

Document downloaded from:

<http://hdl.handle.net/10251/193630>

This paper must be cited as:

Xu, T.; Zhang, W.; Gómez-Hernández, JJ.; 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*. 606:1-17.
<https://doi.org/10.1016/j.jhydrol.2021.127405>



The final publication is available at

<https://doi.org/10.1016/j.jhydrol.2021.127405>

Copyright Elsevier

Additional Information

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

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

^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

25 been used since early on (Gorelick et al., 1983). They seek minimizing the differences be-
26 tween simulated concentrations and measurement observations (e.g., Gorelick et al., 1983;
27 Sidauruk et al., 1998; Sun et al., 2006b,a; Mirghani et al., 2009); probabilistic approaches
28 seek maximizing the posterior probabilities of the source parameters conditioned on obser-
29 vations (e.g., Woodbury and Ulrych, 1996; Woodbury et al., 1998; Cupola et al., 2015; Zeng
30 et al., 2012; Zhang et al., 2015; Butera et al., 2013; Wang and Jin, 2013); and backward-
31 in-time simulation approaches solve the advection-dispersion equation backwards in time to
32 determine the locations and times with the highest probabilities for the source (e.g., Atmadja
33 and Bagtzoglou, 2001; Bagtzoglou and Atmadja, 2003; Skaggs and Kabala, 1995; Bagtzoglou
34 and Atmadja, 2003; Neupauer et al., 2000). The works published can also be classified in
35 three categories as a function of how the source is treated during the identification process:
36 identification of the release history with known source locations (e.g., Gorelick et al., 1983;
37 Skaggs and Kabala, 1994; Atmadja and Bagtzoglou, 2001; Mahar and Datta, 2000), identifi-
38 cation only of the source location(s) (e.g., Dimov et al., 1996; Neupauer and Wilson, 1999),
39 simultaneous identification of both source location and release history (e.g., Aral et al., 2001;
40 Mahinthakumar and Sayeed, 2005; Jamshidi et al., 2020; Xu and Gómez-Hernández, 2016,
41 2018).

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

48 We propose to employ the ensemble smoother with multiple data assimilations (ES-MDA)
49 to solve non-point source identification problems. The work is based on previous works by
50 Xu and Gómez-Hernández (2016, 2018); Chen et al. (2018, 2021); Xu et al. (2021), where

51 both the restart ensemble Kalman filter (r-EnKF) and the ES-MDA were shown as capable
52 to identify point contaminant sources in synthetic and laboratory cases. In this work, we
53 will explore the applicability of the ES-MDA for the identification of the spatial architecture
54 of irregular non-point contaminant sources and their release history. To the best of our
55 knowledge, it is the first time that the ES-MDA is used for the identification of non-point
56 contaminant source information.

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

61 **2. Methodology**

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

74 In this work, the ES-MDA will be used to identify the parameters defining the spatial
75 architecture of the non-point source, and the corresponding release parameters including

76 initial release time Ti [T], release duration ΔT [T], and mass-loading rate Q [MT^{-1}] by
77 assimilating observed concentrations C [ML^{-3}] from all-time steps at a number of locations.
78 We assume that the shape of the non-point source area can be approximated by an ellipse,
79 hence, the parameters describing the spatial architecture of the non-point source include the
80 x and y coordinates of the center point of the ellipse Xs [L] and Ys [L], its semi-major and
81 semi-minor axes of source Ra [L] and Rb [L], and the ellipse clockwise rotation angle B [$^\circ$].
82 We build an augmented model parameter vector S including all the above parameters:

$$S = \begin{bmatrix} Xs \\ Ys \\ Ra \\ Rb \\ B \\ Ti \\ \Delta T \\ Q \end{bmatrix}. \quad (1)$$

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

$$C_j^f = \psi(C_0, S_{j-1}^a). \quad (2)$$

87 In the analysis step, at iteration j , the augmented model parameter vector S_j^a is updated
88 accounting for the misfit $(C^o + \sqrt{a_j}\varepsilon_j - C_j^{f,o})$ between forecasted $C_j^{f,o}$ and observed C^o

89 concentrations for all-time steps. The update equation is

$$S_j^a = S_{j-1}^a + K_j(C^o + \sqrt{a_j}\varepsilon_j - C_j^{f,o}), \quad (3)$$

90 with

$$K_j = D_{SC,j}(D_{CC,j} + a_j R_j)^{-1}, \quad (4)$$

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

97 3. Application

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

104 The boundary conditions shown in Figure 1 are set as follows: north and south boundaries
 105 are impermeable; west and east boundaries are prescribed heads with values of 300 [L]
 106 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 $\ln K$

	Mean	Std. dev.	Variogram	λ_{\max}	λ_{\min}	Anis. angle
$\ln K$	-2	1	Spherical	300	200	135

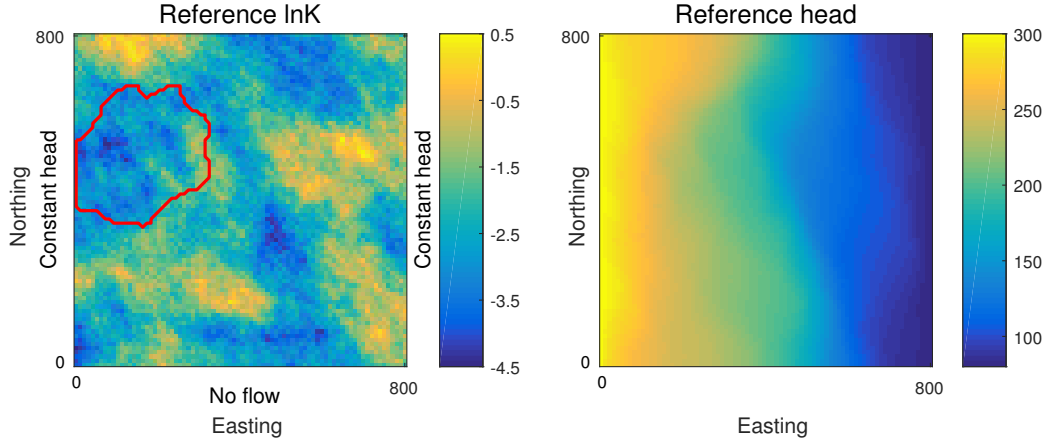


Figure 1: Reference fields for $\ln K$ (left) and piezometric head (right). The red closed line in the reference $\ln K$ field marks the suspect source area.

107 The rest of the parameters controlling transport simulation are homogeneous and take the
 108 following values: porosity, 0.3 [-], longitudinal dispersivity, 3.0 [L], transverse to longitudinal
 109 dispersivity, 0.5. There are 30 observation wells and 2 verification wells within the domain
 110 (see Figure 2). We have analyzed three different contaminant events with sources of different
 111 shapes as shown in Figure 2: an ellipse, a circle and an elongated, wiggly shape. The values
 112 of the parameters describing the shapes are given in Table 2. We assume that groundwater
 113 flow is at steady-state. The total simulation time is 10950 [T]. The transport model is run
 114 for this time in 100 equally-sized time steps (the length of each time step is, therefore, 109.5
 115 [T]). The contaminant enters the aquifer at time 985.5 [T] (around the 10th time step), and
 116 ends at time 3285.0 [T] (around at the 30th time step), with a constant mass-loading rate of
 117 1000 [MT⁻¹]. The release duration is 2299.5 [T] (around 20 time steps). The concentrations
 118 have been recorded in the reference fields at observation wells at each time step until the 50th
 119 time step (around 5475 [T]) and are used as the observation data for the source identification
 120 problem. The numerical transport simulator MT3DMS (e.g., Zheng, 2010; Ma et al., 2012) is
 121 used to solve the transport equation. In this work, we only consider advection and dispersion
 122 as transport mechanisms. The transport equation is (Zheng, 2010)

$$\frac{\partial(\theta C)}{\partial t} = \nabla \cdot [\theta(D_m + \alpha v) \cdot \nabla C] - \nabla \cdot (\theta v C) - q_s C_s, \quad (5)$$

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

$$\nabla \cdot (K\nabla H) + W = 0, \quad (6)$$

130 where W denotes sources and sinks per unit volume [LT^{-1}]. The numerical groundwater
 131 flow simulator MODFLOW (McDonald and Harbaugh, 1988) is used to solve this equation.
 132 The steady-state head field for the reference field is shown in Figure 1; flow is mainly from
 133 west to east, and so is the spreading of the contaminant plume.

134 The objective of the work is to test the capacity of the ES-MDA in the identification
 135 of non-point contaminant sources. For the three scenarios, an ellipse is used as the best
 136 shape approximating the true source. In all three scenarios, the initial release time, the
 137 release duration and the mass-loading to be identified are the same (see Table 2). Note that
 138 to test the need of multiple assimilation of the observation, we show the results after the
 139 1st, 2nd, 4th, and 6th iterations. The location of the three true contaminant source areas is
 140 shown in Figure 2, and the corresponding contaminant plumes in the references at the 10th,
 141 30th, and 50th time steps are shown in Figure 3. It is important to note that the differences
 142 among the three plumes are not too large, which will make difficult for the ES-MDA to
 143 correctly identify each source shape. An ensemble of 500 8-tuplets for the source parameters

144 is generated, each 8-tuplet stores 8 values drawn independently from the following uniform
 145 distributions with a wide range around the true values: x -coordinate of center point of ellipse
 146 $Xs \in \mathcal{U}[110, 210]$, y -coordinate of center point of ellipse $Ys \in \mathcal{U}[460, 560]$, semi-major axis
 147 of ellipse $Ra \in \mathcal{U}[40, 140]$, semi-minor axis of ellipse $Rb \in \mathcal{U}[10, 80]$, clockwise rotation angle
 148 of ellipse major axis $B \in \mathcal{U}[0, 90]$, initial release time $Ti \in \mathcal{U}[0, 3175.5]$, release duration
 149 $\Delta T \in \mathcal{U}[1204.5, 6679.5]$, and mass-loading rate $Q \in \mathcal{U}[950, 1200]$. Table 3 summarizes these
 150 uniform distributions.

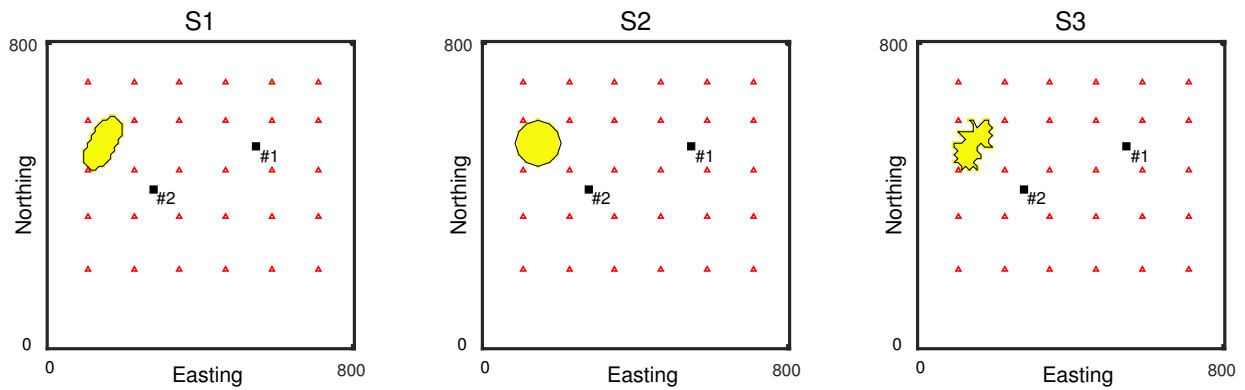


Figure 2: Well locations and the three sources used in the analysis (yellow areas). The red triangles correspond to observation wells; the black squares mark verification wells.

Table 2: Definition of scenarios

Scenario	S1	S2	S3
Number of assimilation iterations [Na]	1,2,4,6	1,2,4,6	1,2,4,6
Contaminant source shape	Ellipse	Circle	Irregular
x -coordinate of center point of ellipse [Xs]	150	150	/
y -coordinate of center point of ellipse [Ys]	540	540	/
Semi-major axis of ellipse [Ra]	80	60	/
Semi-minor axis of an ellipse [Rb]	40	60	/
Clockwise rotation angle [B]	30	/	/
Initial release time [Ti]	985.5	985.5	985.5
Release duration [ΔT]	2299.5	2299.5	2299.5
Mass-loading rate [Q]	1000	1000	1000

Table 3: Suspect range of source parameters

Parameters	Suspect Range
Xs	110 – 210
Ys	460 – 560
Ra	40 – 140
Rb	10 – 80
B	0 – 90
Ti	0 – 3175.5
ΔT	1204.5 – 6679.5
Q	950 – 1200

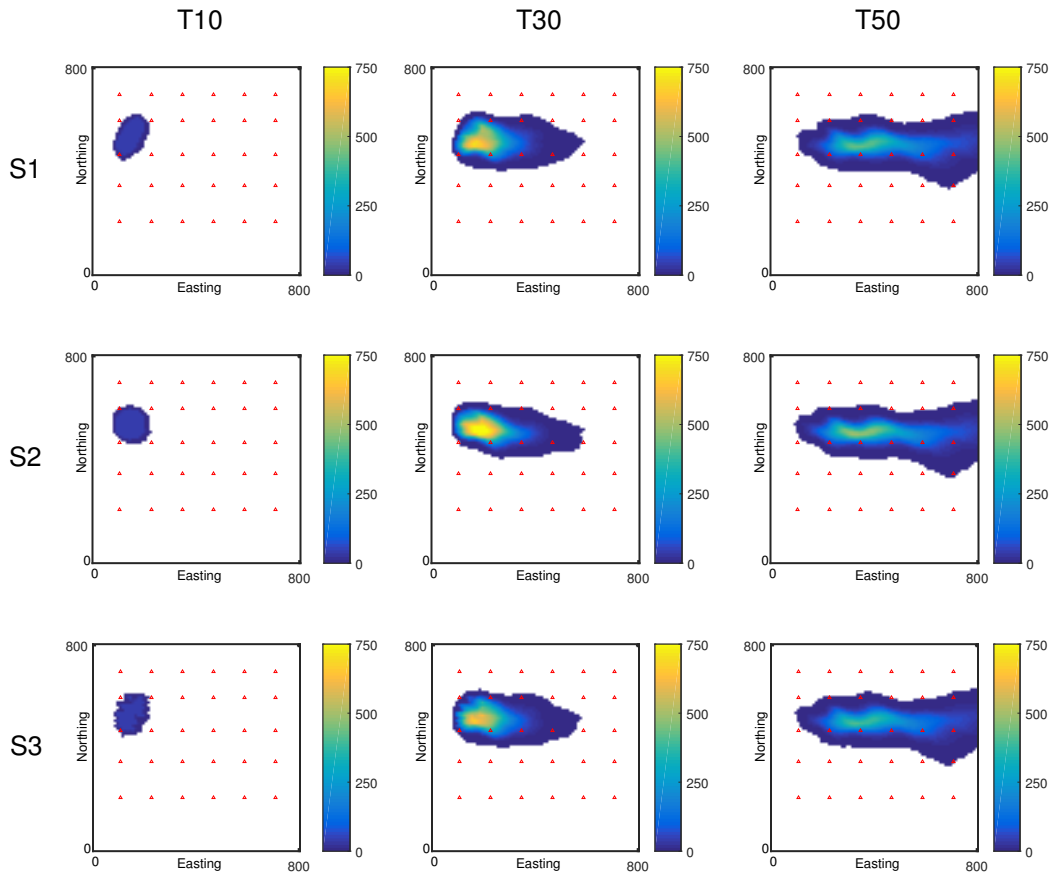


Figure 3: Reference contaminant plumes. Contaminant plumes in the reference $\ln K$ field induced by the sources shown in Figure 2. Ellipse (top row), circle (middle row), and irregular (bottom row) after time steps 10 (beginning of contaminant injection, left column), 30 (end of contaminant injection, middle column), and 50 (end of assimilation, right column) in the reference aquifer. Red triangles mark the observation wells.

151 **4. Results**

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

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

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

$$ESp = \sqrt{\sigma_S^2}, \quad (8)$$

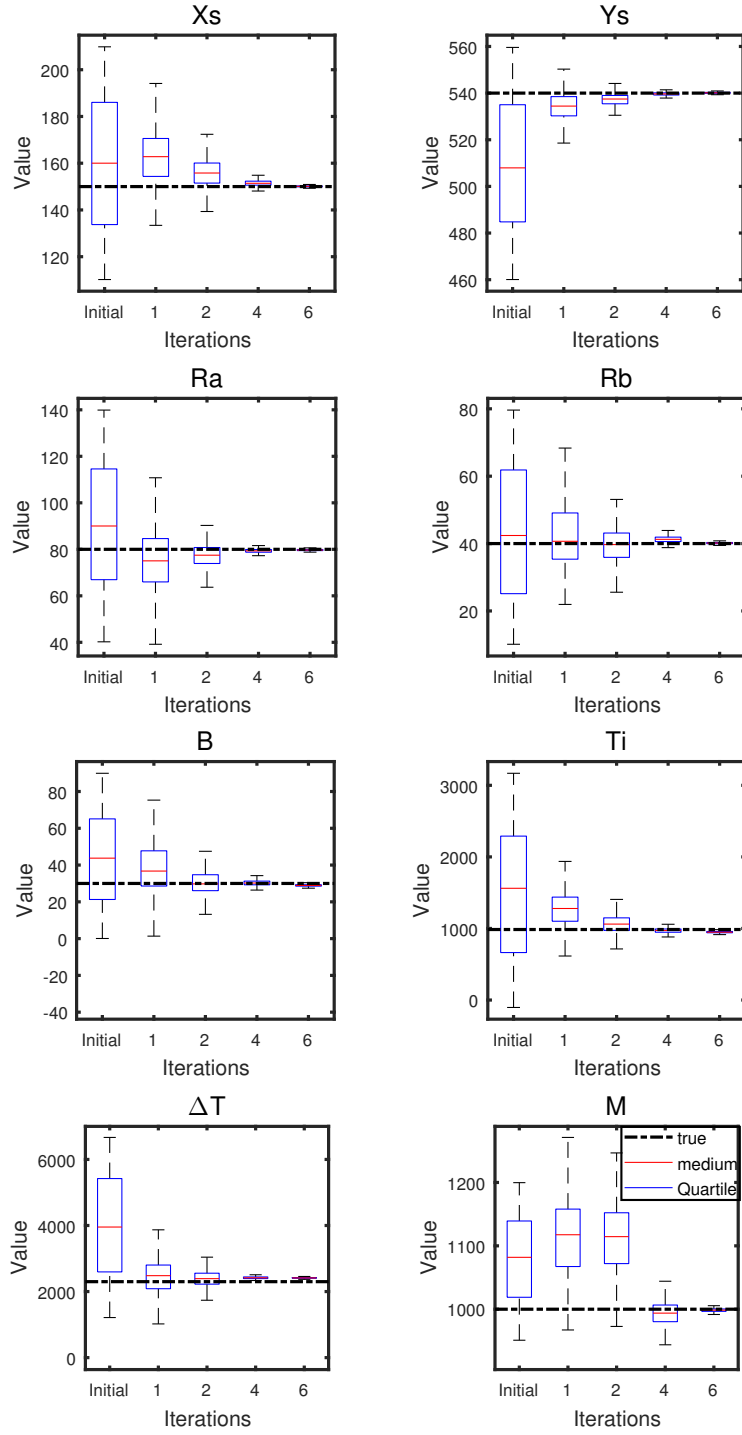


Figure 4: Scenario S1. Boxplots computed with the initial and updated ensemble of source information parameters including X_s , Y_s , R_a , R_b , B , T_i , ΔT and M after the 1st, 2nd, 4th, and 6th data assimilation iterations. The dashed horizontal black line corresponds to the reference value.

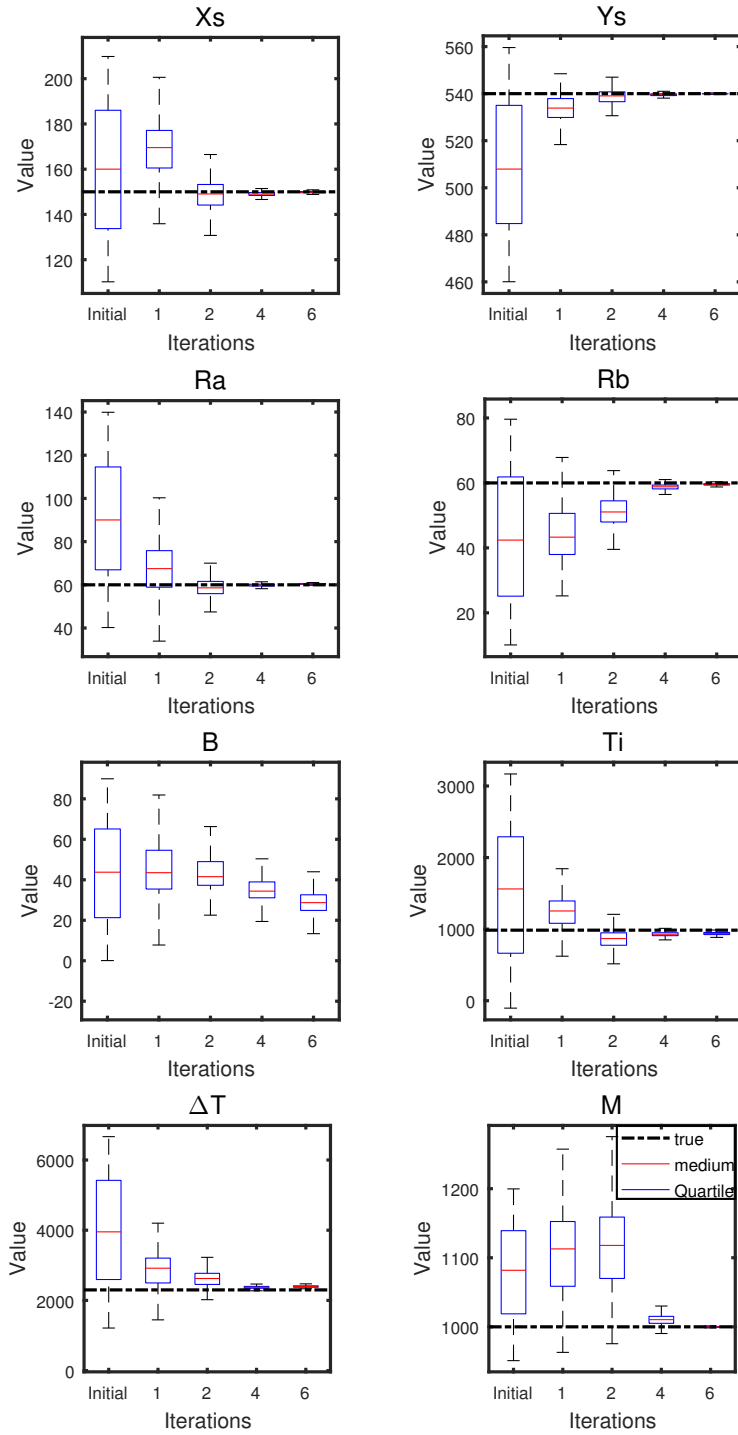


Figure 5: Scenarios S2. Boxplots computed with the initial and updated ensemble of source information parameters including X_s , Y_s , R_a , R_b , B , T_i , ΔT and M after the 1st, 2nd, 4th, and 6th data assimilation iterations. The dashed horizontal black line corresponds to the reference value.

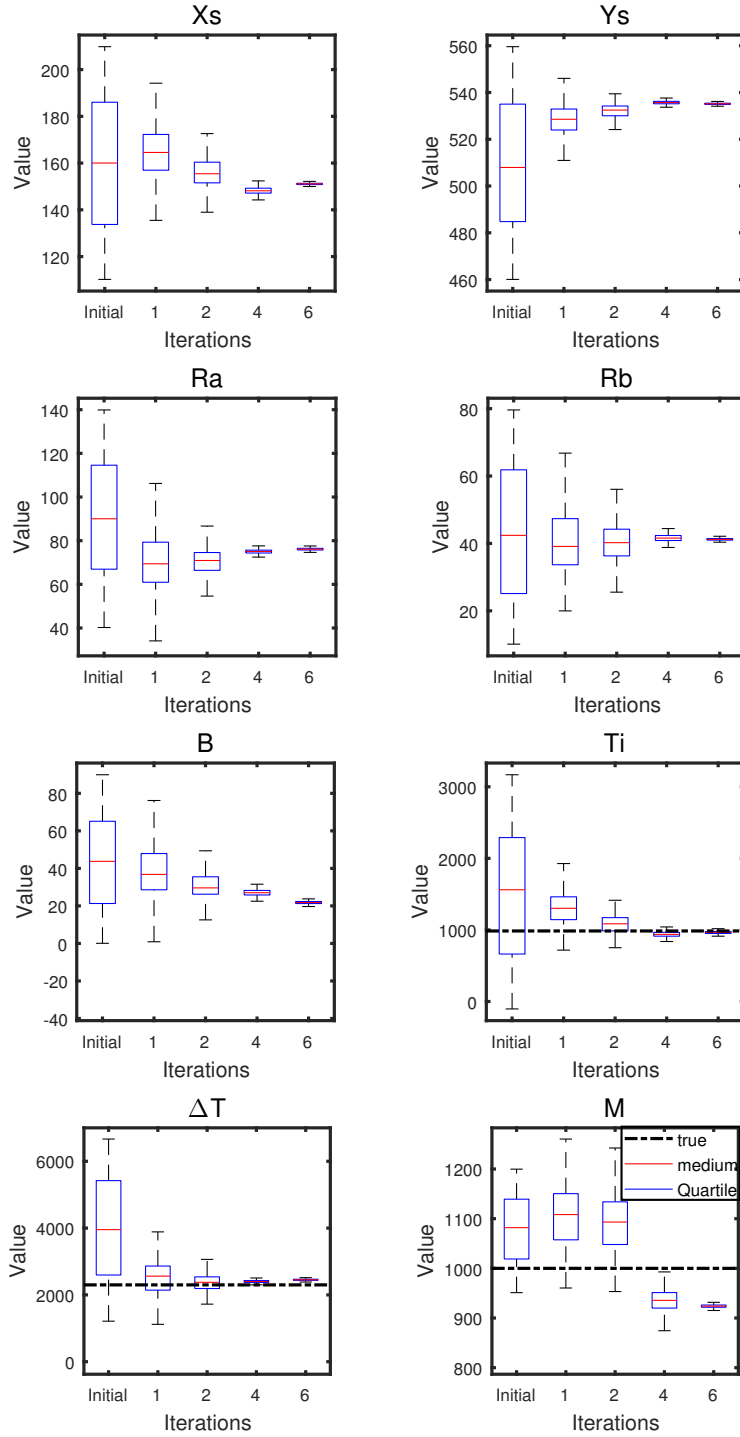


Figure 6: Scenarios S3. Boxplots computed with the initial and updated ensemble of source information parameters including X_s , Y_s , R_a , R_b , B , T_i , ΔT and M after the 1st, 2nd, 4th, and 6th data assimilation iterations. The dashed horizontal black line corresponds to the reference value.

171 where N_r is the number of realizations, S_{ref} is the reference source parameter value, S_j is
 172 the source parameter value for the j^{th} realization in the ensemble, and σ_S^2 is the ensemble
 173 variance of the source parameters. The AAB has not been computed for those parameters
 174 for which there is reference value is undefined. An analysis of the figures shows how both
 175 parameters decrease as the number of assimilation iterations increase reaching a value close
 176 to zero at iteration 6, indicating that the ES-MDA has retrieved successfully the source
 177 parameters with great accuracy and precision. Notice also that the ratio ESp/AAB is close
 178 to 1 for almost all parameters and all iterations, an indication that the filter is performing
 179 well without any filter inbreeding.

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

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

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

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

192 Besides analyzing how well the source parameters are identified by the ES-MDA, it is
 193 important to analyze how well transport is reproduced with the updated parameters. Given

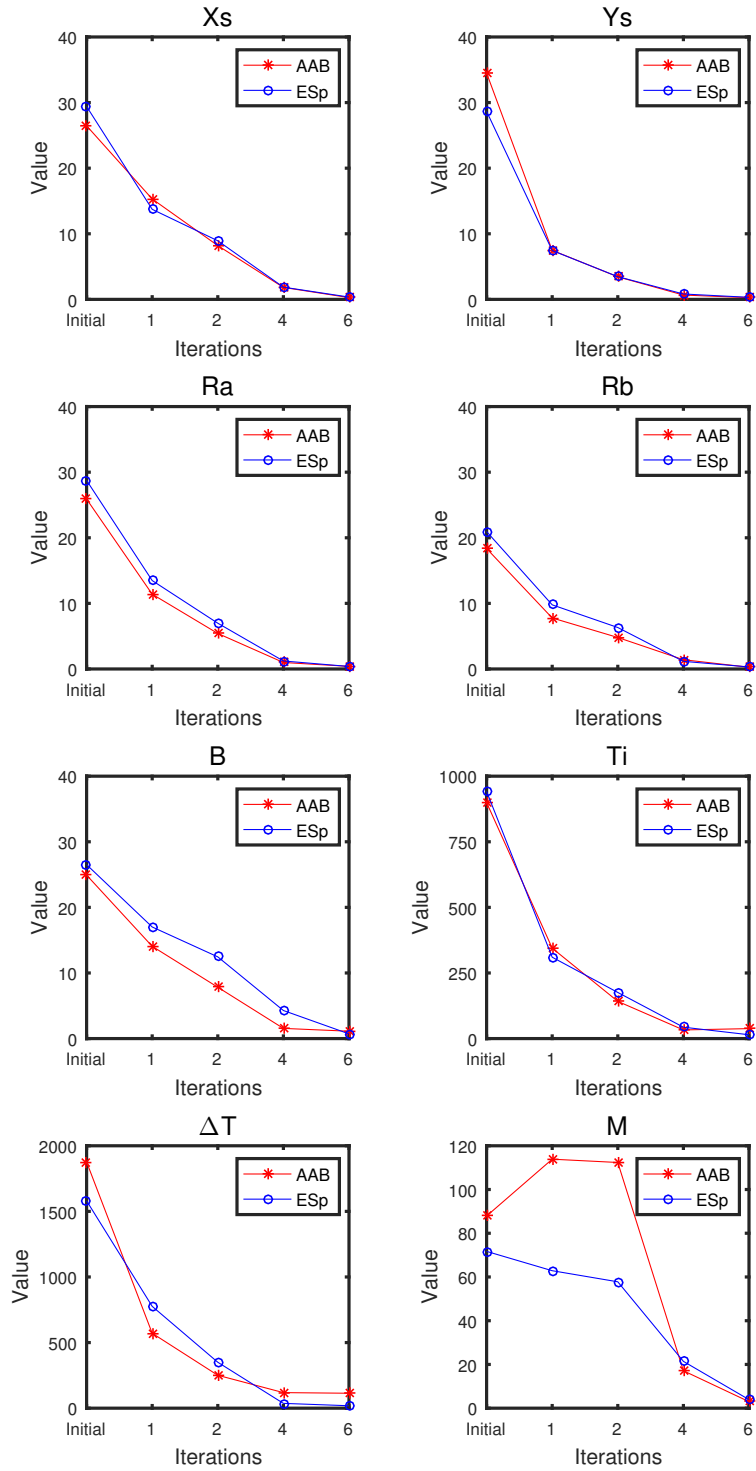


Figure 7: Scenario S1. Average absolute bias (AAB) and ensemble spread (ESp) computed with the initial and updated ensembles of source parameters X_s , Y_s , R_a , R_b , B , T_i , ΔT and M after the 1st, 2nd, 4th, and 6th data assimilation iterations.

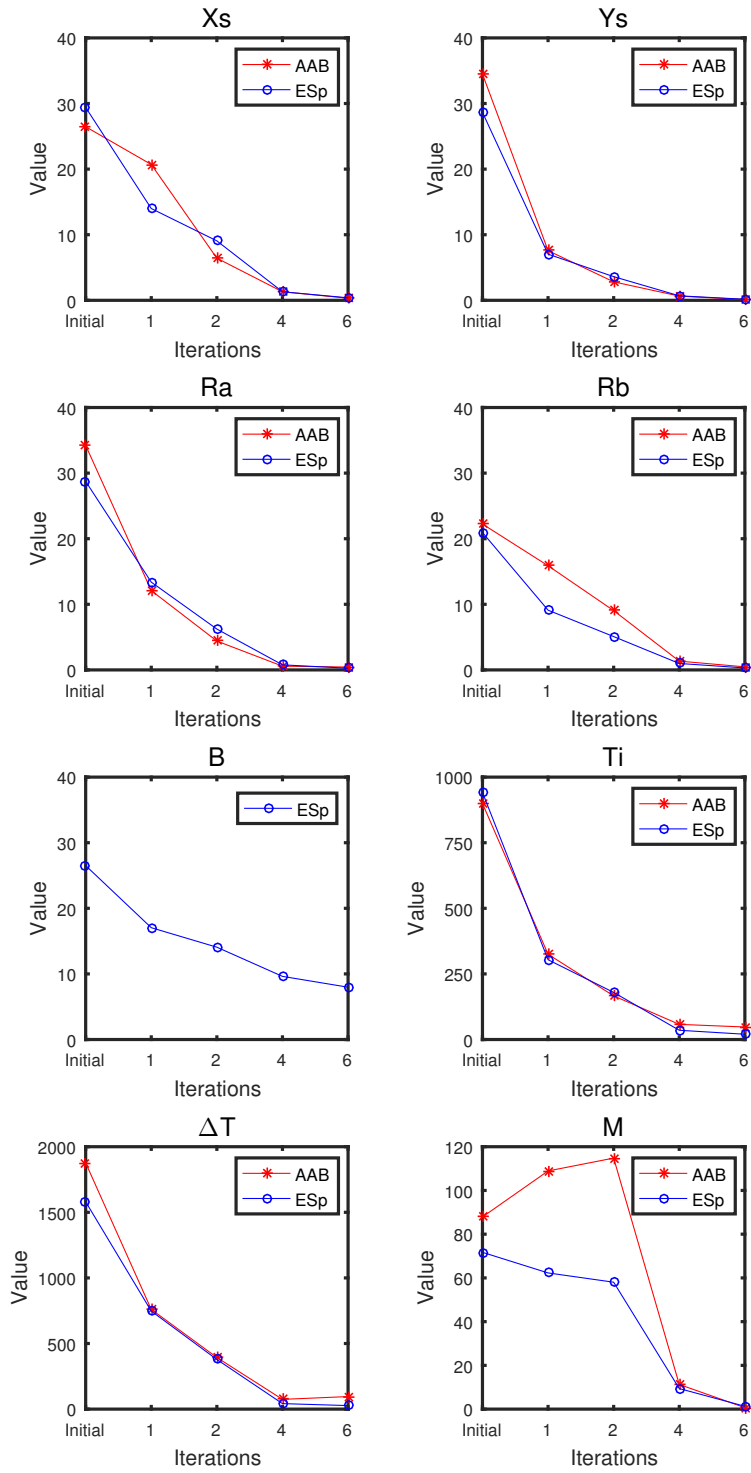


Figure 8: Scenario S2. *AAB* and *ESp* computed with the initial and updated ensembles of source parameters X_s , Y_s , R_a , R_b , B , T_i , ΔT and M after the 1st, 2nd, 4th, and 6th data assimilation iterations.

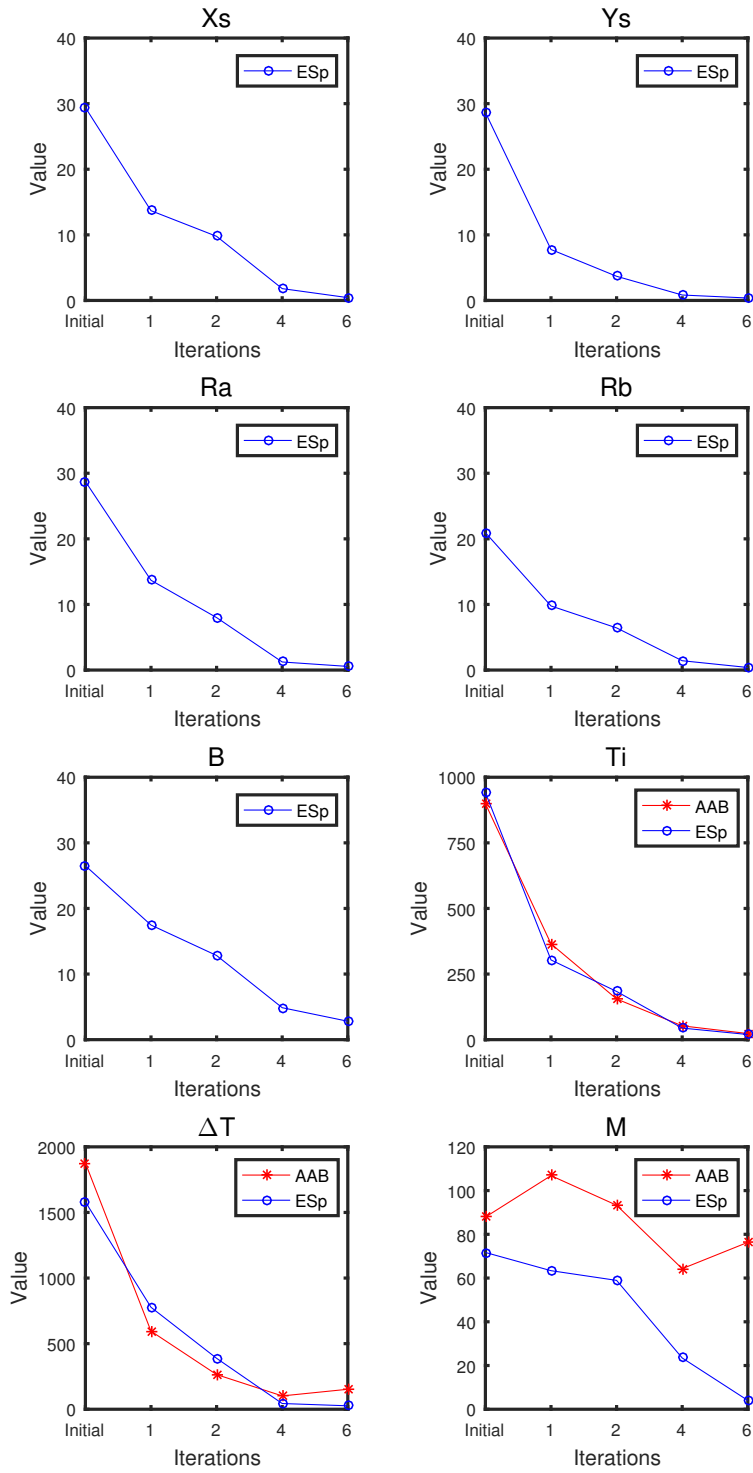


Figure 9: Scenario S3. *AAB* and *ESp* computed with the initial and updated ensembles of source parameters X_s , Y_s , R_a , R_b , B , T_i , ΔT and M after the 1st, 2nd, 4th, and 6th data assimilation iterations.

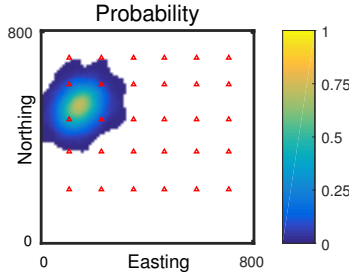


Figure 10: Scenarios S1-S3. Probability of source location as computed from the ellipses given by the initial ensemble of parameters.

194 the very large accuracy and precision of the final estimates, it can be anticipated that
 195 this reproduction will be very good. Figure 12 shows the time evolution of contaminant
 196 concentrations at the two verification wells (#1 and #2) for all scenarios, computed with
 197 the initial ensemble (same for all three scenarios). Figures 13 and 14 show the time evolution
 198 of contaminant concentrations at the two verification wells (#1 and #2) for all scenarios,
 199 computed with the ensembles of updated contaminant source parameters for each scenario
 200 after the 1st, 2st, 4st and 6th assimilation iterations, respectively. Similarly to what happens
 201 with the identification of the source area, uncertainties about predicted concentrations are
 202 very large with the initial ensemble of parameters and during the first iterations, but this
 203 uncertainty reduces considerably after six iterations; up to the point, that the 90% confidence
 204 interval almost collapses onto of the reference concentrations.

205 Figures 15, 16 and 17 show the contaminant plume in realization #300 (top row), the
 206 ensemble mean (middle row) and the ensemble variance (bottom row) of all plumes at the
 207 10th, 30th and 30th simulation time steps for scenarios S1, S2 and S3, respectively. The plumes
 208 are computed using the updated source parameters after the 6th assimilation iteration. As
 209 expected, when compared with the reference contaminant plumes in Figure 3, the shapes
 210 and spatial distribution of solute concentrations are well reproduced.

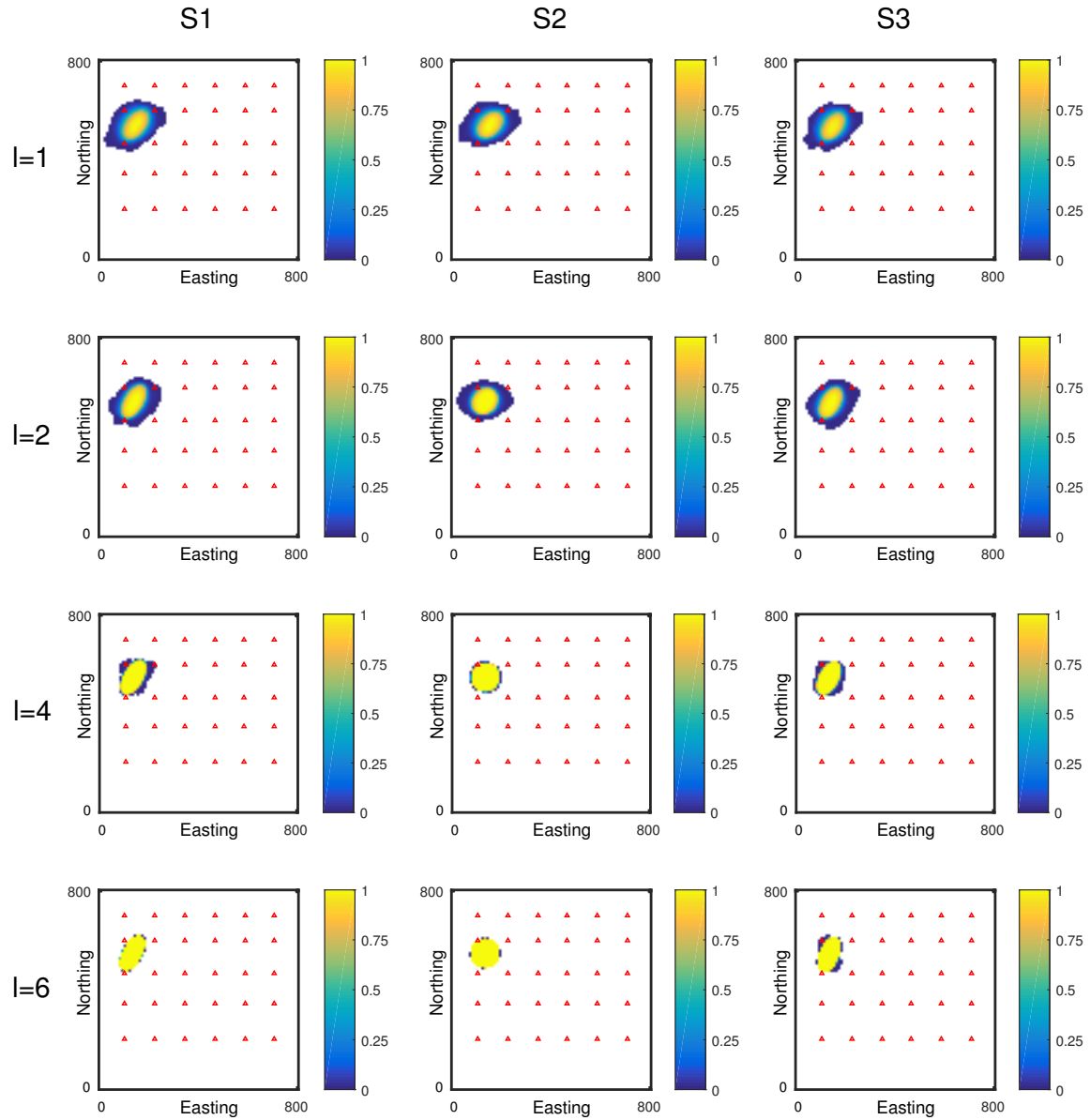


Figure 11: Scenarios S1-S3. Probability of source location as computed from the ellipses given by the parameters updated at the 1st, 2st, 4th and 6th assimilation iterations.

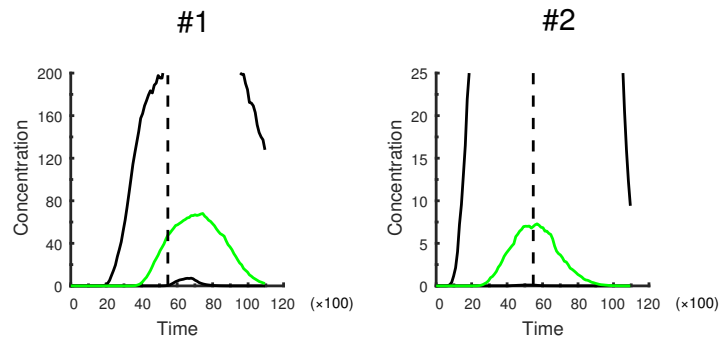


Figure 12: Scenarios S1-S3. Time evolution of contaminant concentrations at the two verification wells #1 and #2 computed with the initial ensemble of source information parameters (same for all three scenarios). The red line corresponds to the evolution of the concentration in the reference; the black 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.

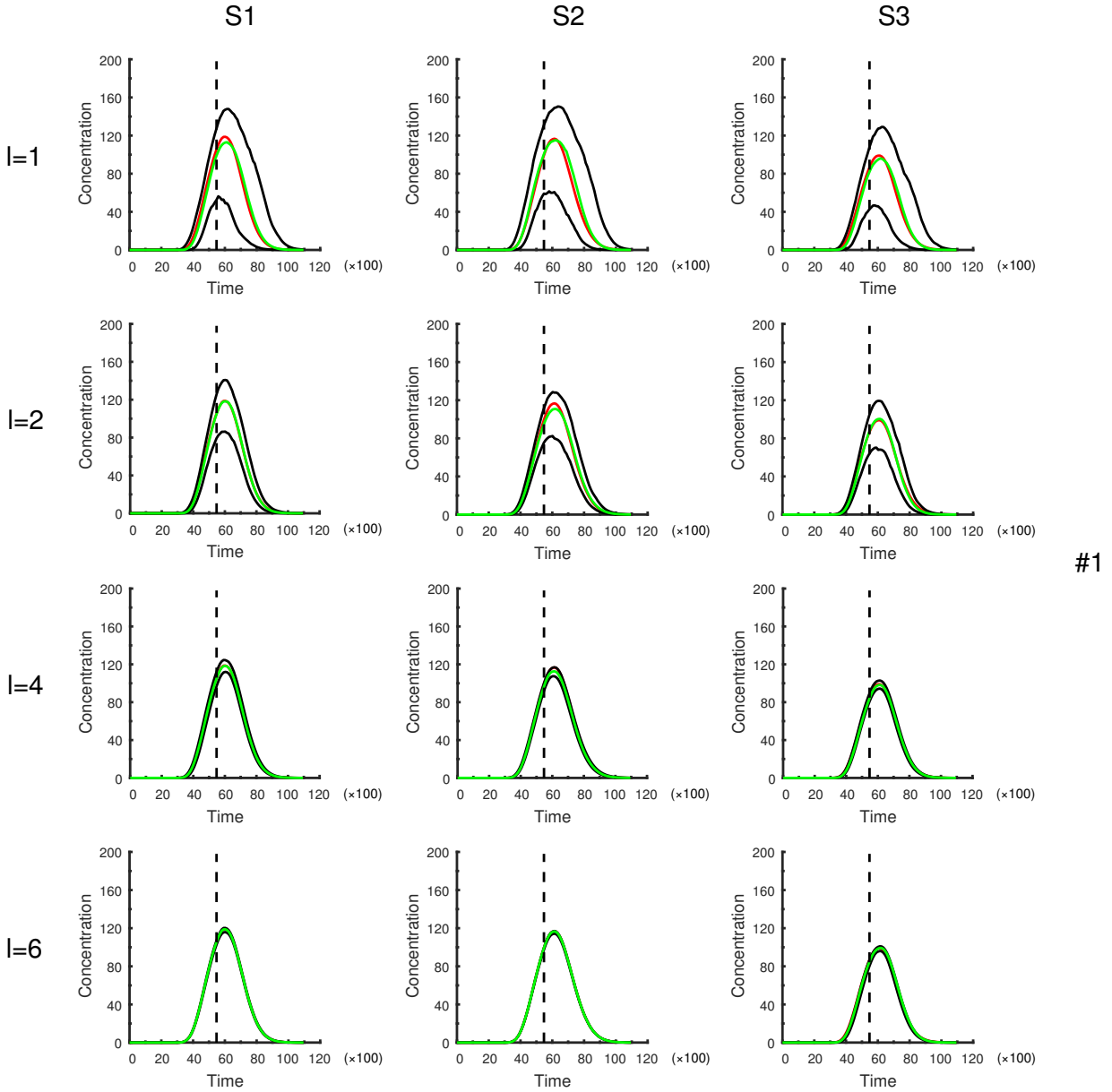


Figure 13: Scenarios S1-S3. Time evolution of the contaminant concentrations at the verification well #1 computed with the updated ensembles of source parameters after the 1st, 2st, 4th and 6th assimilation iterations. The red line corresponds to the evolution of the concentration in the reference; the 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.

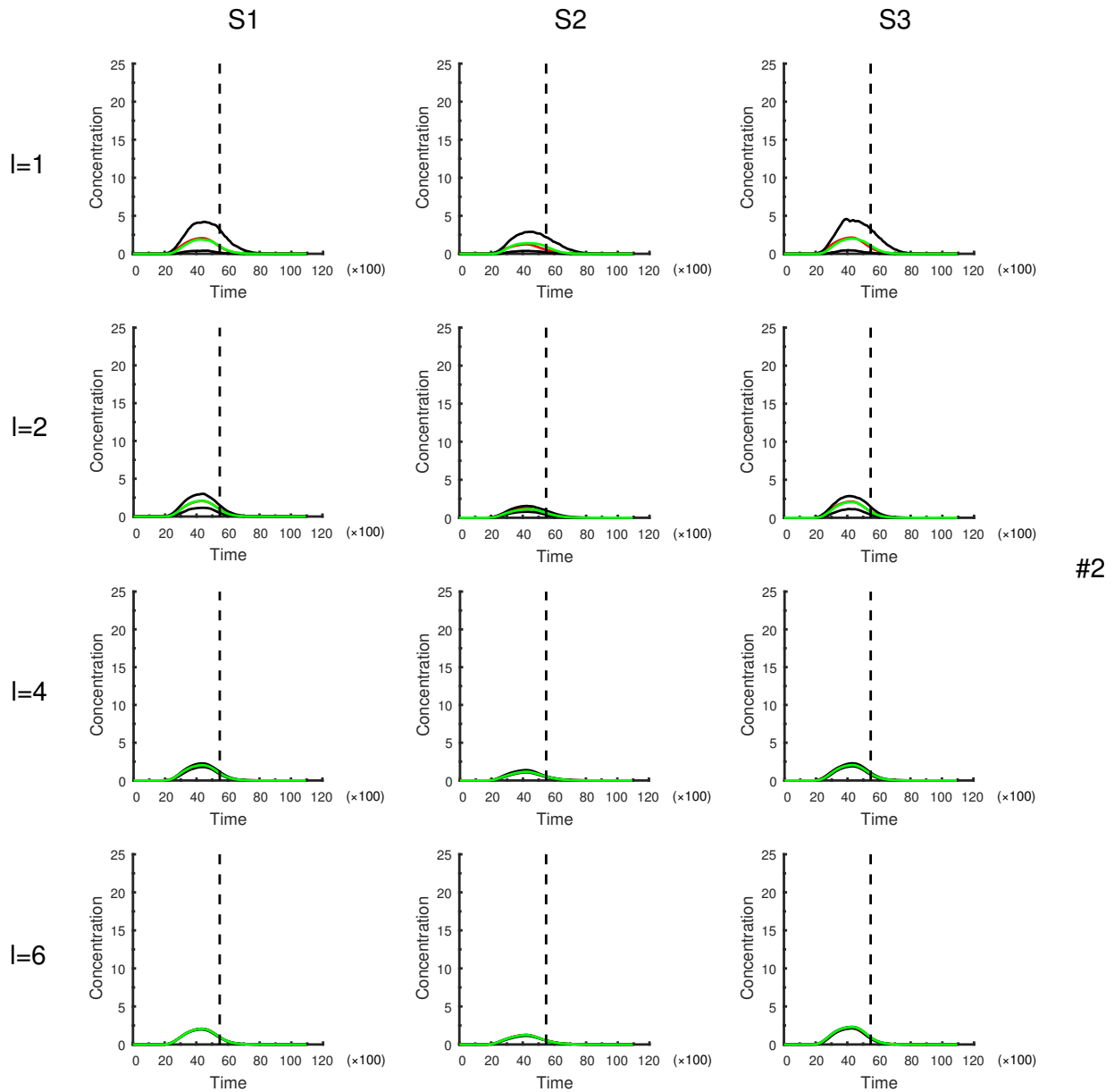


Figure 14: Scenarios S1-S3. Time evolution of the contaminant concentrations at the verification well #2 computed with the updated ensembles of source parameters after the 1st, 2st, 4th and 6th assimilation iterations. The red line corresponds to the evolution of the concentration in the reference; the 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.

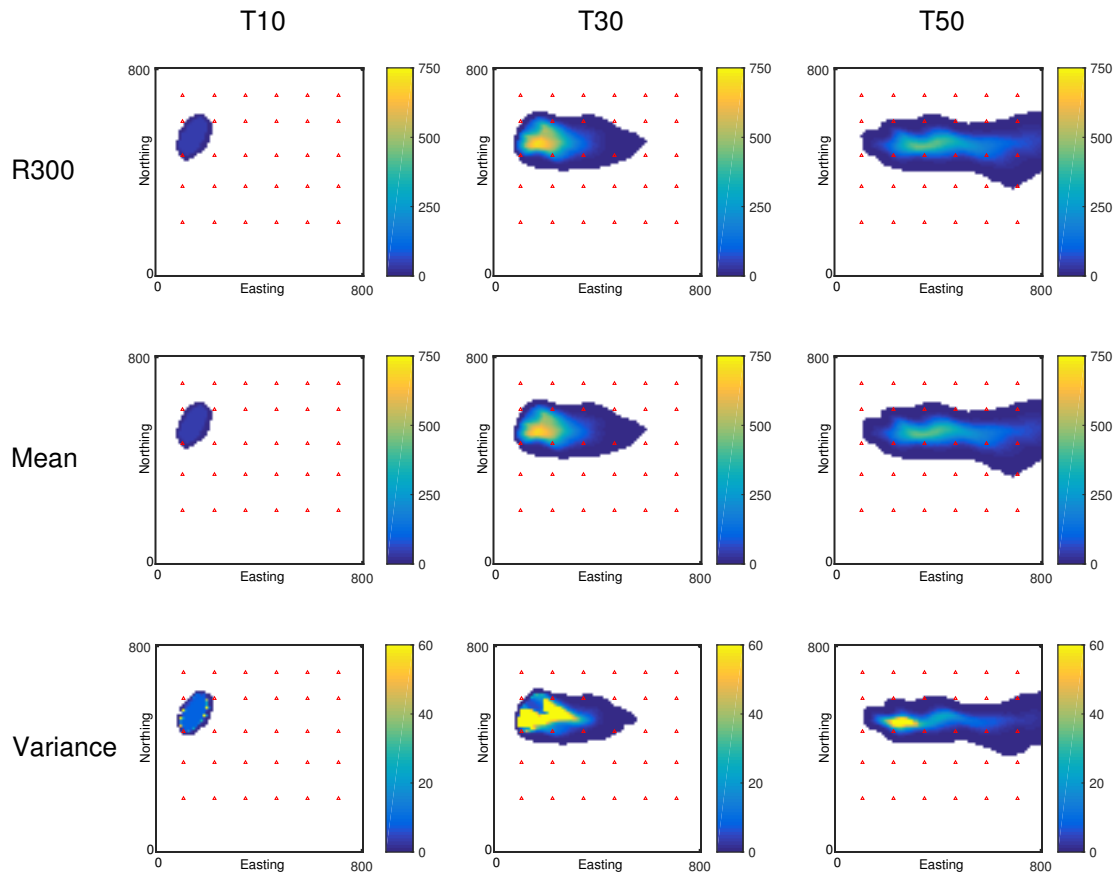


Figure 15: Scenario S1. Contaminant plume at the 10th, 30th and 30th simulation time steps, computed with the updated parameters after the 6th assimilation iteration. From top to bottom, plume in realization #300; ensemble mean of all plumes, and ensemble variance of all plumes.

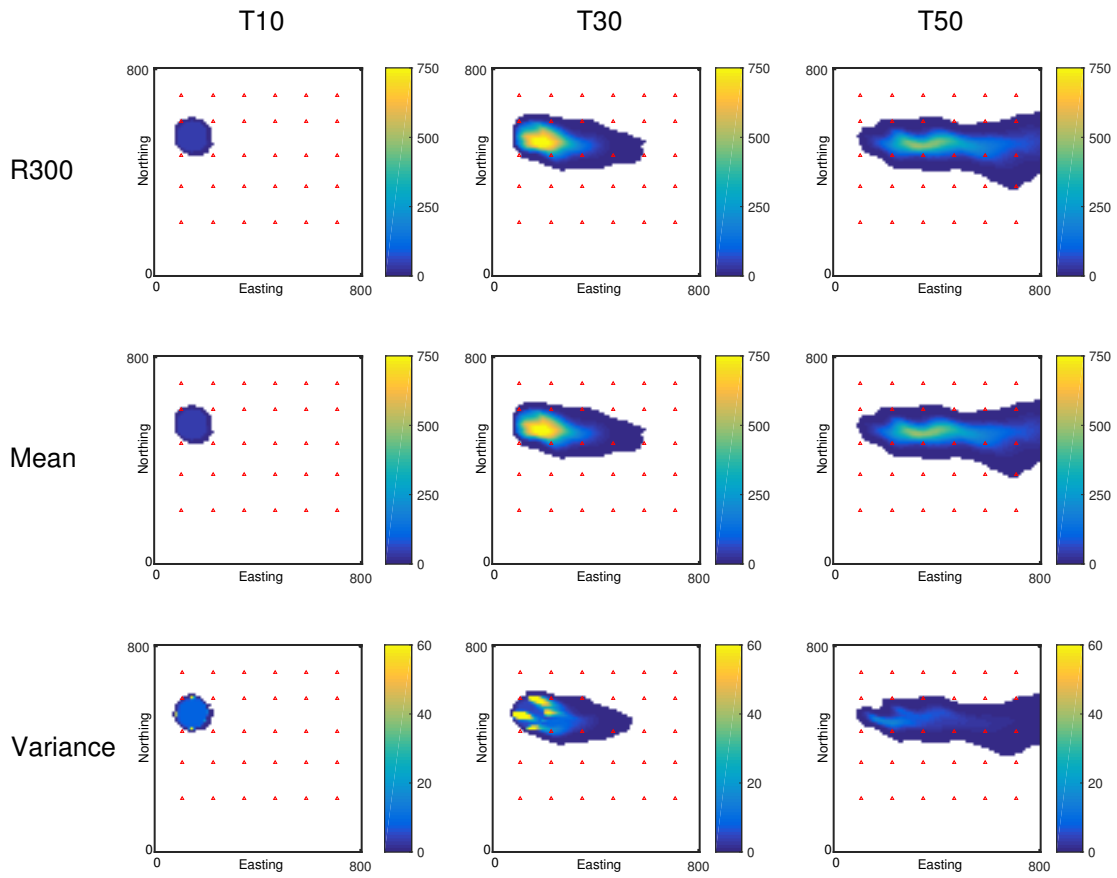


Figure 16: Scenario S2. Contaminant plume at the 10th, 30th and 30th simulation time steps, computed with the updated parameters after the 6th assimilation iteration. From top to bottom, plume in realization #300; ensemble mean of all plumes, and ensemble variance of all plumes.

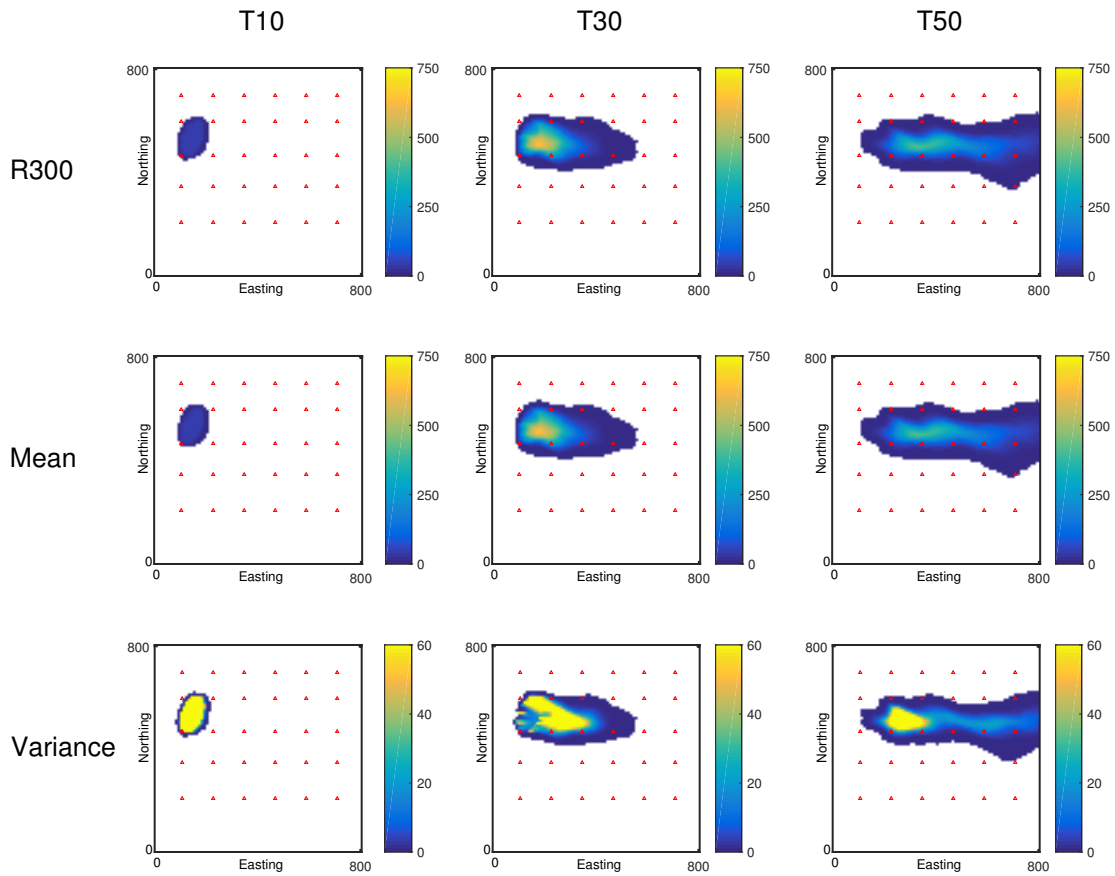


Figure 17: Scenario S3. Contaminant plume at the 10th, 30th and 30th simulation time steps, computed with the updated parameters after the 6th assimilation iteration. From top to bottom, plume in realization #300; ensemble mean of all plumes, and ensemble variance of all plumes.

211 5. Summary and discussion

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

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

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

230 **Acknowledgements** Financial support to carry out this work was received from the
231 financial support from the Fundamental Research Funds for the Central Universities
232 (B200201015) and Jiangsu Specially-Appointed Professor Program from Jiangsu Provincial
233 Department of Education (B19052). J.J. Gómez-Hernández acknowledges the financial
234 support from the Spanish Ministry of Science and Innovation through project number
235 PID2019-109131RB-I00. C. Lu acknowledges the National Natural Science Foundation of

236 China (51879088), Fundamental Research Funds for the Central Universities (B200204002),
237 and the Natural Science Foundation of Jiangsu Province (BK 20190023). Y. Xie
238 acknowledges the Fundamental Research Funds for the Central Universities (B210202018).
239 J. Yang acknowledges the National Natural Science Foundation of China (52009032) and
240 the Fundamental Research Funds for the Central Universities (B210202019)

241 **References**

242 **References**

- 243 Aral, M.M., Guan, J., Maslia, M.L., 2001. Identification of contaminant source location and
244 release history in aquifers. *Journal of hydrologic engineering* 6, 225–234.
- 245 Atmadja, J., Bagtzoglou, A.C., 2001. Pollution source identification in heterogeneous porous
246 media. *Water Resources Research* 37, 2113–2125.
- 247 Ayvaz, M.T., 2007. Simultaneous determination of aquifer parameters and zone structures
248 with fuzzy c-means clustering and meta-heuristic harmony search algorithm. *Advances in*
249 *water resources* 30, 2326–2338.
- 250 Ayvaz, M.T., 2016. A hybrid simulation–optimization approach for solving the areal ground-
251 water pollution source identification problems. *Journal of Hydrology* 538, 161–176.
- 252 Bagtzoglou, A.C., Atmadja, J., 2003. Marching-jury backward beam equation and quasi-
253 reversibility methods for hydrologic inversion: Application to contaminant plume spatial
254 distribution recovery. *Water Resources Research* 39.
- 255 Barzegar, R., Moghaddam, A.A., Tziritis, E., Fakhri, M.S., Soltani, S., 2017. Identification of
256 hydrogeochemical processes and pollution sources of groundwater resources in the marand
257 plain, northwest of iran. *Environmental Earth Sciences* 76, 1–16.

- 258 Butera, I., Tanda, M.G., Zanini, A., 2013. Simultaneous identification of the pollutant release
259 history and the source location in groundwater by means of a geostatistical approach.
260 Stochastic Environmental Research and Risk Assessment 27, 1269–1280.
- 261 Capilla, J.E., Gómez-Hernández, J.J., Sahuquillo, A., 1998. Stochastic simulation of trans-
262 missivity fields conditional to both transmissivity and piezometric head data—3. applica-
263 tion to the culebra formation at the waste isolation pilot plan (wipp), new mexico, usa.
264 Journal of Hydrology 207, 254–269.
- 265 Capilla, J.E., Rodrigo, J., Gómez-Hernández, J.J., 1999. Simulation of non-gaussian trans-
266 missivity fields honoring piezometric data and integrating soft and secondary information.
267 Mathematical Geology 31, 907–927.
- 268 Chen, Z., Gómez-Hernández, J.J., Xu, T., Zanini, A., 2018. Joint identification of contam-
269 inant source and aquifer geometry in a sandbox experiment with the restart ensemble
270 kalman filter. Journal of Hydrology 564, 1074–1084.
- 271 Chen, Z., Xu, T., Gómez-Hernández, J.J., Zanini, A., 2021. Contaminant spill in a sandbox
272 with non-gaussian conductivities: Simultaneous identification by the restart normal-score
273 ensemble kalman filter. Mathematical Geosciences , 1–29.
- 274 Crestani, E., Camporese, M., Baú, D., Salandin, P., 2013. Ensemble Kalman filter versus
275 ensemble smoother for assessing hydraulic conductivity via tracer test data assimilation.
276 Hydrology and Earth System Sciences 17, 1517.
- 277 Cupola, F., Tanda, M.G., Zanini, A., 2015. Contaminant release history identification in
278 2-d heterogeneous aquifers through a minimum relative entropy approach. SpringerPlus
279 4, 656.
- 280 Dimov, I., Jaekel, U., Vereecken, H., 1996. A numerical approach for determination of sources
281 in transport equations. Computers & Mathematics with Applications 32, 31–42.

- 282 Emerick, A.A., Reynolds, A.C., 2013. Ensemble smoother with multiple data assimilation.
283 Computers & Geosciences 55, 3–15.
- 284 Franssen, H.H., Gómez-Hernández, J., 2002. 3d inverse modelling of groundwater flow at
285 a fractured site using a stochastic continuum model with multiple statistical populations.
286 Stochastic Environmental Research and Risk Assessment 16, 155–174.
- 287 Gómez-Hernández, J.J., Journel, A.G., 1993. Joint sequential simulation of Multi-Gaussian
288 fields, in: Soares, A. (Ed.), Geostatistics Tróia '92, Kluwer Academic Publishers, Dor-
289 drecht. pp. 85–94.
- 290 Gómez-Hernández, J.J., Xu, T., 2021. Contaminant source identification in aquifers: A
291 critical view. Mathematical Geosciences accepted.
- 292 Gorelick, S.M., Evans, B., Remson, I., 1983. Identifying sources of groundwater pollution:
293 an optimization approach. Water Resources Research 19, 779–790.
- 294 Jamshidi, A., Samani, J.M.V., Samani, H.M.V., Zanini, A., Tanda, M.G., Mazaheri, M.,
295 2020. Solving inverse problems of unknown contaminant source in groundwater-river in-
296 tegrated systems using a surrogate transport model based optimization. Water 12, 2415.
- 297 Jin, X., Mahinthakumar, G.K., Zechman, E.M., Ranjithan, R.S., 2009. A genetic algorithm-
298 based procedure for 3D source identification at the Borden emplacement site. Journal of
299 Hydroinformatics 11, 51–64.
- 300 Li, L., Zhou, H., Hendricks Franssen, H., Gómez-Hernández, J., 2012. Groundwater flow
301 inverse modeling in non-multigaussian media: performance assessment of the normal-score
302 ensemble kalman filter. Hydrology and Earth System Sciences 16, 573.
- 303 Ma, R., Zheng, C., Zachara, J.M., Tonkin, M., 2012. Utility of bromide and heat tracers

304 for aquifer characterization affected by highly transient flow conditions. *Water Resources*
305 *Research* 48.

306 Mahar, P.S., Datta, B., 2000. Identification of pollution sources in transient groundwater
307 systems. *Water Resources Management* 14, 209–227.

308 Mahinthakumar, G., Sayeed, M., 2005. Hybrid genetic algorithm-local search methods for
309 solving groundwater source identification inverse problems. *Journal of water resources*
310 *planning and management* 131, 45–57.

311 McDonald, M.G., Harbaugh, A.W., 1988. A modular three-dimensional finite-difference
312 ground-water flow model. volume 6. US Geological Survey Reston, VA.

313 Mirghani, B.Y., Mahinthakumar, K.G., Tryby, M.E., Ranjithan, R.S., Zechman, E.M.,
314 2009. A parallel evolutionary strategy based simulation–optimization approach for solving
315 groundwater source identification problems. *Advances in Water Resources* 32, 1373–1385.

316 Neupauer, R.M., Borchers, B., Wilson, J.L., 2000. Comparison of inverse methods for re-
317 constructing the release history of a groundwater contamination source. *Water Resources*
318 *Research* 36, 2469–2475.

319 Neupauer, R.M., Wilson, J.L., 1999. Adjoint method for obtaining backward-in-time location
320 and travel time probabilities of a conservative groundwater contaminant. *Water Resources*
321 *Research* 35, 3389–3398.

322 Sidauruk, P., Cheng, A.D., Ouazar, D., 1998. Ground water contaminant source and trans-
323 port parameter identification by correlation coefficient optimization. *Ground Water* 36,
324 208–214.

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

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

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

331 Sun, A.Y., Painter, S.L., Wittmeyer, G.W., 2006b. A robust approach for iterative contam-
332 inant source location and release history recovery. *Journal of contaminant hydrology* 88,
333 181–196.

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

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

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

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

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

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

- 350 Xu, T., Gómez-Hernández, J.J., 2018. Simultaneous identification of a contaminant source
351 and hydraulic conductivity via the restart normal-score ensemble kalman filter. *Advances*
352 *in Water Resources* 112, 106–123.
- 353 Xu, T., Gómez-Hernández, J.J., Chen, Z., Lu, C., 2021. A comparison between es-mda and
354 restart enkf for the purpose of the simultaneous identification of a contaminant source and
355 hydraulic conductivity. *Journal of Hydrology* 595, 125681.
- 356 Xu, T., Gómez-Hernández, J.J., Zhou, H., Li, L., 2013. The power of transient piezometric
357 head data in inverse modeling: An application of the localized normal-score EnKF with
358 covariance inflation in a heterogenous bimodal hydraulic conductivity field. *Advances in*
359 *Water Resources* 54, 100–118.
- 360 Zeng, L., Shi, L., Zhang, D., Wu, L., 2012. A sparse grid based bayesian method for
361 contaminant source identification. *Advances in Water Resources* 37, 1–9.
- 362 Zhang, J., Zeng, L., Chen, C., Chen, D., Wu, L., 2015. Efficient bayesian experimental
363 design for contaminant source identification. *Water Resources Research* 51, 576–598.
- 364 Zheng, C., 2010. MT3DMS v5. 3Supplemental users guide: Tuscaloosa, Ala., University of
365 Alabama Department of Geological Sciences. Technical Report. Technical Report to the
366 US Army Engineer Research and Development Center.
- 367 Zhou, H., Gómez-Hernández, J.J., Li, L., 2012. A pattern-search-based inverse method.
368 *Water Resources Research* 48.