given location. This probability is approximated, cell by cell, by the fraction of realizations in which the source is present

$$P_i = \frac{1}{N_r} \sum_{j=1}^{N_r} I_{j,i}. \quad (9)$$

where $P_i$ is the probability that the source is present at a cell $i$ and $I_{j,i}$ is an indicator function valued 1 if the source is present at cell $i$ for realization $j$, 0, if not.

Figure 10 only displays one map since the ensemble of initial ellipses is the same for all three scenarios; whereas Figure 11 shows the evolution for each scenario as observations are assimilated. In the three scenarios, we can notice how the initial ensemble and the first iterations display considerable uncertainty about the location of the source. Uncertainty that disappears at iteration six, where the area of probability 1 identifies almost perfectly the source. Only S3 with the irregular-shape source shows some uncertainty at the edges of the ellipse.

Besides analyzing how well the source parameters are identified by the ES-MDA, it is important to analyze how well transport is reproduced with the updated parameters. Given
Figure 7: Scenario S1. Average absolute bias (AAB) and ensemble spread (ESp) computed with the initial and updated ensembles of source parameters Xs, Ys, Ra, Rb, B, Ti, ΔT and M after the 1st, 2nd, 4th, and 6th data assimilation iterations.
Figure 8: Scenario S2. AAB and ESp computed with the initial and updated ensembles of source parameters $X_s$, $Y_s$, $Ra$, $Rb$, $B$, $Ti$, $\Delta T$ and $M$ after the 1$^{st}$, 2$^{nd}$, 4$^{th}$, and 6$^{th}$ data assimilation iterations.
Figure 9: Scenario S3. AAB and ESp computed with the initial and updated ensembles of source parameters $Xs, Ys, Ra, Rb, B, Ti, \Delta T$ and $M$ after the 1st, 2nd, 4th, and 6th data assimilation iterations.
the very large accuracy and precision of the final estimates, it can be anticipated that
this reproduction will be very good. Figure 12 shows the time evolution of contaminant
concentrations at the two verification wells (#1 and #2) for all scenarios, computed with
the initial ensemble (same for all three scenarios). Figures 13 and 14 show the time evolution
of contaminant concentrations at the two verification wells (#1 and #2) for all scenarios,
computed with the ensembles of updated contaminant source parameters for each scenario
after the 1st, 2nd, 4th and 6th assimilation iterations, respectively. Similarly to what happens
with the identification of the source area, uncertainties about predicted concentrations are
very large with the initial ensemble of parameters and during the first iterations, but this
uncertainty reduces considerably after six iterations; up to the point, that the 90% confidence
interval almost collapses onto of the reference concentrations.

Figures 15, 16 and 17 show the contaminant plume in realization #300 (top row), the
ensemble mean (middle row) and the ensemble variance (bottom row) of all plumes at the
10th, 30th and 30th simulation time steps for scenarios S1, S2 and S3, respectively. The plumes
are computed using the updated source parameters after the 6th assimilation iteration. As
expected, when compared with the reference contaminant plumes in Figure 3, the shapes
and spatial distribution of solute concentrations are well reproduced.
Figure 11: Scenarios S1-S3. Probability of source location as computed from the ellipses given by the parameters updated at the 1st, 2nd, 4th and 6th assimilation iterations.
Figure 12: Scenarios S1-S3. Time evolution of contaminant concentrations at the two verification wells #1 and #2 computed with the initial ensemble of source information parameters (same for all three scenarios). The red line corresponds to the evolution of the concentration in the reference; the black lines correspond to the ensemble 5 and 95 percentiles of all realizations; the green line corresponds to the ensemble median; the vertical dashed lines mark the end of the assimilation period.
Figure 13: Scenarios S1-S3. Time evolution of the contaminant concentrations at the verification well #1 computed with the updated ensembles of source parameters after the 1st, 2nd, 4th and 6th assimilation iterations. The red line corresponds to the evolution of the concentration in the reference; the black lines correspond to the ensemble 5 and 95 percentiles of all realizations; the green line corresponds to the ensemble median; the vertical dashed lines mark the end of the assimilation period.
Figure 14: Scenarios S1-S3. Time evolution of the contaminant concentrations at the verification well #2 computed with the updated ensembles of source parameters after the 1st, 2nd, 4th and 6th assimilation iterations. The red line corresponds to the evolution of the concentration in the reference; the black lines correspond to the ensemble 5 and 95 percentiles of all realizations; the green line corresponds to the ensemble median; the vertical dashed lines mark the end of the assimilation period.
Figure 15: Scenario S1. Contaminant plume at the 10\textsuperscript{th}, 30\textsuperscript{th} and 30\textsuperscript{th} simulation time steps, computed with the updated parameters after the 6\textsuperscript{th} assimilation iteration. From top to bottom, plume in realization #300; ensemble mean of all plumes, and ensemble variance of all plumes.
Figure 16: Scenario S2. Contaminant plume at the 10th, 30th and 30th simulation time steps, computed with the updated parameters after the 6th assimilation iteration. From top to bottom, plume in realization #300; ensemble mean of all plumes, and ensemble variance of all plumes.
Figure 17: Scenario S3. Contaminant plume at the 10th, 30th and 30th simulation time steps, computed with the updated parameters after the 6th assimilation iteration. From top to bottom, plume in realization #300; ensemble mean of all plumes, and ensemble variance of all plumes.
5. Summary and discussion

The main objective of this work is to analyze the capacity of the ES-MDA for the identification of non-point contaminant sources. We have demonstrated that the ES-MDA is capable to identify the shape of the source area (approximated as an ellipse, which is defined with five geometrical parameters), the initial release time, the release duration, and the mass-loading rate, in three scenarios using an elliptical, circular and irregular shape source. We have shown that the ellipse can not only characterize the regular source area (ellipse and circle) but also successfully approximate the irregular source area; however, when we use the ellipse to approximate the irregular source area, the final estimate may give an ellipse covering some extra nodes and, in order to conserve mass, the mass-load rate be underestimated.

Besides, we also demonstrate that increasing the number of data assimilation iterations is very helpful to improve the performance of the ES-MDA for the purpose of identifying the source, but at the cost of higher computation.

Although we have successfully demonstrated the ability of the ES-MDA for the non-point contaminant source identification, there is still a long way until it could be applied in practice. The next step is to couple the identification of the source with that of the underlying heterogeneity of hydraulic conductivities, and then to devise a technique that can be applied to the identification of truly irregular shapes.

Acknowledgements Financial support to carry out this work was received from the financial support from the Fundamental Research Funds for the Central Universities (B200201015) and Jiangsu Specially-Appointed Professor Program from Jiangsu Provincial Department of Education (B19052). J.J. Gómez-Hernández acknowledges the financial support from the Spanish Ministry of Science and Innovation through project number 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). Y. Xie acknowledges the Fundamental Research Funds for the Central Universities (B210202018). J. Yang acknowledges the National Natural Science Foundation of China (52009032) and the Fundamental Research Funds for the Central Universities (B210202019)

References


