Ensemble smoother with multiple data assimilation to simultaneously estimate the source location and the release history of a contaminant spill in an aquifer

Valeria Todaro[a,b,*], Marco D'Oria[a], Maria Giovanna Tanda[a], J. Jaime Gómez-Hernández[a,b]

[a]*Department of Engineering and Architecture, University of Parma, Parma, 43124, Italy*
[b]*Institute for Water and Environmental Engineering, Universitat Politènica de València, València, 46022, Spain*

Abstract

The source location and the time history of a pollutant released in an aquifer are very relevant information for the design of effective remediation strategies. Usually, their identification requires solving an inverse problem when the only available information about the groundwater contamination event is a sparse set of concentration data collected in the aquifer at a few points downstream from the source. Here, a novel approach is proposed to solve the inverse problem: the use of the Ensemble Smoother with Multiple Data Assimilation (ES-MDA) in the context of source contamination identification. This method is used for the simultaneous determination of the time history and the source location of a pollutant release based on observed concentration data and a calibrated numerical model of groundwater flow and mass

*Corresponding author at: Department of Engineering and Architecture, University of Parma, 43124, Parma, Italy.

Email addresses: valeria.todaro@unipr.it (Valeria Todaro), marco.doria@unipr.it (Marco D'Oria), mariagiovanna.tanda@unipr.it (Maria Giovanna Tanda), jaime@dihma.upv.es (J. Jaime Gómez-Hernández)

transport in the aquifer. The ES-MDA is demonstrated in two case studies. The first one is based on an analytical solution of the flow and transport equations, aimed at the estimation of the source location and the release history of a nonreactive pollutant spreading in a two-dimensional homogeneous aquifer from a point source. For this case, different alternatives are considered for the spatial distribution of the observation points, the concentration sampling frequency, the ensemble size and the use of covariance localization and covariance inflation techniques in the formulation of the smoother. The purpose of this case is to test the new approach, analyze its performance and also to identify the conditions that render the problem ill-posed and, therefore, without solution; also, in this case, a new spatiotemporal iterative localization is presented. In the second case study, we use real data collected in a laboratory sandbox that reproduces a vertical cross-section of an unconfined aquifer with two-dimensional quasi-parallel flow between constant-head boundaries. The results show that the location, time and number of observations, the ensemble size and the application of covariance localization and covariance inflation techniques have an impact on the final solution. A well-designed monitoring network and the application of covariance corrections improve the performance of the ES-MDA and help avoiding ill-posedness and equifinality. The application to laboratory data validates the potential of ES-MDA to simultaneously estimate the time history and the source location of a pollutant released in groundwater in real cases.

Keywords: Inverse modeling, Ensemble Kalman filter method, Groundwater contaminant source, Covariance localization, Stochastic analysis

1. Introduction

Monitoring, protection and restoration of aquifers have received a lot of attention in the past decades, thanks to the growing interest in environmental issues and the importance of groundwater quality for water supply. The first steps in any remediation strategies of a polluted aquifer should be the identification of the source location and the release history of the contaminant. They would allow to identify the cause of the contamination, to implement an effective remediation plan and to share the costs among the responsible parties.

When groundwater contamination is first detected, the source location and the release history are usually unknown. Recovering these variables from sparse data of the spatial distribution of the pollutant concentration in the aquifer is a type of inverse problem. Inverse problems are inherently ill-posed, which means that the solution is generally non-unique and could be not stable to small perturbations of the data. Several deterministic and stochastic methods have been proposed to solve this problem. The first category includes Tikhonov regularization (Skaggs and Kabala, 1994); nonlinear optimization with embedding (Mahar and Datta, 1997); non-regularized nonlinear least squares (Alapati and Kabala, 2000); progressive genetic algorithms (Aral et al., 2001); constrained robust least squares (Sun et al., 2006) and heuristic harmony search algorithms (Ayvaz, 2010). The second category includes probability-based methods such as statistical pattern recognition (Datta et al., 1989); minimum relative entropy (Woodbury and Ulrych, 1996; Woodbury et al., 1998; Cupola et al., 2015); geostatistical approaches (Snodgrass and Kitanidis, 1997; Michalak and Kitanidis, 2004a,b;

Neupauer et al., 2000; Butera and Tanda, 2003; Butera et al., 2006, 2012; Gzyl et al., 2014; Cupola et al., 2015); empirical Bayesian methods combined with Akaike's Bayesian Information Criterion (Zanini and Woodbury, 2016); Bayesian global optimization (Pirot et al., 2019) and ensemble Kalman filter methods (Xu and Gómez-Hernández, 2016, 2018; Chen et al., 2018; Xu et al., 2020).

However, only a few of the presented studies allow to simultaneously identify the source location and the release history of a groundwater contaminant. The method proposed by Aral et al. (2001) used a progressive genetic algorithm to solve an iterative nonlinear optimization problem, in which the source location and release history were explicitly defined as continuous unknown variables and contaminant concentrations were used as observations. Sun et al. (2006) combined a constrained robust least squares estimator with a global optimization solver for iteratively identifying release histories and source locations on the basis of concentration measurements. Ayvaz (2010) used an optimization method based on the heuristic harmony search algorithm to identify locations and release histories for pollution sources, minimizing residuals between the simulated and measured contaminant concentrations. All these methods are deterministic and do not allow to quantify the uncertainty of the results.

Butera et al. (2012) applied a Bayesian geostatistical approach for the simultaneous identification of the release function and the source location based on concentration data. The methodology has then been tested by Cupola et al. (2015) on real data collected in a laboratory sandbox. The method requires a preliminary delineation of possible sources and some hy-

potheses about the structure of the unknown release function. The approach aims to recover the contaminant release history considering all the possible sources simultaneously and selecting the location where the highest amount of pollutant is estimated. The method adopts a transfer function approach for the solution of the forward problem (Butera et al., 2006).

We propose a new procedure for the joint identification of the source location and the release history of a pollutant in an aquifer: the use of an Ensemble Smoother with Multiple Data Assimilation (ES-MDA) in the context of contaminant source identification. The ES-MDA, introduced by Emerick and Reynolds (2012, 2013a), has been mainly applied to reservoir history matching problems (Emerick and Reynolds, 2013b; Fokker et al., 2016; Zhao et al., 2016), but its popularity is growing also in hydrology (Lan et al., 2018; Li et al., 2018, 2019; Kang et al., 2019; Song et al., 2019; Todaro et al., 2019; Bao et al., 2020). It is an iterative data assimilation method based on the Ensemble Kalman Filter (EnKF), initially proposed by Evensen (1994). In particular, the ES-MDA is a variant of the Ensemble Smoother (ES) proposed by van Leeuwen and Evensen (1996). Unlike the EnKF, which performs a sequential update one step at a time assimilating the data as they are collected, the ES and the ES-MDA simultaneously assimilate all the available observation data. Also, the ES-MDA iteratively assimilates the same data multiple times leading to better results for strongly nonlinear problems than the ES, which performs a single global update (Evensen, 2018).

The main advantages of the ES-MDA are: i) its capability to be used with almost any forward model for the solution of inverse problems; ii) the possibility of being implemented with parallel computing, and iii) its ca-

pability to select a best estimate under different criteria and to assess its uncertainty, through the analysis of an ensemble of realizations. Compared with the Bayesian geostatistical approach (Butera et al., 2012), the ES-MDA does not require the explicit time-consuming calculation of sensitivity matrices to solve the inverse problem, since they are embedded in the covariance matrices of the ensemble. Moreover, it allows the simulation of groundwater flow and mass transport even in complex cases.

As all the inverse approaches, also the proposed method computes the unknown parameters based on the knowledge of observed data. In this work, the parameters to identify are represented by the spatial coordinates of the contaminant source location and the time-discretized release history; the observations are sparse concentration data measured at different monitoring locations and times. Notice that, in general, piezometric head data will be available, which could also be assimilated and used in the solution of the inverse problem; it is not the case in the laboratory experiment described next, for which no piezometric head data were available.

Two applications of the ES-MDA are presented. First, the ES-MDA is used to solve a synthetic case from the literature with the purpose of showing its capabilities and to obtain guidelines for its application to real cases. Second, the ES-MDA is used to validate the methodology on experimental data collected in a laboratory sandbox that mimics an unconfined aquifer.

The synthetic case study allows to investigate in detail the inverse procedure with a limited computational effort. In particular, we evaluated the impact of the observation sampling scheme and different algorithm settings in the context of ill-posedness of inverse problems. The ill-conditioning in-

creases as uncertainties about the model increase and as the quantity and quality of the observed data decrease. Therefore, it is important to design a monitoring network that makes a good compromise between valuable information about the concentration evolution and the costs of monitoring actions, which would limit the number of monitoring points.

The study also addresses the problem of undersampling present in ensemble-based methods; it occurs when the ensemble size is so small that it is not statistically representative of the variability of the unknowns. Although large ensembles mitigates this problem, the computational cost increases with the ensemble size; therefore, it is advantageous to solve the problem with the smallest possible ensemble. Covariance localization has been developed to overcome this problem; it helps in removing long-range spurious correlations and mitigates the ensemble rank deficiency, allowing the use of a small number of realizations. Localization can be achieved by different ways (Houtekamer and Mitchell, 1998; Hamill et al., 2001; Anderson, 2007b; Chen and Oliver, 2009). Covariance localization is generally based on the spatial distance between parameter locations and observations; in this study, parameters and observations are also time-dependent, furthermore the distance between them is not fixed since the source position is unknown, what complicates the use of standard localization techniques. Todaro et al. (2019) proposed a temporal localization considering time lapses rather than spatial distances. A new localization approach is presented, which takes into account both spatial and temporal distances and iteratively updates the distance between the unknown parameters and the observations. Covariance inflation is also considered to overcome undersampling problems (Anderson and An-

derson, 1999; Anderson, 2007a; Li et al., 2009; Liang et al., 2011; Wang and Bishop, 2003; Zheng, 2009); it modifies the original ES-MDA adjusting the ensemble spread to avoid smoother divergence.

Hence, the presented study aims to provide an efficient methodology to solve the contaminant source identification problem. The manuscript is organized as follows: first, the forward problem, its solution and the ES-MDA procedure are described. Then, the synthetic and the laboratory case study are presented and discussed. The manuscript ends with some conclusions.

2. Methods

2.1. Forward problem

The forward problem is based on the groundwater flow and mass transport equations. In particular, we consider an incompressible fluid in saturated porous media and a non-reactive contaminant injected in the aquifer at a point subject to advection and dispersion (Bear, 1972; Bear and Verruijt, 1987). Assuming a uniform porosity, initial condition $C(\mathbf{x}, 0) = 0$, and boundary condition, $C(\infty, t) = 0$, where $C(\mathbf{x}, t)\,[\mathrm{ML}^{-3}]$ is the solute concentration, the transport equation can be solved by the convolution integral

$$C(\mathbf{x}, t) = \int_0^t s(\mathbf{x_0}, \tau)\, g(\mathbf{x}, t - \tau)\, d\tau. \tag{1}$$

The term $s(\mathbf{x_0}, t)\,[\mathrm{MT}^{-1}]$ is the the contaminant flux injected into the aquifer through the source located at $\mathbf{x_0}$ given by

$$s(\mathbf{x_0}, t) = C_0(t) \cdot q_0(\mathbf{x_0}, t), \tag{2}$$

where $C_0(t)\,[\mathrm{ML}^{-3}]$ is the concentration of the released pollutant at time t and $q_0(\mathbf{x_0}, t)\,[\mathrm{L}^3\mathrm{T}^{-1}]$ is the injection flow rate. The term $g(\mathbf{x}, t - \tau)$ is a

Kernel function that represents the response at location $\mathbf{x}$ and time t to a pulse injection at the source location $\mathbf{x_0}$ and time τ.

Defining with $\mathbf{D}(\mathbf{x})\,[\mathrm{L^2T^{-1}}]$ the hydrodynamic dispersion coefficient tensor and with $\mathbf{v}(\mathbf{x},t)\ [\mathrm{LT^{-1}}]$ the effective flow velocity, in two-dimensional cases, with uniform flow, $v_y = 0$ and constant dispersion coefficients, the Kernel function can be determined analytically. With these assumptions, the solution of Eq. (1) is

$$C(x,y,t) = \int_0^t s(x_0,y_0,\tau)\,\frac{1}{4\pi\sqrt{D_x D_y\,(t-\tau)}} \cdot \exp\left[-\frac{((x-x_0)-v_x\,(t-\tau))^2}{4D_x\,(t-\tau)} - \frac{(y-y_0)^2}{4D_y\,(t-\tau)}\right] d\tau. \tag{3}$$

For complex cases in which the flow field is not uniform (for instance, non-isotropic and heterogeneous aquifers), the advection-dispersion equation can not be solved analytically and it is necessary to employ numerical methods. Here, for the second case study for which the analytical solution cannot be used, the flow equation is solved using the numerical model MODFLOW (Harbaugh, 2005), and the transport equation with MT3DMS (Zheng and Wang, 1999).

2.2. Ensemble smoother with multiple data assimilation

In this work, the iterative Ensemble Smoother with Multiple Data Assimilation method (ES-MDA) is used to solve a parameter estimation problem in which the unknown parameters are updated based on the available observations. The ES-MDA procedure is extensively described by Emerick and Reynolds (2013a) and Evensen (2018); here, an overview of the method and the scheme to perform the spatiotemporal iterative localization are presented.

The vector of unknown parameters is defined as: $\mathbf{X} = (x_s, y_s, s_1, s_2, ..., s_k)^T$, where x_s is the x-coordinate of the source, y_s is the y-coordinate and $(s_1, s_2, ..., s_k)$ is the discretized-in time release history; the number of parameters to be estimated depends on the duration of the groundwater pollution event to be simulated and the time step selected for the discretization. The vector of observations ($\mathbf{D}$) is composed of measured concentrations at different times and monitoring locations. A first fundamental assumption is that a reliable forward model is available since the relationship between parameters and observations must be known; in our case, the forward model is represented by a calibrated groundwater flow and solute transport model, that is, the parameters of both models will not be subject of further identification. Having a calibrated flow and transport model is probably not a very realistic assumption but the purpose of the current paper is the testing of the ES-MDA for the identification of time-varying point contaminant sources. The simultaneous estimation of the parameters controlling the flow and transport equations is left for further investigation.

The ES-MDA scheme can be summarized in the three following steps:

1. Initialization step.

   An initial ensemble of parameters must be defined taking into account all the available prior information. Often, no data are available and the ensemble is generated using prior distributions based on expert knowledge. The release history is modeled as a continuous function of time and, for this reason, imposing some degree of continuity in the initial realizations will facilitate the identification process. This can be achieved with proper parameterization of the time functions to

be generated. The ensemble of the spatial coordinates of the source is generated using random values selected over a uniform distribution wide enough to bound the true location. After the initialization step, the number of iterations has to be decided and the next two steps are repeated as many times as iterations there are.

2. Forecast step.

   Each realization j of the ensemble is used as input to the forward model and an ensemble of predictions ($\mathbf{Y}$) at measurement locations over time is obtained. For the first iteration, $\mathbf{Y}$ is generated using the initial ensemble of parameters; then the ensemble of predictions is generated using the updated parameters from the last iteration,

$$\mathbf{Y}_{j,i} = \psi\left(\mathbf{X}_{j,i}\right). \tag{4}$$

   The operator $\psi\left(\cdot\right)$ denotes the forward model and i is the iteration index.

3. Update step.

   Parameters are updated for each realization of the ensemble j and iteration i according to the following equation

$$\mathbf{X}_{j,i+1} = \mathbf{X}_{j,i} + \mathbf{C}^i_{\mathbf{XY}}(\mathbf{C}^i_{\mathbf{YY}} + \alpha_i\mathbf{R})^{-1}\left(\mathbf{D} + \sqrt{\alpha_i}\varepsilon_j - \mathbf{Y}_{j,i}\right), \tag{5}$$

   where ε_j is the observation error, which is drawn from a Gaussian distribution of mean zero and covariance matrix $\mathbf{R}$, $\mathcal{N}\left(0, \mathbf{R}\right)$; α_i is a coefficient that, at each iteration i, inflates the measurement error and its covariance matrix. The values of α_i are chosen following a decreasing sequence; in this way, the magnitude of the updates for the first iterations, when the misfit between predictions and observation may be

too large, is small to reduce the magnitude of the initial updates; also, the coefficients α_i must satisfy the following expression (Emerick and Reynolds, 2013a)

$$\sum_{i=1}^{N} \frac{1}{\alpha_i} = 1, \tag{6}$$

where N is the total number of iterations. $\mathbf{C}^i_{\mathbf{XY}}$ is the cross-covariance matrix between parameters and predictions and $\mathbf{C}^i_{\mathbf{YY}}$is the autocovariance matrix of predictions. They are computed from the ensemble at each iteration i as

$$\mathbf{C}^i_{\mathbf{XY}} = \frac{1}{N_e - 1} \sum_{j=1}^{N_e} \left(\mathbf{X}_{j,i} - \overline{\mathbf{X}}_i\right) \left(\mathbf{Y}_{j,i} - \overline{\mathbf{Y}}_i\right)^T, \tag{7}$$

$$\mathbf{C}^i_{\mathbf{YY}} = \frac{1}{N_e - 1} \sum_{j=1}^{N_e} \left(\mathbf{Y}_{j,i} - \overline{\mathbf{Y}}_i\right) \left(\mathbf{Y}_{j,i} - \overline{\mathbf{Y}}_i\right)^T, \tag{8}$$

where N_e is the total number of ensemble realizations, $\overline{\mathbf{X}}_i$ is the ensemble mean of the parameters and $\overline{\mathbf{Y}}_i$ is the ensemble mean of the predictions. When covariance localization is applied, Eq. (7) and (8) are modified as follows

$$\widetilde{\mathbf{C}}^i_{\mathbf{XY}} = \rho^i_{\mathbf{XY}} \circ \mathbf{C}^i_{\mathbf{XY}}, \tag{9}$$

$$\widetilde{\mathbf{C}}^i_{\mathbf{YY}} = \rho_{\mathbf{YY}} \circ \mathbf{C}^i_{\mathbf{YY}}. \tag{10}$$

where $\circ$ represents the elementwise multiplication and $\rho^i_{\mathbf{XY}}$ and $\rho_{\mathbf{YY}}$ are correlation matrices based on spatial and temporal distances between parameters and observations and between observations and observations, respectively. The correlations in space ($\rho^i_{\mathbf{XY},s}$, $\rho_{\mathbf{YY},s}$) and

time ($\rho_{\mathbf{XY},t}$, $\rho_{\mathbf{YY},t}$) are computed independently and then coupled via a Schur product

$$\rho^i_{\mathbf{XY}} = \rho^i_{\mathbf{XY},s} \circ \rho_{\mathbf{XY},t}, \tag{11}$$

$$\rho_{\mathbf{YY}} = \rho_{\mathbf{YY},s} \circ \rho_{\mathbf{YY},t}. \tag{12}$$

We use the fifth-order correlation function introduced by Gaspari and Cohn (1999), which smoothly reduces the correlations between points for increasing distances and cuts off long-range correlations above a specific distance

$$\rho = \begin{cases} -\frac{1}{4}\left(\frac{\delta}{b}\right)^5 + \frac{1}{2}\left(\frac{\delta}{b}\right)^4 + \frac{5}{8}\left(\frac{\delta}{b}\right)^3 - \frac{5}{3}\left(\frac{\delta}{b}\right)^2 + 1, & 0 \le \delta \le b, \\ \frac{1}{12}\left(\frac{\delta}{b}\right)^5 - \frac{1}{2}\left(\frac{\delta}{b}\right)^4 + \frac{5}{8}\left(\frac{\delta}{b}\right)^3 + \frac{5}{3}\left(\frac{\delta}{b}\right)^2 & \\ -5\left(\frac{\delta}{b}\right) + 4 - \frac{2}{3}\left(\frac{\delta}{b}\right)^{-1}, & b \le \delta \le 2b, \\ 0 & \delta \ge 2b, \end{cases} \tag{13}$$

where δ represents the parameter-observation or observation-observation distances in space ($\delta^i_{XY,s}$, $\delta_{YY,s}$) or time ($\delta_{XY,t}$, $\delta_{YY,t}$). The spatial distances between parameters and observations are unknown since the coordinates of the source are to be estimated; therefore, $\delta^i_{XY,s}$ must be updated at each iteration i considering the source located at the coordinates given by the ensemble means of x_0 and y_0. The coefficient b characterizes the space (b_s) or time (b_t) distance at which the covariances become zero.

At the end of each update step, linear relaxation and covariance inflation are used to prevent smoother divergence. Linear relaxation reduces the magnitude of the update at the end of an iteration. When linear

relaxation is used the expression of Eq. (5) is replaced with

$$\widetilde{\mathbf{X}}_{j,i+1} = (1-w)\,\mathbf{X}_{j,i+1} + w\mathbf{X}_{j,i}, \tag{14}$$

where w is a relaxation coefficient between 0 and 1. Covariance inflation is applied using the scheme proposed by Anderson and Anderson (1999) where the ensemble is linearly inflated around its mean by an inflation factor (r) slightly larger than 1

$$\widetilde{\mathbf{X}}_{j,i+1} = r\left(\mathbf{X}_{j,i+1} - \overline{\mathbf{X}}_{i+1}\right) + \overline{\mathbf{X}}_{i+1}. \tag{15}$$

In this work, the update step is performed in log-space in order to prevent the appearance of unphysical negative values. The vector of parameters is log transformed before the update step and back transformed into the parameter space before the forecast step.

Then, the scheme is repeated from step 2, after setting $\mathbf{X}_{j,i} = \mathbf{X}_{j,i-1}$, until the last iteration.

3. Case studies

The proposed approach is demonstrated on two case studies. First, the ES-MDA is applied to an analytical case study with the aim to show the capabilities of the method to simultaneously identify a contaminant source location and its release history in an aquifer. In this case, the forward model requires a small computational time and the results can be compared with a reference solution. This also allows to investigate different configurations of the inverse algorithm, in order to determine the optimal setting to be used for real cases. The second application validates the methodology on experimental data collected in a laboratory sandbox experiment.

3.1. Analytical case

The analytical case simulates a pollution event in an infinite homogeneous two-dimensional aquifer, with uniform flow, as result of the injection of a nonreactive contaminant at a point (Butera and Tanda, 2003). It is assumed that the water discharge $q_0(\mathbf{x_0}, t)$ is of unit value and small enough such that it does not affect the uniform groundwater flow. Therefore, the release history $s(\mathbf{x_0}, t)$, defined in Eq. (2), is equivalent to the concentration history $C_0(t)$. All quantities are considered with unspecified but consistent units. The uniform velocity and the dispersion coefficients are assumed known: $v = 1$, $D_x = 1$ and $D_y = 0.1$. We use the same expression for the release function $s_r(\mathbf{x_0}, t)$ used elsewhere (Skaggs and Kabala, 1994; Woodbury and Ulrych, 1996; Snodgrass and Kitanidis, 1997; Butera and Tanda, 2003; Butera et al., 2012; Zanini and Woodbury, 2016) to define the reference solution

$$\begin{aligned} s_r(\mathbf{x_0}, t) = &\exp\left(-\frac{(t-130)^2}{50}\right) + 0.3\exp\left(-\frac{(t-150)^2}{200}\right) \\ &+ 0.5\exp\left(-\frac{(t-190)^2}{98}\right). \end{aligned} \tag{16}$$

The actual source location $\mathbf{x_0}$ is $x_0 = 50$ and $y_0 = 20$. The concentration history has a total duration of 300; it is discretized into 101 intervals with a time step of $\Delta t = 3$ resulting in a total number of parameters to be estimated $N_p = 103$ (the two spatial coordinates plus the 101 temporal solute fluxes). The reference release function, depicted in Fig. 1, is used to obtain the reference observations, which are computed by evaluating Eq. (3) using numerical integration.

Different test cases are carried out to investigate the impact of the observation sampling scheme, ensemble size, covariance localization and inflation

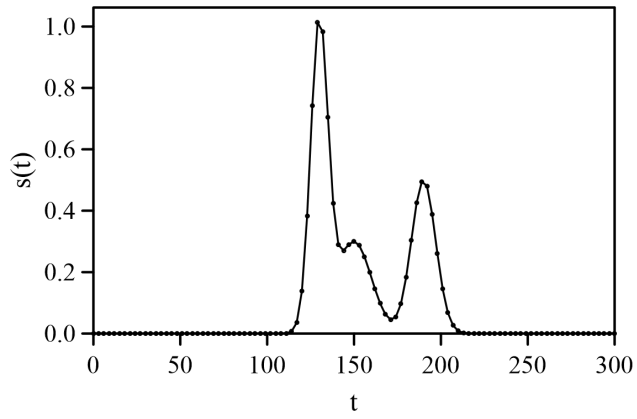

Figure 1: Analytical case: reference release history.

techniques. The test cases will be evaluated in terms of equifinality, that is, when different source functions are identified that are consistent with the observations, and in terms of sensitivity to the initial ensemble values. For this purposes, for each test case, 100 experiments were performed to identify the source history changing only the random component of the initial ensemble and the observation measurement errors. At the end of each experiment, the performance of the method is evaluated using the following metrics:

- The Nash-Sutcliffe efficiency criterion (NSE) to evaluate the agreement between the actual and estimated release history:

$$NSE = \left(1 - \frac{\sum_{i=1}^{N_p-2}\left(\overline{X}_i - s_{r,i}\right)^2}{\sum_{i=1}^{N_p-2}\left(s_{r,i} - \overline{s}_r\right)^2}\right) \cdot 100, \tag{17}$$

where $N_p - 2$ is equal to 101, the number of intervals used to discretize $s(t)$; $s_{r,i}$ represents the discretized source function and is the i-th actual amount of released contaminant, $\overline{s}_{r,i}$ is the time average of the reference release history ($\frac{1}{N_p-2}\sum_{i=1}^{N_p-2} s_{r,d}$) and $\overline{X}_i$ is the ensemble mean of the i-th estimated amount of released contaminant ($\frac{1}{N_e}\sum_{j=1}^{N_e} X_i^j$, with X_i^j

the final estimate of parameter X_i in realization j). The closer to 100, the better.

- The root mean square error ($RMSE$) between observations and model predictions:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m} \left(D_i - \overline{Y}_i\right)^2}{m}} \tag{18}$$

where D_i is the i-th observed concentration and $\overline{Y}_i$ is the ensemble mean of the i-th predicted concentration ($\frac{1}{N_e}\sum_{j=1}^{N_e} Y_i^j$, with Y_i^j the prediction of Y_i in realization j). The closer to zero, the better.

- The spatial distance between the true and estimated source location (L):

$$L = \sqrt{\left(\overline{x}_s - x_0\right)^2 + \left(\overline{y}_s - y_0\right)^2} \tag{19}$$

where $\overline{x}_s$ and $\overline{y}_s$ are the ensemble means of the estimated spatial coordinates of the source and (x_0, y_0) is the true source location. The closer to zero, the better.

These metrics are compared with reference threshold values to evaluate the performance of the method. We consider three cases: i) good performance when the reproduction of the observed concentrations is good, the identification of the source location is good and the identification of the release function is good; ii) equifinality performance, when reproduction of the observed concentrations is good, but neither the source location nor the release function are well identified; iii) poor performance, otherwise:

i) Good performance when

$RMSE < RMSE_{thr}$ and $NSE > NSE_{thr1}$ and $L < L_{thr}$

Table 1: Threshold values used to define test criteria.

$RMSE_{thr}$	4σ
NSE_{thr1}	70
NSE_{thr2}	60
L_{thr}	5

ii) Equifinality performance when

$RMSE < RMSE_{thr}$ and ($NSE < NSE_{thr2}$ or $L > L_{thr}$)

iii) Otherwise, fail.

There is not a standard criterion for the definition of metric thresholds to assess goodness-of-fit (see e.g. Moriasi et al., 2007; Ritter and Muñoz-Carpena, 2013). In this study, we consider the performance of the method to be good if NSE>0.7 and unsatisfactory if NSE<0.6. The fit between predictions and observations is considered to be good when the $RMSE$ is less than the maximum assumed error. Since the observation errors are normally distributed, the maximum error is defined as 4σ, where σ is its standard deviation. The selected threshold values ($RMSE_{thr}$, NSE_{thr1}, NSE_{thr2}, L_{thr}) are summarized in Table 1. With these criteria, it is possible to define the percentage of successful tests, tests with multiple solutions and failed tests for each case, on the basis of the 100 experiments.

3.1.1. Impact of the observation network geometry and sampling frequency

The effect of the spatial distribution of the observation points is evaluated. For this case, a large ensemble was used to avoid the need of using localization or inflation techniques in the implementation of ES-MDA. The observation

network geometries used, displayed in Fig. 2, are:

A. Concentrations collected at two monitoring points, located on the same line as the source ($y = 20$) at points (150, 20) and (200, 20), and 31 sampling times from $T = 0$ up to $T = 450$ with a time step $\Delta t = 15$. The total number of observations is $m = 2 \cdot 31 = 62$.

B. Concentrations collected at 21 monitoring points distributed on the same line of the source ($y = 20$) at uniform intervals between $x = 90$ and $x = 290$; only one observation from each location at time $T = 300$. The total number of observations is $m = 22 \cdot 1 = 22$.

C. Concentrations collected at four monitoring points distributed on the same line of the source ($y = 20$) at x-coordinates 80, 115, 150 and 185, and the same 31 sampling times of set A. The total number of observations is $m = 4 \cdot 31 = 124$.

D. Concentrations collected at four monitoring points vertically distributed on the line $x = 150$ and at y-coordinates 11, 16, 21 adn 26; the sampling times are the same as for sets A and C. The total number of observations is $m = 4 \cdot 31 = 124$.

A random observation error ε normally distributed with zero mean and variance $5{\cdot}10^{-8}$ for all the performed tests is considered. The initial ensemble of parameters is composed of 1000 realizations. The realizations of the source coordinates are uniformly distributed random values selected in the range [5, 80] for x and [10, 30] for y. The realizations of the release history are normal

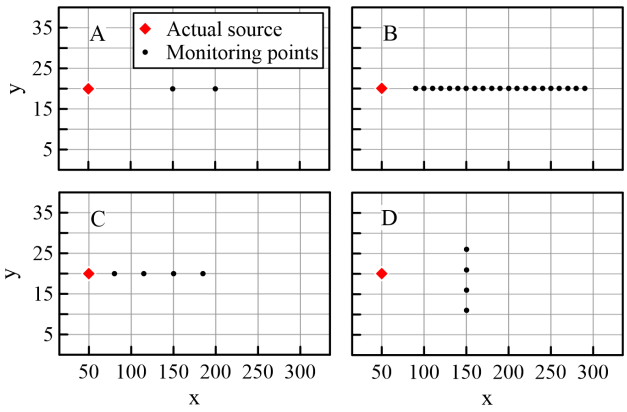

Figure 2: Analytical case: location of the measurement points for sets A, B, C and D; the red diamond is the actual source location.

functions described by the following expression:

$$f(t) = \Delta + \Gamma \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)}, \tag{20}$$

where t is the time, Δ is a base amount of released concentration, Γ is the volume under the Gaussian function of mean μ and variance σ^2. These coefficients are selected randomly from uniform distributions, $\Delta \in U[1{\cdot}10^{-10}, 1{\cdot}10^{-3}]$, $\Gamma \in U[10, 40]$, $\mu \in U[89, 210]$ and $\sigma \in U[6, 59]$. The ES-MDA is run with 10 iterations and a decreasing series of α values following the sequence [113.33; 75.55; 50.37; 33.58; 22.39; 14,92; 9.95; 6.63; 4.42; 2.95].

Table 2 summarizes the results of the four test cases, T denotes the percentage of successful tests over the 100 synthetic experiments and E indicates the percentage of synthetic experiments in which equifinality is detected.

The observation network geometry greatly impacts the final results. The synthetic experiments that give reliable solutions (NSE>70 and L<5) are less than 21% for observation sets A, B and C. Furthermore, equifinality occurs in large proportions for cases A and B, and to a lesser extent for case

Table 2: ES-MDA performance for observations sets A, B, C and D and ensemble size N_e=1000. T indicates the percentage of successful tests and E the percentage of tests that present equifinality.

A	B	C	D
T:10%	T:19%	T:21%	T:98%
E:53%	E:34%	E:12%	E:0%

C. Only in case D, the ES-MDA is able to identify successfully the source location and the release function without equifinality.

3.1.2. Impact of the ensemble size and application of localization and inflation techniques

The test cases designed to investigate the impact of the ensemble size, covariance localization and inflation techniques make use of the observation set D. We tested five ensemble sizes N_e of 1000, 500, 250, 100 and 50 with and without covariance corrections. The number of iterations, α values, and distributions used to generate the initial ensembles are the same ones used in the previous section. Covariance localization is applied using the coefficients b_s equal to 210 and b_t equal to 300. The factor r used for the covariance inflation is equal to 1.01. The results obtained from each set of 100 synthetic experiments are reported in Table 3. The ES-MDA performs better for increasing ensemble sizes and when covariance inflation and localization techniques are applied. The percentage of successful tests is high for large ensembles, with even better numbers when covariance corrections are applied. The presence of equifinality is detected when the ensemble size reduces, but the corrections on the algorithm help to reduce it. The effects of

Table 3: ES-MDA performance for observation set D and ensemble sizes of 1000, 500, 250, 100 and 50, with and without corrections on the covariance calculation. T indicates the percentage of succesful tests and E the percentage of tests that present equifinality.

N_e	without corrections	with corrections
1000	T:98%	T:100%
	E:0%	E:0%
500	T:85%	T:96%
	E:8%	E:0%
250	T:71%	T:87%
	E:19%	E:4%
100	T:46%	T:64%
	E:43%	E:14%
50	T:20%	T:45%
	E:60%	E:29%

covariance and inflation techniques are more evident for small ensemble sizes; considering N_e equal to 100, the percentage of successful tests is 46% for the experiments without corrections and 64% for those with corrections; multiple solutions are detected for 43% of the experiments without corrections and for 14% of those with corrections. The tests computed with the smaller ensemble size (N_e=50) lead to unsatisfactory results with a percentage of successful tests lower than 45% and a high probability of equifinality.

For the sake of brevity, we show only the results of one of the tests performed with a small ensemble size of 100 realizations and with corrections in the computation of the covariance. Among the 100 synthetic experiments, we selected as the best estimate of the release function the median of the

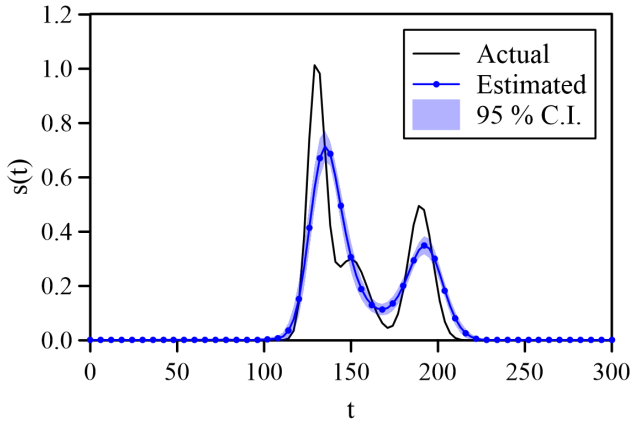

Figure 3: Analytical case: actual and estimated release history with 95% uncertainty interval resulting from a test performed with $N_e = 100$ and observation set D.

successful tests, and we use the set of successful tests to build uncertainty intervals about the median. In Fig. 3 the reference solution and the ensemble median with its 95% uncertainty interval are depicted. Figure 4 shows a comparison between observed and predicted concentrations at observation locations. The ES-MDA reproduces quite well the release history and the source location estimate is very close to the true one (x_0=50, y_0=20). The NSE is 80.46% and the ensemble means of x and y coordinates are, respectively, equal to 52.66 ($\pm$1.78, 95% uncertainty interval) and 20.00 ($\pm$0.06, 95% uncertainty interval). The test leads to a good match between observations and predictions with an RMSE at the last iteration equal to $3.3{\cdot}10^{-4}$ and a narrow 95% uncertainty interval.

3.2. Experimental case

The second case study uses a laboratory experimental dataset following the work by Cupola et al. (2014). The experimental device is a sandbox that

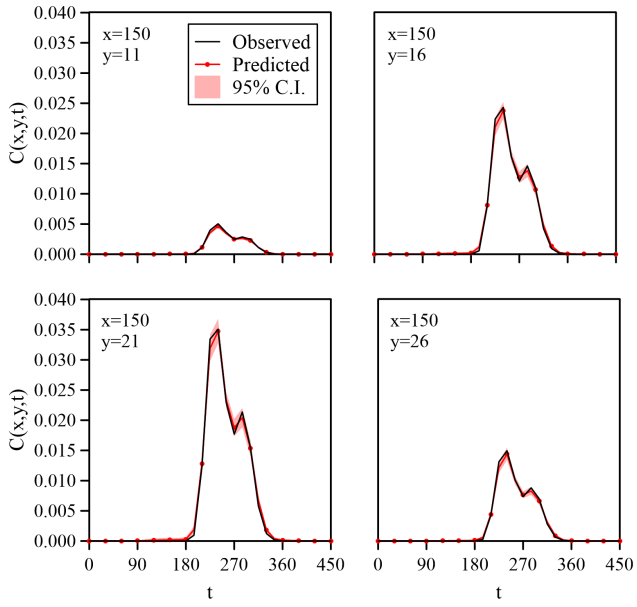

Figure 4: Analytical case: observed and predicted concentrations with 95% uncertainty interval.

reproduces an unconfined aquifer characterized by two-dimensional flow in a vertical plane. The sandbox has external dimensions of 120 cm $\times$ 14 cm $\times$ 73 cm and it is made of three parts along the longitudinal direction: upstream and downstream tanks and an internal chamber of 95 cm $\times$ 10 cm $\times$ 70 cm, which contains the porous media consisting of glass beads with diameter in the range between 0.75 mm and 1 mm. The flow is governed by constant upstream and downstream water levels equal to 59.9 cm and 53.6 cm above the horizontal bottom of the tank, respectively. Fluorescein sodium salt was used as tracer solution and it was injected at a variable mass rate through an injector located in the upstream part of the sandbox at coordinates $x = 14.25$ cm and $y = 32.75$ cm, that extends through the entire thickness of the sandbox. The test had a duration of 2200 s; the injection started at time 310

s and ended at 1800 s; the concentration of the fluorescein sodium salt is constant and equal to 20 mg·l^{-1}, while the flow rate changes over time. The resulting mass rate ranges from 0 to about 55 μg·l^{-1} and presents three peaks of different magnitude. The observed concentrations are recorded over the entire sandbox by taking pictures with a digital camera and then converting luminosity into concentration through image processing techniques (for more details see Citarella et al. (2015)). Modeling is performed in two dimensions, since no lateral movement orthogonal to the sandbox plane is expected. A comparison between the results obtained with a two-dimensional model and a three-dimensional one is reported by Uribe-Asarta (2019), showing no differences between the two models.

The inverse methodology requires a calibrated numerical model able to describe as accurately as possible the forward process. Groundwater flow was modeled with MODFLOW 2005 (Harbaugh, 2005) and mass transport with MT3DMS (Zheng and Wang, 1999). The effect of the injection on the background flow is not negligible; therefore, a transient flow model is considered. The numerical model was preliminaryly calibrated by an inverse procedure not reported here for brevity. After the calibration, and for the purposes of the source identification, this model is used throughout. Table 4 summarizes the parameters of the flow and transport models and Figure 5 shows the hydraulic conductivity field after the calibration process. The estimated field is slightly heterogeneous and conductivity is anisotropic, even though the sandbox was filled with glass beads of almost the same size with the intention of reproducing an isotropic homogeneous field. Our interpretation of the lower conductivity values towards the bottom of the sandbox is

Table 4: Transport and hydraulic parameters of the numerical model.

Porosity	0.37
Average hydraulic conductivity (cm s^{-1})	0.673
Ratio of horizontal to vertical conductivity (Kh/Kv)	3.267
Specific storage coefficient (cm^{-1})	10^{-4}
Longitudinal dispersivity (cm)	0.178
Transverse dispersivity (cm)	0.065

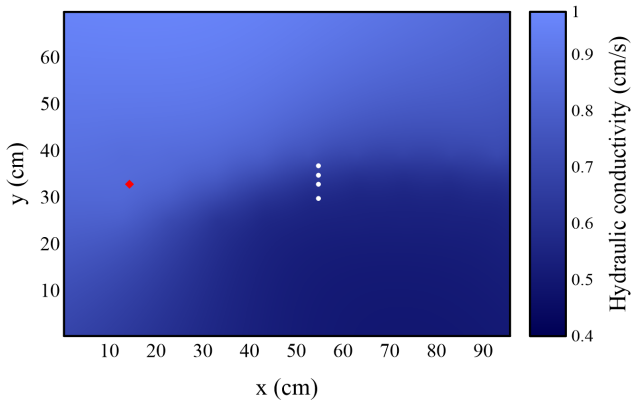

Figure 5: Hydraulic conductivity field. The red diamonds denotes the actual source location. The white dots are the monitoring points.

that it is due to additional compaction during the filling process.

Since the concentration of the contaminant is known, the estimation of the release history is limited to identifying the injected flow rate. The release duration is discretized into 72 intervals with a time step of $\Delta t = 3$ s resulting in a total number of parameters $N_p = 74$, of which two are the spatial coordinates of the source. The initial ensemble of parameters is made up of 81 realizations ($N_e = 81$); the spatial coordinates of the source are random values selected from uniform distributions $x \in U[5, 30]$ cm, and $y \in U[30, 34]$

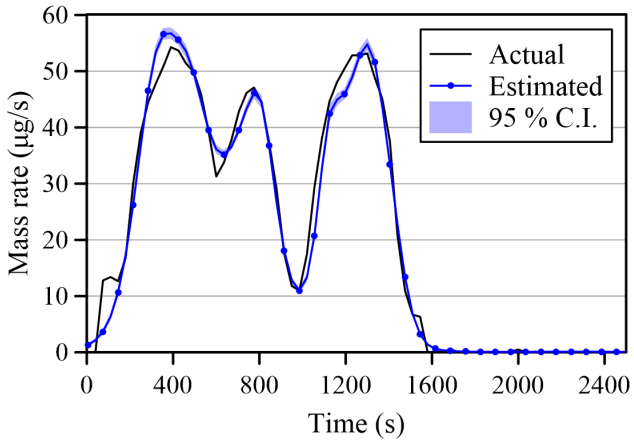

Figure 6: Experimental case: actual and estimated release history with 95% uncertainty interval. Time 0 s represents the time at which injection starts.

cm. The initial realizations of the injected flow rate history follow expression Eq. (20), with parameters selected randomly from the following uniform distributions, $\Delta \in \mathcal{U}[1 \cdot 10^{-10},\ 1{\cdot}10^{-1}]$, $\Gamma \in U[800, 1000]$, $\mu \in U[490, 1400]$ and $\sigma \in U[60, 365]$. The four monitoring points are vertically distributed on the line $x = 54.75$ cm and at y-coordinates 29.00, 32.75, 34.75 and 36.75 cm. For each monitoring point, the observed concentrations are recorded at 45 sampling times from $T = 0$ s to $T = 2200$ s (total number of monitoring data is $m = 180$). The random measurement error ε is assumed normally distributed with zero mean and variance $1{\cdot}10^{-2}$ $(\text{mg}{\cdot}\text{l}^{-1})^2$. The ES-MDA with 6 iterations and decreasing α=[63.0; 31.5; 15.8; 7.88 3.9; 2.0] is used for the inversion. Covariance localization and covariance inflation are applied using the coefficients b_s=200, b_t=2500 and r=1.01, and linear relaxation with the coefficient w=0.1.

Figure 6 shows the results of the experimental case; the ensemble mean of

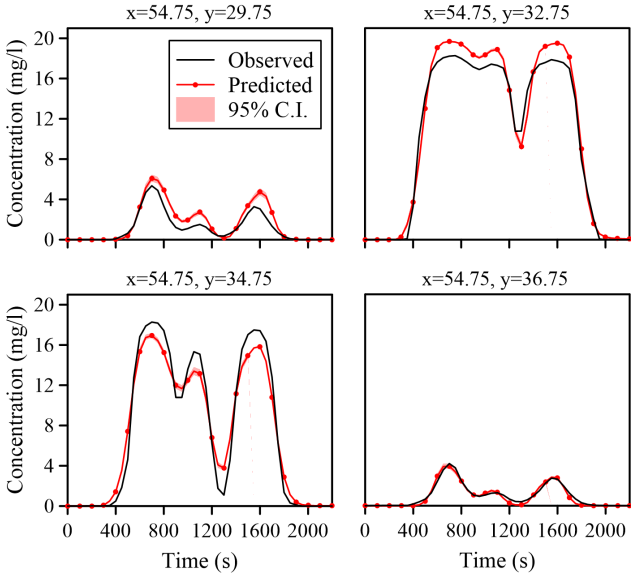

Figure 7: Experimental case: observed and predicted concentrations with 95% uncertainty intervals. Time 0 s represents the time at which injection starts.

the release history with its 95% confidence interval and the true solution are depicted. The ES-MDA leads to a good agreement between the two curves with an NSE value equal to 98.34% and with a satisfactory representation of peak magnitudes and times. The ensemble means of the x and y coordinates of the source are, respectively, equal to 14.71 cm ($\pm$0.45 cm, 95% uncertainty interval) and 32.91 cm ($\pm$0.14 cm, 95% uncertainty interval); the distance between the true and estimated source location is less than 0.5 cm. In Fig. 7, the experimental and predicted observations are compared. The retrieved source parameters reproduce quite well the observed concentrations with a narrow 95% uncertainty interval; the RMSE at the last iteration is equal to 0.96 mg$\cdot$l^{-1}, which is comparable with the experimental observation errors.

4. Discussion and Conclusions

In this paper, a novel application of the Ensemble Smoother with Multiple Data Assimilation (ES-MDA) is proposed for the simultaneous identification of the source location and the release history of a groundwater contamination event from observed sparse concentration data collected downstream from the spill. The procedure is tested by means of an analytical case study and an experimental one.

The analytical case serves to demonstrate the capability of the ES-MDA to solve this type of inverse problem and to analyze the impact of the different settings on the final identification. The impact of the observation network geometry and density, ensemble size, covariance and inflation techniques and also the effect of different sets of initial realizations are investigated. The aim was to find out a configuration that leads to a reliable solution and mitigates the ill-conditioned nature of inverse problems. Equifinality is analyzed in the analytical case, finding that there are some network geometries that may lead to acceptable results (in terms of reproduction of the observed concentrations) but with very different release functions.

The effect of the observation network geometry and density is evaluated considering four sets of observed concentrations, a large ensemble size (N_e=1000) and the other factors being the same. The results show that location, time and number of observations significantly impact the final solution obtained by the ES-MDA; for the sets in which the observations are located in a line parallel to the main flow direction, the percentage of successful tests is low and equifinality is detected. Instead, for the set with the observations in a line orthogonal to the main flow direction, the number of successful tests

is 98% and the algorithm simultaneously estimates the release history and the source location. We find that placing the observation locations in a line orthogonal to the main flow directions is more informative than placing the observation locations along the same line. In the latter case, it is easy to think of multiple solutions that should lead to the same observations, for instance, by estimating the source location in the direction orthogonal to flow symmetrically with respect to the line of observations. This indicates the importance of a good design of the observation network, since if observations provide poor information, the ill-posed inverse problem is difficult to solve and the impact of random factors increases; it is also noteworthy that, in real cases, only a limited number of concentration measurements are available given the field sampling costs; for this reason, an optimal design of new monitoring points has a great relevance.

The observation set orthogonal to the flow direction is used to check the effect of the ensemble size and the application of covariance localization and covariance inflation techniques in the performance of the ES-MDA. In this paper, a new procedure to apply the covariance localization is presented. Covariance localization was commonly performed taking into account the fixed spatial distance between observation-observation and parameter-observation only; here, the spatial and temporal distances are both considered and, furthermore, the parameter-observation spatial distance is iteratively updated since the location of the parameters is an unknown of the problem.

The results show that the ES-MDA works better when large ensembles and the correction to the covariances are used, demonstrating the capability of the proposed spatiotemporal iterative localization to improve the ES-MDA

performance. The percentage of successful tests increases with the ensemble size and the covariance corrections and, at the same time, the chances that equifinality happens decrease. Covariance inflation and, in particular, covariance localization, overcome the undersampling problems noticed in the ensemble-based methods; and for this reason, their effects are more evident for small ensemble sizes. The tests performed with an ensemble size of 50 realizations lead to unreasonable results with a low percentage of passed tests and a high percentage of tests with multiple solutions. We suggest to use, for this type of problems, ensemble sizes greater than the number of unknown parameters to identify.

It is noteworthy to point out that another aspect to take into account is the impact on the solution of the errors on both the observations and the model structure. Small measurement errors can improve the ES-MDA results when the model is perfect and the observations are uncorrupted, as in the synthetic case study. However, overfitting problems and ensemble collapse can arise for real cases, which are always affected by uncertainty in the forward model and measurement noises. In these cases, the modeler should use an appropriate level of fit based on the quality of the available observation and the model. The effects of the errors on the ES-MDA performance will be investigated in future works.

The experimental case study uses real data collected in a laboratory test. The experimental device is a sandbox that reproduces an unconfined aquifer under controlled conditions; it allows to validate the ES-MDA methodology in a real test case. The algorithm parameters, such as the monitoring network and the ensemble size, were chosen after the results of the analytical study.

For this case, the initial ensemble of source coordinates has been generated considering a limited suspect area, which guarantees that all the realizations of the ensemble are representative. This decision was taken based on preliminary tests performed with large suspect areas. Even if it is not mandatory that the initial ensemble contains the solution, a well designed ensemble helps to reach better results.

The results prove the capability of ES-MDA to solve this type of inverse problem in a real cases, when the available observations are usually noisy. The method reproduces very well both the contaminant release history and the spatial coordinates of the source; the NSE is about 98% and the distance between the true and estimated source location is less than 0.5 cm.

To the best of our knowledge, this is the first work that uses a stochastic method for the simultaneous identification of the source location and the release history. It allows to assess the estimation uncertainty and to directly estimate the spatial coordinates of the source, unlike, for example, the Bayesian geostatistical approach that only identifies the most probable location among a set of possible source points defined a priori.

Another innovative aspect of this work is the use of the ES-MDA method for the estimation of time-dependent parameters. In hydrogeology, ensemble Kalman methods are usually applied for the investigation of groundwater field parameters that are time-independent such as porosity or hydraulic conductivity. In this study, the parameters to be estimated are identified performing a discretization in time of the release history of a contaminant into an aquifer, which is time dependent.

In summary, the proposed procedure is a novelty method able to simulta-

neously recover the release history and the source location of a groundwater pollutant on the basis of sparse observed concentration data. A well-designed monitoring network and the application of covariance localization and covariance inflation techniques lead to satisfactory results and reduce the inherent equifinality encountered in parameter estimation problems.

Acknowledgements

The TeachinParma initiative, co-funded by Fondazione Cariparma and University of Parma (http://www.teachinparma.com/about/) supported Prof. J. Jaime GÃşmez- HernÃąndez as Visiting Professor at the University of Parma. Project PID2019-109131RB-I00 financed by the Spanish Ministry of Science and Innovation is also gratefully acknowledged.

References

Alapati, S., Kabala, Z.J., 2000. Recovering the release history of a groundwater contaminant using a non-linear least-squares method. Hydrological Processes 14, 1003–1016. doi:10.1002/(SICI)1099-1085(20000430)14:6<1003::AID-HYP981>3.0.CO;2-W.

Anderson, J.L., 2007a. An adaptive covariance inflation error correction algorithm for ensemble filters. Tellus A: Dynamic Meteorology and Oceanography 59, 210–224. doi:10.1111/j.1600-0870.2006.00216.x.

Anderson, J.L., 2007b. Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter. Physica D: Nonlinear Phenomena 230, 99–111. doi:10.1016/j.physd.2006.02.011.

Anderson, J.L., Anderson, S.L., 1999. A monte carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. Monthly Weather Review 127, 2741–2758. doi:10.1175/1520-0493(1999)127<2741:amciot>2.0.co;2.

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. doi:10.1061/(asce)1084-0699(2001)6:3(225).

Ayvaz, M.T., 2010. A linked simulation–optimization model for solving the unknown groundwater pollution source identification problems. Journal of Contaminant Hydrology 117, 46–59. doi:10.1016/j.jconhyd.2010.06.004.

Bao, J., Li, L., Redoloza, F., 2020. Coupling ensemble smoother and deep learning with generative adversarial networks to deal with non-gaussianity in flow and transport data assimilation. Journal of Hydrology 590, 125443.

Bear, J., 1972. Dynamics of fluids in porous media. American Elsevier Publishing Company, New York.

Bear, J., Verruijt, A., 1987. Modeling groundwater flow and pollution. volume 2. Springer Science & Business Media.

Butera, I., Tanda, M.G., 2003. A geostatistical approach to recover the release history of groundwater pollutants. Water Resources Research 39. doi:10.1029/2003WR002314.

Butera, I., Tanda, M.G., Zanini, A., 2006. Use of numerical modelling to identify the transfer function and application to the geostatistical proce-

dure in the solution of inverse problems in groundwater. Journal of Inverse and Ill-posed Problems 14, 547–572. doi:10.1163/156939406778474532.

Butera, I., Tanda, M.G., Zanini, A., 2012. Simultaneous identification of the pollutant release history and the source location in groundwater by means of a geostatistical approach. Stochastic Environmental Research and Risk Assessment 27, 1269–1280. doi:10.1007/s00477-012-0662-1.

Chen, Y., Oliver, D.S., 2009. Cross-covariances and localization for EnKF in multiphase flow data assimilation. Computational Geosciences 14, 579–601. doi:10.1007/s10596-009-9174-6.

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. doi:10.1016/j.jhydrol.2018.07.073.

Citarella, D., Cupola, F., Tanda, M.G., Zanini, A., 2015. Evaluation of dispersivity coefficients by means of a laboratory image analysis. Journal of Contaminant Hydrology 172, 10–23. doi:10.1016/j.jconhyd.2014.11.001.

Cupola, F., Tanda, M.G., Zanini, A., 2014. Laboratory sandbox validation of pollutant source location methods. Stochastic Environmental Research and Risk Assessment 29, 169–182. doi:10.1007/s00477-014-0869-4.

Cupola, F., Tanda, M.G., Zanini, A., 2015. Contaminant release history identification in 2-d heterogeneous aquifers through a minimum relative entropy approach. SpringerPlus 4. doi:10.1186/s40064-015-1465-x.

661 Datta, B., Beegle, J.E., Kavvas, M.L., Orlob, G.T., 1989. Development of
662 an expert-system embedding pattern-recognition techniques for pollution-
663 source identification. Report for 30 September 1987-29 November 1989.

664 Emerick, A.A., Reynolds, A.C., 2012. History matching time-lapse seismic
665 data using the ensemble kalman filter with multiple data assimilations.
666 Computational Geosciences 16, 639–659. doi:10.1007/s10596-012-9275-5.

667 Emerick, A.A., Reynolds, A.C., 2013a. Ensemble smoother with
668 multiple data assimilation. Computers & Geosciences 55, 3–15.
669 doi:10.1016/j.cageo.2012.03.011.

670 Emerick, A.A., Reynolds, A.C., 2013b. History-matching production and
671 seismic data in a real field case using the ensemble smoother with multiple
672 data assimilation, in: SPE Reservoir Simulation Symposium, Society of
673 Petroleum Engineers. doi:10.2118/163675-ms.

674 Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-
675 geostrophic model using monte carlo methods to forecast error statistics.
676 Journal of Geophysical Research 99, 10143. doi:10.1029/94jc00572.

677 Evensen, G., 2018. Analysis of iterative ensemble smoothers for
678 solving inverse problems. Computational Geosciences 22, 885–908.
679 doi:10.1007/s10596-018-9731-y.

680 Fokker, P., Wassing, B., van Leijen, F., Hanssen, R., Nieuwland, D., 2016.
681 Application of an ensemble smoother with multiple data assimilation to
682 the bergermeer gas field, using PS-InSAR. Geomechanics for Energy and
683 the Environment 5, 16–28. doi:10.1016/j.gete.2015.11.003.

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. doi:10.1002/qj.49712555417.

Gzyl, G., Zanini, A., Frączek, R., Kura, K., 2014. Contaminant source and release history identification in groundwater: A multi-step approach. Journal of Contaminant Hydrology 157, 59–72. doi:10.1016/j.jconhyd.2013.11.006.

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. doi:10.1175/1520-0493(2001)129<2776:ddfobe>2.0.co;2.

Harbaugh, A.W., 2005. MODFLOW-2005: the u.s. geological survey modular ground-water model–the ground-water flow process. doi:10.3133/tm6a16.

Houtekamer, P.L., Mitchell, H.L., 1998. Data assimilation using an ensemble kalman filter technique. Monthly Weather Review 126, 796–811. doi:10.1175/1520-0493(1998)126<0796:dauaek>2.0.co;2.

Kang, X., Shi, X., Revil, A., Cao, Z., Li, L., Lan, T., Wu, J., 2019. Coupled hydrogeophysical inversion to identify non-gaussian hydraulic conductivity field by jointly assimilating geochemical and time-lapse geophysical data. Journal of Hydrology 578, 124092. doi:10.1016/j.jhydrol.2019.124092.

Lan, T., Shi, X., Jiang, B., Sun, Y., Wu, J., 2018. Joint inversion of physical and geochemical parameters in groundwater models by sequential ensemble-based optimal design. Stochastic Environmental Research and Risk Assessment 32, 1919–1937. doi:10.1007/s00477-018-1521-5.

van Leeuwen, P.J., Evensen, G., 1996. Data assimilation and inverse methods in terms of a probabilistic formulation. Monthly Weather Review 124, 2898–2913. doi:10.1175/1520-0493(1996)124<2898:daaimi>2.0.co;2.

Li, H., Kalnay, E., Miyoshi, T., 2009. Simultaneous estimation of covariance inflation and observation errors within an ensemble kalman filter. Quarterly Journal of the Royal Meteorological Society 135, 523–533. doi:10.1002/qj.371.

Li, L., Bao, J., Cao, Z., Cui, F., 2019. Soil hydraulic parameters estimation using gpr data via es-mda. AGUFM 2019, H43F–2047.

Li, L., Stetler, L., Cao, Z., Davis, A., 2018. An iterative normal-score ensemble smoother for dealing with non-gaussianity in data assimilation. Journal of Hydrology 567, 759–766.

Liang, X., Zheng, X., Zhang, S., Wu, G., Dai, Y., Li, Y., 2011. Maximum likelihood estimation of inflation factors on error covariance matrices for ensemble kalman filter assimilation. Quarterly Journal of the Royal Meteorological Society 138, 263–273. doi:10.1002/qj.912.

Mahar, P.S., Datta, B., 1997. Optimal monitoring network and ground-water–pollution source identification. Journal of Water Resources Planning and Management 123, 199–207. doi:10.1061/(asce)0733-9496(1997)123:4(199).

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. doi:10.1080/00221680409500042.

730 Michalak, A.M., Kitanidis, P.K., 2004b. Estimation of historical ground-
731 water contaminant distribution using the adjoint state method ap-
732 plied to geostatistical inverse modeling. Water Resources Research 40.
733 doi:10.1029/2004wr003214.

734 Moriasi, D.N., Arnold, J.G., Liew, M.W.V., Bingner, R.L., Harmel, R.D.,
735 Veith, T.L., 2007. Model evaluation guidelines for systematic quantification
736 of accuracy in watershed simulations. Transactions of the ASABE 50, 885–
737 900.

738 Neupauer, R.M., Borchers, B., Wilson, J.L., 2000. Comparison of in-
739 verse methods for reconstructing the release history of a groundwa-
740 ter contamination source. Water Resources Research 36, 2469–2475.
741 doi:10.1029/2000wr900176.

742 Pirot, G., Krityakierne, T., Ginsbourger, D., Renard, P., 2019. Contaminant
743 source localization via bayesian global optimization. Hydrology and Earth
744 System Sciences 23, 351–369. doi:10.5194/hess-23-351-2019.

745 Ritter, A., Muñoz-Carpena, R., 2013. Performance evaluation of hy-
746 drological models: Statistical significance for reducing subjectivity
747 in goodness-of-fit assessments. Journal of Hydrology 480, 33–45.
748 doi:10.1016/j.jhydrol.2012.12.004.

749 Skaggs, T.H., Kabala, Z.J., 1994. Recovering the release history of
750 a groundwater contaminant. Water Resources Research 30, 71–79.
751 doi:10.1029/93wr02656.

Snodgrass, M.F., Kitanidis, P.K., 1997. A geostatistical approach to contaminant source identification. Water Resources Research 33, 537–546. doi:10.1029/96wr03753.

Song, X., Chen, X., Ye, M., Dai, Z., Hammond, G., Zachara, J.M., 2019. Delineating facies spatial distribution by integrating ensemble data assimilationand indicator geostatistics with level-set transformation. Water Resources Research 55, 2652–2671. doi:10.1029/2018wr023262.

Sun, A.Y., Painter, S.L., Wittmeyer, G.W., 2006. A robust approach for iterative contaminant source location and release history recovery. Journal of Contaminant Hydrology 88, 181–196. doi:10.1016/j.jconhyd.2006.06.006.

Todaro, V., D'Oria, M., Tanda, M.G., Gómez-Hernández, J.J., 2019. Ensemble smoother with multiple data assimilation for reverse flow routing. Computers & Geosciences 131, 32–40. doi:10.1016/j.cageo.2019.06.002.

Uribe-Asarta, J., 2019. Modelación numérica de un experimento de transporte de masa en un tanque de arena de laboratorio. Master's thesis. Universitat Politècnica de València.

Wang, X., Bishop, C.H., 2003. A comparison of breeding and ensemble transform kalman filter ensemble forecast schemes. Journal of the Atmospheric Sciences 60, 1140–1158. doi:10.1175/1520-0469(2003)060<1140:acobae>2.0.co;2.

Woodbury, A., Sudicky, E., Ulrych, T.J., Ludwig, R., 1998. Three-dimensional plume source reconstruction using minimum relative en-

tropy inversion. Journal of Contaminant Hydrology 32, 131–158. doi:10.1016/s0169-7722(97)00088-0.

Woodbury, A.D., Ulrych, T.J., 1996. Minimum relative entropy inversion: Theory and application to recovering the release history of a groundwater contaminant. Water Resources Research 32, 2671–2681. doi:10.1029/95wr03818.

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. doi:10.1002/2016wr019111.

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. doi:10.1016/j.advwatres.2017.12.011.

Xu, T., Gómez-Hernández, J.J., Chen, Z., Lu, C., 2020. 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 , 125681doi:10.1016/j.jhydrol.2020.125681.

Zanini, A., Woodbury, A.D., 2016. Contaminant source reconstruction by empirical bayes and akaikes bayesian information criterion. Journal of Contaminant Hydrology 185-186, 74–86. doi:10.1016/j.jconhyd.2016.01.006.

Zhao, Y., Forouzanfar, F., Reynolds, A.C., 2016. History matching of multi-facies channelized reservoirs using ES-MDA with common basis DCT.

797 Computational Geosciences 21, 1343–1364. doi:10.1007/s10596-016-9604-
798 1.

799 Zheng, C., Wang, P.P., 1999. MT3DMS : a modular three-dimensional multi-
800 species transport model for simulation of advection, dispersion, and chemi-
801 cal reactions of contaminants in groundwater systems; documentation and
802 user's guide.

803 Zheng, X., 2009. An adaptive estimation of forecast error covariance pa-
804 rameters for kalman filtering data assimilation. Advances in Atmospheric
805 Sciences 26, 154–160. doi:10.1007/s00376-009-0154-5.

819 Zhao, Y., Forouzanfar, F., Reynolds, A.C., 2016. History matching of multi-
820 facies channelized reservoirs using ES-MDA with common basis DCT.
821 Computational Geosciences 21, 1343–1364. doi:10.1007/s10596-016-9604-
822 1.

823 Zheng, C., Wang, P.P., 1999. MT3DMS : a modular three-dimensional multi-
824 species transport model for simulation of advection, dispersion, and chemi-
825 cal reactions of contaminants in groundwater systems; documentation and
826 user's guide.

827 Zheng, X., 2009. An adaptive estimation of forecast error covariance pa-
828 rameters for kalman filtering data assimilation. Advances in Atmospheric
829 Sciences 26, 154–160. doi:10.1007/s00376-009-0154-5.