



UNIVERSITÀ DI PARMA

ARCHIVIO DELLA RICERCA

University of Parma Research Repository

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

This is the peer reviewed version of the following article:

Original

Ensemble smoother with multiple data assimilation to simultaneously estimate the source location and the release history of a contaminant spill in an aquifer / Todaro, V.; D'Oria, M.; Tanda, M. G.; Gomez-Hernandez, J. J.. - In: JOURNAL OF HYDROLOGY. - ISSN 0022-1694. - 598:(2021), p. 126215. [10.1016/j.jhydrol.2021.126215]

Availability:

This version is available at: 11381/2893620 since: 2022-09-27T13:53:28Z

Publisher:

Elsevier B.V.

Published

DOI:10.1016/j.jhydrol.2021.126215

Terms of use:

Anyone can freely access the full text of works made available as "Open Access". Works made available

Publisher copyright

note finali coverpage

(Article begins on next page)

02 May 2026

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 Politècnica 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;

26 Neupauer et al., 2000; Butera and Tanda, 2003; Butera et al., 2006, 2012;
27 Gzyl et al., 2014; Cupola et al., 2015); empirical Bayesian methods combined
28 with Akaike’s Bayesian Information Criterion (Zanini and Woodbury, 2016);
29 Bayesian global optimization (Piroot et al., 2019) and ensemble Kalman filter
30 methods (Xu and Gómez-Hernández, 2016, 2018; Chen et al., 2018; Xu et al.,
31 2020).

32 However, only a few of the presented studies allow to simultaneously
33 identify the source location and the release history of a groundwater contam-
34 inant. The method proposed by Aral et al. (2001) used a progressive genetic
35 algorithm to solve an iterative nonlinear optimization problem, in which the
36 source location and release history were explicitly defined as continuous un-
37 known variables and contaminant concentrations were used as observations.
38 Sun et al. (2006) combined a constrained robust least squares estimator with
39 a global optimization solver for iteratively identifying release histories and
40 source locations on the basis of concentration measurements. Ayvaz (2010)
41 used an optimization method based on the heuristic harmony search algo-
42 rithm to identify locations and release histories for pollution sources, mini-
43 mizing residuals between the simulated and measured contaminant concen-
44 trations. All these methods are deterministic and do not allow to quantify
45 the uncertainty of the results.

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

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

56 We propose a new procedure for the joint identification of the source
57 location and the release history of a pollutant in an aquifer: the use of
58 an Ensemble Smoother with Multiple Data Assimilation (ES-MDA) in the
59 context of contaminant source identification. The ES-MDA, introduced by
60 Emerick and Reynolds (2012, 2013a), has been mainly applied to reservoir
61 history matching problems (Emerick and Reynolds, 2013b; Fokker et al.,
62 2016; Zhao et al., 2016), but its popularity is growing also in hydrology
63 (Lan et al., 2018; Li et al., 2018, 2019; Kang et al., 2019; Song et al., 2019;
64 Todaro et al., 2019; Bao et al., 2020). It is an iterative data assimilation
65 method based on the Ensemble Kalman Filter (EnKF), initially proposed
66 by Evensen (1994). In particular, the ES-MDA is a variant of the Ensemble
67 Smoother (ES) proposed by van Leeuwen and Evensen (1996). Unlike the
68 EnKF, which performs a sequential update one step at a time assimilating the
69 data as they are collected, the ES and the ES-MDA simultaneously assimilate
70 all the available observation data. Also, the ES-MDA iteratively assimilates
71 the same data multiple times leading to better results for strongly nonlinear
72 problems than the ES, which performs a single global update (Evensen, 2018).

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

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

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

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

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

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

106 The study also addresses the problem of undersampling present in
107 ensemble-based methods; it occurs when the ensemble size is so small that
108 it is not statistically representative of the variability of the unknowns. Al-
109 though large ensembles mitigate this problem, the computational cost in-
110 creases with the ensemble size; therefore, it is advantageous to solve the prob-
111 lem with the smallest possible ensemble. Covariance localization has been
112 developed to overcome this problem; it helps in removing long-range spuri-
113 ous correlations and mitigates the ensemble rank deficiency, allowing the use
114 of a small number of realizations. Localization can be achieved by different
115 ways (Houtekamer and Mitchell, 1998; Hamill et al., 2001; Anderson, 2007b;
116 Chen and Oliver, 2009). Covariance localization is generally based on the
117 spatial distance between parameter locations and observations; in this study,
118 parameters and observations are also time-dependent, furthermore the dis-
119 tance between them is not fixed since the source position is unknown, what
120 complicates the use of standard localization techniques. Todaro et al. (2019)
121 proposed a temporal localization considering time lapses rather than spatial
122 distances. A new localization approach is presented, which takes into account
123 both spatial and temporal distances and iteratively updates the distance be-
124 tween the unknown parameters and the observations. Covariance inflation
125 is also considered to overcome undersampling problems (Anderson and An-

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

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

134 2. Methods

135 2.1. Forward problem

136 The forward problem is based on the groundwater flow and mass trans-
 137 port equations. In particular, we consider an incompressible fluid in satu-
 138 rated porous media and a non-reactive contaminant injected in the aquifer
 139 at a point subject to advection and dispersion (Bear, 1972; Bear and Ver-
 140 ruijt, 1987). Assuming a uniform porosity, initial condition $C(\mathbf{x}, 0) = 0$, and
 141 boundary condition, $C(\infty, t) = 0$, where $C(\mathbf{x}, t)$ [ML⁻³] is the solute concen-
 142 tration, the transport equation can be solved by the convolution integral

$$C(\mathbf{x}, t) = \int_0^t s(\mathbf{x}_0, \tau) g(\mathbf{x}, t - \tau) d\tau. \quad (1)$$

143 The term $s(\mathbf{x}_0, t)$ [MT⁻¹] is the the contaminant flux injected into the aquifer
 144 through the source located at \mathbf{x}_0 given by

$$s(\mathbf{x}_0, t) = C_0(t) \cdot q_0(\mathbf{x}_0, t), \quad (2)$$

145 where $C_0(t)$ [ML⁻³] is the concentration of the released pollutant at time t
 146 and $q_0(\mathbf{x}_0, t)$ [L³T⁻¹] is the injection flow rate. The term $g(\mathbf{x}, t - \tau)$ is a

147 Kernel function that represents the response at location \mathbf{x} and time t to a
 148 pulse injection at the source location \mathbf{x}_0 and time τ .

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

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

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

161 2.2. Ensemble smoother with multiple data assimilation

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

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

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

185 1. Initialization step.

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

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

198 2. Forecast step.

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

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

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

206 3. Update step.

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

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

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

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

$$\sum_{i=1}^N \frac{1}{\alpha_i} = 1, \quad (6)$$

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

$$\mathbf{C}_{\mathbf{XY}}^i = \frac{1}{N_e - 1} \sum_{j=1}^{N_e} (\mathbf{X}_{j,i} - \bar{\mathbf{X}}_i) (\mathbf{Y}_{j,i} - \bar{\mathbf{Y}}_i)^T, \quad (7)$$

222

$$\mathbf{C}_{\mathbf{YY}}^i = \frac{1}{N_e - 1} \sum_{j=1}^{N_e} (\mathbf{Y}_{j,i} - \bar{\mathbf{Y}}_i) (\mathbf{Y}_{j,i} - \bar{\mathbf{Y}}_i)^T, \quad (8)$$

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

$$\tilde{\mathbf{C}}_{\mathbf{XY}}^i = \rho_{\mathbf{XY}}^i \circ \mathbf{C}_{\mathbf{XY}}^i, \quad (9)$$

227

$$\tilde{\mathbf{C}}_{\mathbf{YY}}^i = \rho_{\mathbf{YY}} \circ \mathbf{C}_{\mathbf{YY}}^i. \quad (10)$$

228 where \circ represents the elementwise multiplication and $\rho_{\mathbf{XY}}^i$ and $\rho_{\mathbf{YY}}$
 229 are correlation matrices based on spatial and temporal distances be-
 230 tween parameters and observations and between observations and ob-
 231 servations, respectively. The correlations in space ($\rho_{\mathbf{XY},s}^i$, $\rho_{\mathbf{YY},s}$) and

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

$$234 \quad \rho_{\mathbf{XY}}^i = \rho_{\mathbf{XY},s}^i \circ \rho_{\mathbf{XY},t}, \quad (11)$$

$$\rho_{\mathbf{YY}} = \rho_{\mathbf{YY},s} \circ \rho_{\mathbf{YY},t}. \quad (12)$$

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

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

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

247 At the end of each update step, linear relaxation and covariance infla-
 248 tion are used to prevent smoother divergence. Linear relaxation reduces
 249 the magnitude of the update at the end of an iteration. When linear

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

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

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

$$\tilde{\mathbf{X}}_{j,i+1} = r (\mathbf{X}_{j,i+1} - \bar{\mathbf{X}}_{i+1}) + \bar{\mathbf{X}}_{i+1}. \quad (15)$$

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

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

261 3. Case studies

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

271 *3.1. Analytical case*

272 The analytical case simulates a pollution event in an infinite homogeneous
 273 two-dimensional aquifer, with uniform flow, as result of the injection of a
 274 nonreactive contaminant at a point (Butera and Tanda, 2003). It is assumed
 275 that the water discharge $q_0(\mathbf{x}_0, t)$ is of unit value and small enough such
 276 that it does not affect the uniform groundwater flow. Therefore, the release
 277 history $s(\mathbf{x}_0, t)$, defined in Eq. (2), is equivalent to the concentration history
 278 $C_0(t)$. All quantities are considered with unspecified but consistent units.
 279 The uniform velocity and the dispersion coefficients are assumed known:
 280 $v = 1$, $D_x = 1$ and $D_y = 0.1$. We use the same expression for the release
 281 function $s_r(\mathbf{x}_0, t)$ used elsewhere (Skaggs and Kabala, 1994; Woodbury and
 282 Ulrych, 1996; Snodgrass and Kitanidis, 1997; Butera and Tanda, 2003; Butera
 283 et al., 2012; Zanini and Woodbury, 2016) to define the reference solution

$$\begin{aligned}
 s_r(\mathbf{x}_0, t) = & \exp\left(-\frac{(t-130)^2}{50}\right) + 0.3 \exp\left(-\frac{(t-150)^2}{200}\right) \\
 & + 0.5 \exp\left(-\frac{(t-190)^2}{98}\right).
 \end{aligned}
 \tag{16}$$

284 The actual source location \mathbf{x}_0 is $x_0 = 50$ and $y_0 = 20$. The concentration
 285 history has a total duration of 300; it is discretized into 101 intervals with
 286 a time step of $\Delta t = 3$ resulting in a total number of parameters to be esti-
 287 mated $N_p = 103$ (the two spatial coordinates plus the 101 temporal solute
 288 fluxes). The reference release function, depicted in Fig. 1, is used to obtain
 289 the reference observations, which are computed by evaluating Eq. (3) using
 290 numerical integration.

291 Different test cases are carried out to investigate the impact of the obser-
 292 vation sampling scheme, ensemble size, covariance localization and inflation

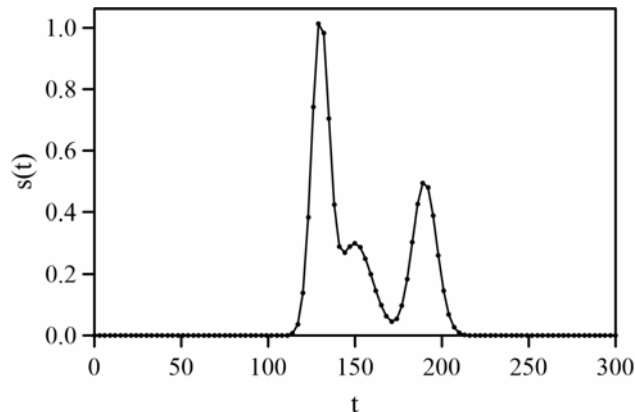


Figure 1: Analytical case: reference release history.

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

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

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

302 where $N_p - 2$ is equal to 101, the number of intervals used to discretize
 303 $s(t)$; $s_{r,i}$ represents the discretized source function and is the i -th actual
 304 amount of released contaminant, \bar{s}_r is the time average of the reference
 305 release history ($\frac{1}{N_p-2} \sum_{i=1}^{N_p-2} s_{r,i}$) and \bar{X}_i is the ensemble mean of the
 306 i -th estimated amount of released contaminant ($\frac{1}{N_e} \sum_{j=1}^{N_e} X_i^j$, with X_i^j

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

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

$$RMSE = \sqrt{\frac{\sum_{i=1}^m (D_i - \bar{Y}_i)^2}{m}} \quad (18)$$

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

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

$$L = \sqrt{(\bar{x}_s - x_0)^2 + (\bar{y}_s - y_0)^2} \quad (19)$$

316 where \bar{x}_s and \bar{y}_s are the ensemble means of the estimated spatial coor-
 317 dinates of the source and (x_0, y_0) is the true source location. The closer
 318 to zero, the better.

319 These metrics are compared with reference threshold values to evaluate
 320 the performance of the method. We consider three cases: i) good perfor-
 321 mance when the reproduction of the observed concentrations is good, the
 322 identification of the source location is good and the identification of the re-
 323 lease function is good; ii) equifinality performance, when reproduction of
 324 the observed concentrations is good, but neither the source location nor the
 325 release function are well identified; iii) poor performance, otherwise:

326 i) Good performance when

327 $RMSE < RMSE_{thr}$ 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

328 ii) Equifinality performance when

329 $RMSE < RMSE_{thr}$ and ($NSE < NSE_{thr2}$ or $L > L_{thr}$)

330 iii) Otherwise, fail.

331 There is not a standard criterion for the definition of metric thresholds to as-
 332 sess goodness-of-fit (see e.g. Moriasi et al., 2007; Ritter and Muñoz-Carpena,
 333 2013). In this study, we consider the performance of the method to be good if
 334 $NSE > 0.7$ and unsatisfactory if $NSE < 0.6$. The fit between predictions and
 335 observations is considered to be good when the $RMSE$ is less than the max-
 336 imum assumed error. Since the observation errors are normally distributed,
 337 the maximum error is defined as 4σ , where σ is its standard deviation. The
 338 selected threshold values ($RMSE_{thr}$, NSE_{thr1} , NSE_{thr2} , L_{thr}) are summa-
 339 rized in Table 1. With these criteria, it is possible to define the percentage of
 340 successful tests, tests with multiple solutions and failed tests for each case,
 341 on the basis of the 100 experiments.

342 *3.1.1. Impact of the observation network geometry and sampling frequency*

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

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

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

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

355 C. Concentrations collected at four monitoring points distributed on the
356 same line of the source ($y = 20$) at x -coordinates 80, 115, 150 and
357 185, and the same 31 sampling times of set A. The total number of
358 observations is $m = 4 \cdot 31 = 124$.

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

363 A random observation error ε normally distributed with zero mean and
364 variance $5 \cdot 10^{-8}$ for all the performed tests is considered. The initial ensemble
365 of parameters is composed of 1000 realizations. The realizations of the source
366 coordinates are uniformly distributed random values selected in the range $[5,$
367 $80]$ for x and $[10, 30]$ for y . The realizations of the release history are normal

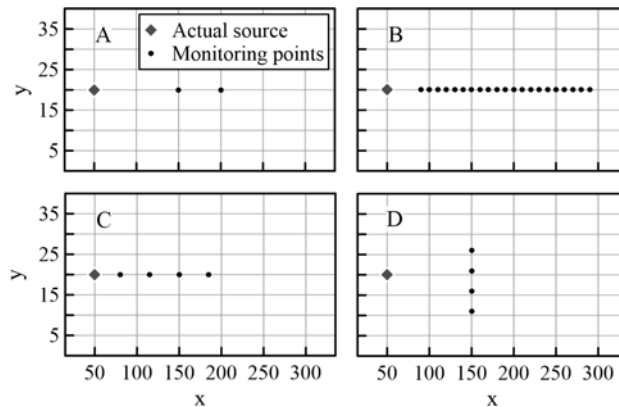


Figure 2: Analytical case: location of the measurement points for sets A, B, C and D; the red diamond is the actual source location.

368 functions described by the following expression:

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

369 where t is the time, Δ is a base amount of released concentration, Γ is
 370 the volume under the Gaussian function of mean μ and variance σ^2 . These
 371 coefficients are selected randomly from uniform distributions, $\Delta \in U[1 \cdot 10^{-10},$
 372 $1 \cdot 10^{-3}]$, $\Gamma \in U[10, 40]$, $\mu \in U[89, 210]$ and $\sigma \in U[6, 59]$. The ES-MDA is run
 373 with 10 iterations and a decreasing series of α values following the sequence
 374 $[113.33; 75.55; 50.37; 33.58; 22.39; 14.92; 9.95; 6.63; 4.42; 2.95]$.

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

378 The observation network geometry greatly impacts the final results. The
 379 synthetic experiments that give reliable solutions ($NSE > 70$ and $L < 5$) are
 380 less than 21% for observation sets A, B and C. Furthermore, equifinality
 381 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%

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

384 *3.1.2. Impact of the ensemble size and application of localization and infla-*
385 *tion techniques*

386 The test cases designed to investigate the impact of the ensemble size,
387 covariance localization and inflation techniques make use of the observation
388 set D. We tested five ensemble sizes N_e of 1000, 500, 250, 100 and 50 with
389 and without covariance corrections. The number of iterations, α values,
390 and distributions used to generate the initial ensembles are the same ones
391 used in the previous section. Covariance localization is applied using the
392 coefficients b_s equal to 210 and b_t equal to 300. The factor r used for the
393 covariance inflation is equal to 1.01. The results obtained from each set of
394 100 synthetic experiments are reported in Table 3. The ES-MDA performs
395 better for increasing ensemble sizes and when covariance inflation and lo-
396 calization techniques are applied. The percentage of successful tests is high
397 for large ensembles, with even better numbers when covariance corrections
398 are applied. The presence of equifinality is detected when the ensemble size
399 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%

400 covariance and inflation techniques are more evident for small ensemble sizes;
401 considering N_e equal to 100, the percentage of successful tests is 46% for the
402 experiments without corrections and 64% for those with corrections; multiple
403 solutions are detected for 43% of the experiments without corrections and for
404 14% of those with corrections. The tests computed with the smaller ensemble
405 size ($N_e=50$) lead to unsatisfactory results with a percentage of successful
406 tests lower than 45% and a high probability of equifinality.

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

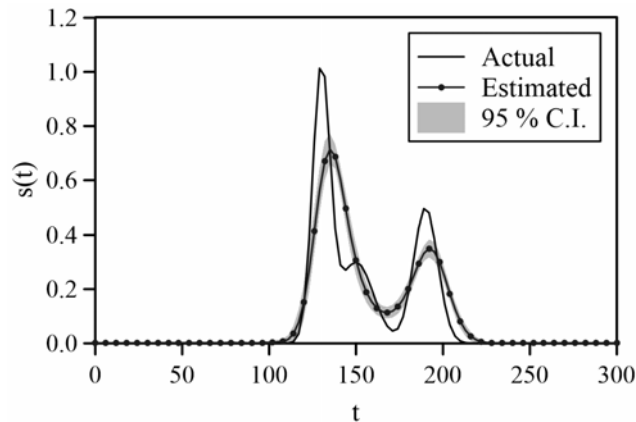


Figure 3: Analytical case: actual and estimated release history with 95% uncertainty interval resulting from a test performed with $N_e = 100$ and observation set D.

411 successful tests, and we use the set of successful tests to build uncertainty
 412 intervals about the median. In Fig. 3 the reference solution and the ensem-
 413 ble median with its 95% uncertainty interval are depicted. Figure 4 shows
 414 a comparison between observed and predicted concentrations at observation
 415 locations. The ES-MDA reproduces quite well the release history and the
 416 source location estimate is very close to the true one ($x_0=50$, $y_0=20$). The
 417 NSE is 80.46% and the ensemble means of x and y coordinates are, respec-
 418 tively, equal to 52.66 (± 1.78 , 95% uncertainty interval) and 20.00 (± 0.06 ,
 419 95% uncertainty interval). The test leads to a good match between observa-
 420 tions and predictions with an RMSE at the last iteration equal to $3.3 \cdot 10^{-4}$
 421 and a narrow 95% uncertainty interval.

422 3.2. Experimental case

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

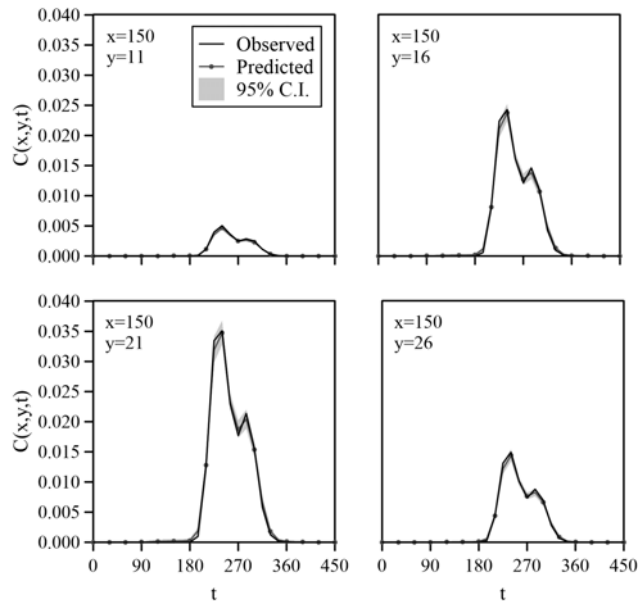


Figure 4: Analytical case: observed and predicted concentrations with 95% uncertainty interval.

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

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

448 The inverse methodology requires a calibrated numerical model able to
449 describe as accurately as possible the forward process. Groundwater flow
450 was modeled with MODFLOW 2005 (Harbaugh, 2005) and mass transport
451 with MT3DMS (Zheng and Wang, 1999). The effect of the injection on the
452 background flow is not negligible; therefore, a transient flow model is con-
453 sidered. The numerical model was preliminarily calibrated by an inverse
454 procedure not reported here for brevity. After the calibration, and for the
455 purposes of the source identification, this model is used throughout. Table
456 4 summarizes the parameters of the flow and transport models and Figure
457 5 shows the hydraulic conductivity field after the calibration process. The
458 estimated field is slightly heterogeneous and conductivity is anisotropic, even
459 though the sandbox was filled with glass beads of almost the same size with
460 the intention of reproducing an isotropic homogeneous field. Our interpreta-
461 tion 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

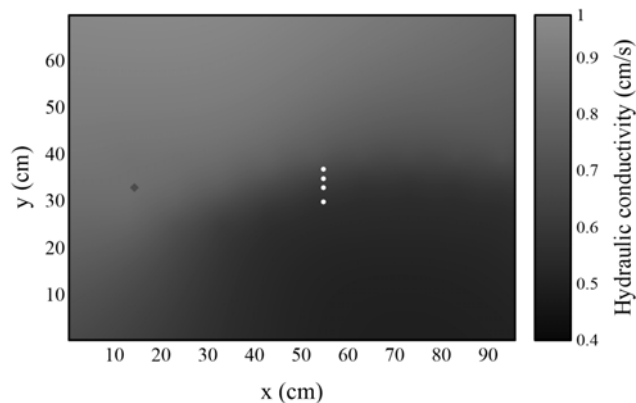


Figure 5: Hydraulic conductivity field. The red diamonds denotes the actual source location. The white dots are the monitoring points.

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

463 Since the concentration of the contaminant is known, the estimation of
464 the release history is limited to identifying the injected flow rate. The release
465 duration is discretized into 72 intervals with a time step of $\Delta t = 3$ s result-
466 ing in a total number of parameters $N_p = 74$, of which two are the spatial
467 coordinates of the source. The initial ensemble of parameters is made up of
468 81 realizations ($N_e = 81$); the spatial coordinates of the source are random
469 values selected from uniform distributions $x \in U[5, 30]$ cm, and $y \in U[30, 34]$

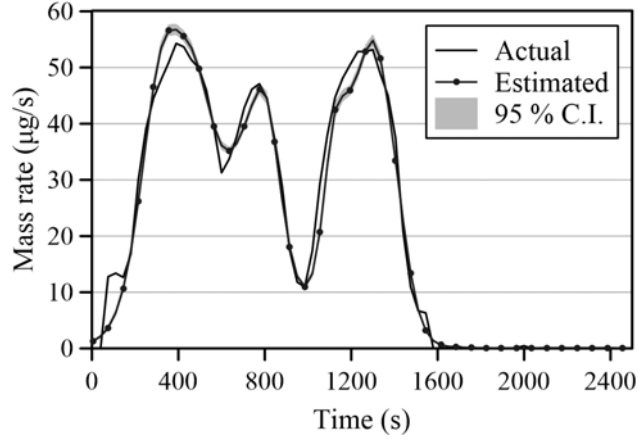


Figure 6: Experimental case: actual and estimated release history with 95% uncertainty interval. Time 0 s represents the time at which injection starts.

470 cm. The initial realizations of the injected flow rate history follow expres-
471 sion Eq. (20), with parameters selected randomly from the following uniform
472 distributions, $\Delta \in \mathcal{U}[1 \cdot 10^{-10}, 1 \cdot 10^{-1}]$, $\Gamma \in U[800, 1000]$, $\mu \in U[490, 1400]$
473 and $\sigma \in U[60, 365]$. The four monitoring points are vertically distributed on
474 the line $x = 54.75$ cm and at y-coordinates 29.00, 32.75, 34.75 and 36.75 cm.
475 For each monitoring point, the observed concentrations are recorded at 45
476 sampling times from $T = 0$ s to $T = 2200$ s (total number of monitoring data
477 is $m = 180$). The random measurement error ε is assumed normally dis-
478 tributed with zero mean and variance $1 \cdot 10^{-2} (\text{mg} \cdot \text{l}^{-1})^2$. The ES-MDA with
479 6 iterations and decreasing $\alpha = [63.0; 31.5; 15.8; 7.88 \ 3.9; 2.0]$ is used for the
480 inversion. Covariance localization and covariance inflation are applied using
481 the coefficients $b_s = 200$, $b_t = 2500$ and $r = 1.01$, and linear relaxation with the
482 coefficient $w = 0.1$.

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

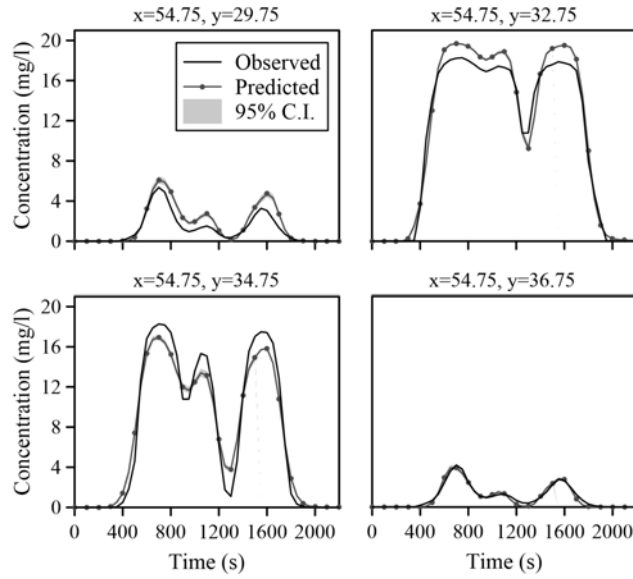


Figure 7: Experimental case: observed and predicted concentrations with 95% uncertainty intervals. Time 0 s represents the time at which injection starts.

484 the release history with its 95% confidence interval and the true solution are
 485 depicted. The ES-MDA leads to a good agreement between the two curves
 486 with an NSE value equal to 98.34% and with a satisfactory representation of
 487 peak magnitudes and times. The ensemble means of the x and y coordinates
 488 of the source are, respectively, equal to 14.71 cm (± 0.45 cm, 95% uncertainty
 489 interval) and 32.91 cm (± 0.14 cm, 95% uncertainty interval); the distance
 490 between the true and estimated source location is less than 0.5 cm. In Fig.
 491 7, the experimental and predicted observations are compared. The retrieved
 492 source parameters reproduce quite well the observed concentrations with a
 493 narrow 95% uncertainty interval; the RMSE at the last iteration is equal to
 494 $0.96 \text{ mg}\cdot\text{l}^{-1}$, which is comparable with the experimental observation errors.

495 4. Discussion and Conclusions

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

600 **Acknowledgements**

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

606 **References**

- 607 Alapati, S., Kabala, Z.J., 2000. Recovering the release his-
608 tory of a groundwater contaminant using a non-linear least-squares
609 method. *Hydrological Processes* 14, 1003–1016. doi:10.1002/(SICI)1099-
610 1085(20000430)14:6<1003::AID-HYP981>3.0.CO;2-W.
- 611 Anderson, J.L., 2007a. An adaptive covariance inflation error correction al-
612 gorithm for ensemble filters. *Tellus A: Dynamic Meteorology and Oceanog-*
613 *raphy* 59, 210–224. doi:10.1111/j.1600-0870.2006.00216.x.
- 614 Anderson, J.L., 2007b. Exploring the need for localization in ensemble data
615 assimilation using a hierarchical ensemble filter. *Physica D: Nonlinear*
616 *Phenomena* 230, 99–111. doi:10.1016/j.physd.2006.02.011.

- 617 Anderson, J.L., Anderson, S.L., 1999. A monte carlo implementation of
618 the nonlinear filtering problem to produce ensemble assimilations and
619 forecasts. *Monthly Weather Review* 127, 2741–2758. doi:10.1175/1520-
620 0493(1999)127<2741:amciot>2.0.co;2.
- 621 Aral, M.M., Guan, J., Maslia, M.L., 2001. Identification of contaminant
622 source location and release history in aquifers. *Journal of Hydrologic En-*
623 *gineering* 6, 225–234. doi:10.1061/(asce)1084-0699(2001)6:3(225).
- 624 Ayvaz, M.T., 2010. A linked simulation–optimization model for solving the
625 unknown groundwater pollution source identification problems. *Journal of*
626 *Contaminant Hydrology* 117, 46–59. doi:10.1016/j.jconhyd.2010.06.004.
- 627 Bao, J., Li, L., Redoloza, F., 2020. Coupling ensemble smoother and deep
628 learning with generative adversarial networks to deal with non-gaussianity
629 in flow and transport data assimilation. *Journal of Hydrology* 590, 125443.
- 630 Bear, J., 1972. *Dynamics of fluids in porous media*. American Elsevier
631 Publishing Company, New York.
- 632 Bear, J., Verruijt, A., 1987. *Modeling groundwater flow and pollution*. vol-
633 *ume 2*. Springer Science & Business Media.
- 634 Butera, I., Tanda, M.G., 2003. A geostatistical approach to recover the
635 release history of groundwater pollutants. *Water Resources Research* 39.
636 doi:10.1029/2003WR002314.
- 637 Butera, I., Tanda, M.G., Zanini, A., 2006. Use of numerical modelling to
638 identify the transfer function and application to the geostatistical proce-

- 639 dure in the solution of inverse problems in groundwater. *Journal of Inverse*
640 and Ill-posed Problems 14, 547–572. doi:10.1163/156939406778474532.
- 641 Butera, I., Tanda, M.G., Zanini, A., 2012. Simultaneous identification of the
642 pollutant release history and the source location in groundwater by means
643 of a geostatistical approach. *Stochastic Environmental Research and Risk*
644 Assessment 27, 1269–1280. doi:10.1007/s00477-012-0662-1.
- 645 Chen, Y., Oliver, D.S., 2009. Cross-covariances and localization for EnKF
646 in multiphase flow data assimilation. *Computational Geosciences* 14, 579–
647 601. doi:10.1007/s10596-009-9174-6.
- 648 Chen, Z., Gómez-Hernández, J.J., Xu, T., Zanini, A., 2018. Joint identifica-
649 tion of contaminant source and aquifer geometry in a sandbox experiment
650 with the restart ensemble kalman filter. *Journal of Hydrology* 564, 1074–
651 1084. doi:10.1016/j.jhydrol.2018.07.073.
- 652 Citarella, D., Cupola, F., Tanda, M.G., Zanini, A., 2015. Evaluation of
653 dispersivity coefficients by means of a laboratory image analysis. *Journal*
654 of Contaminant Hydrology 172, 10–23. doi:10.1016/j.jconhyd.2014.11.001.
- 655 Cupola, F., Tanda, M.G., Zanini, A., 2014. Laboratory sandbox validation
656 of pollutant source location methods. *Stochastic Environmental Research*
657 and Risk Assessment 29, 169–182. doi:10.1007/s00477-014-0869-4.
- 658 Cupola, F., Tanda, M.G., Zanini, A., 2015. Contaminant release history
659 identification in 2-d heterogeneous aquifers through a minimum relative
660 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.

- 684 Gaspari, G., Cohn, S.E., 1999. Construction of correlation functions in two
685 and three dimensions. *Quarterly Journal of the Royal Meteorological So-*
686 *ciety* 125, 723–757. doi:10.1002/qj.49712555417.
- 687 Gzyl, G., Zanini, A., Frączek, R., Kura, K., 2014. Contaminant source and re-
688 lease history identification in groundwater: A multi-step approach. *Journal*
689 *of Contaminant Hydrology* 157, 59–72. doi:10.1016/j.jconhyd.2013.11.006.
- 690 Hamill, T.M., Whitaker, J.S., Snyder, C., 2001. Distance-dependent fil-
691 tering of background error covariance estimates in an ensemble kalman
692 filter. *Monthly Weather Review* 129, 2776–2790. doi:10.1175/1520-
693 0493(2001)129<2776:ddfobe>2.0.co;2.
- 694 Harbaugh, A.W., 2005. MODFLOW-2005: the u.s. geological survey modular
695 ground-water model—the ground-water flow process. doi:10.3133/tm6a16.
- 696 Houtekamer, P.L., Mitchell, H.L., 1998. Data assimilation using an en-
697 semble kalman filter technique. *Monthly Weather Review* 126, 796–811.
698 doi:10.1175/1520-0493(1998)126<0796:dauaek>2.0.co;2.
- 699 Kang, X., Shi, X., Revil, A., Cao, Z., Li, L., Lan, T., Wu, J., 2019. Coupled
700 hydrogeophysical inversion to identify non-gaussian hydraulic conductivity
701 field by jointly assimilating geochemical and time-lapse geophysical data.
702 *Journal of Hydrology* 578, 124092. doi:10.1016/j.jhydrol.2019.124092.
- 703 Lan, T., Shi, X., Jiang, B., Sun, Y., Wu, J., 2018. Joint inversion of
704 physical and geochemical parameters in groundwater models by sequen-
705 tial ensemble-based optimal design. *Stochastic Environmental Research*
706 *and Risk Assessment* 32, 1919–1937. doi:10.1007/s00477-018-1521-5.

- 707 van Leeuwen, P.J., Evensen, G., 1996. Data assimilation and inverse methods
708 in terms of a probabilistic formulation. *Monthly Weather Review* 124,
709 2898–2913. doi:10.1175/1520-0493(1996)124<2898:daaimi>2.0.co;2.
- 710 Li, H., Kalnay, E., Miyoshi, T., 2009. Simultaneous estimation of co-
711 variance inflation and observation errors within an ensemble kalman fil-
712 ter. *Quarterly Journal of the Royal Meteorological Society* 135, 523–533.
713 doi:10.1002/qj.371.
- 714 Li, L., Bao, J., Cao, Z., Cui, F., 2019. Soil hydraulic parameters estimation
715 using gpr data via es-mda. *AGUFM 2019*, H43F–2047.
- 716 Li, L., Stetler, L., Cao, Z., Davis, A., 2018. An iterative normal-score ensem-
717 ble smoother for dealing with non-gaussianity in data assimilation. *Journal*
718 *of Hydrology* 567, 759–766.
- 719 Liang, X., Zheng, X., Zhang, S., Wu, G., Dai, Y., Li, Y., 2011. Maximum
720 likelihood estimation of inflation factors on error covariance matrices for
721 ensemble kalman filter assimilation. *Quarterly Journal of the Royal Mete-*
722 *orological Society* 138, 263–273. doi:10.1002/qj.912.
- 723 Mahar, P.S., Datta, B., 1997. Optimal monitoring network and
724 ground-water–pollution source identification. *Journal of Water Re-*
725 *sources Planning and Management* 123, 199–207. doi:10.1061/(asce)0733-
726 9496(1997)123:4(199).
- 727 Michalak, A.M., Kitanidis, P.K., 2004a. Application of geostatistical inverse
728 modeling to contaminant source identification at dover AFB, delaware.
729 *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.

- 752 Snodgrass, M.F., Kitanidis, P.K., 1997. A geostatistical approach to con-
753 taminant source identification. *Water Resources Research* 33, 537–546.
754 doi:10.1029/96wr03753.
- 755 Song, X., Chen, X., Ye, M., Dai, Z., Hammond, G., Zachara, J.M., 2019.
756 Delineating facies spatial distribution by integrating ensemble data as-
757 similations and indicator geostatistics with level-set transformation. *Water*
758 *Resources Research* 55, 2652–2671. doi:10.1029/2018wr023262.
- 759 Sun, A.Y., Painter, S.L., Wittmeyer, G.W., 2006. A robust approach for iter-
760 ative contaminant source location and release history recovery. *Journal of*
761 *Contaminant Hydrology* 88, 181–196. doi:10.1016/j.jconhyd.2006.06.006.
- 762 Todaro, V., D’Oria, M., Tanda, M.G., Gómez-Hernández, J.J., 2019. En-
763 semble smoother with multiple data assimilation for reverse flow routing.
764 *Computers & Geosciences* 131, 32–40. doi:10.1016/j.cageo.2019.06.002.
- 765 Uribe-Asarta, J., 2019. Modelación numérica de un experimento de trans-
766 porte de masa en un tanque de arena de laboratorio. Master’s thesis.
767 Universitat Politècnica de València.
- 768 Wang, X., Bishop, C.H., 2003. A comparison of breeding and en-
769 semble transform kalman filter ensemble forecast schemes. *Jour-*
770 *nal of the Atmospheric Sciences* 60, 1140–1158. doi:10.1175/1520-
771 0469(2003)060<1140:acobae>2.0.co;2.
- 772 Woodbury, A., Sudicky, E., Ulrych, T.J., Ludwig, R., 1998. Three-
773 dimensional plume source reconstruction using minimum relative en-

- 774 tropy inversion. *Journal of Contaminant Hydrology* 32, 131–158.
775 doi:10.1016/s0169-7722(97)00088-0.
- 776 Woodbury, A.D., Ulrych, T.J., 1996. Minimum relative entropy inver-
777 sion: Theory and application to recovering the release history of a
778 groundwater contaminant. *Water Resources Research* 32, 2671–2681.
779 doi:10.1029/95wr03818.
- 780 Xu, T., Gómez-Hernández, J.J., 2016. Joint identification of contaminant
781 source location, initial release time, and initial solute concentration in an
782 aquifer via ensemble kalman filtering. *Water Resources Research* 52, 6587–
783 6595. doi:10.1002/2016wr019111.
- 784 Xu, T., Gómez-Hernández, J.J., 2018. Simultaneous identification of a
785 contaminant source and hydraulic conductivity via the restart normal-
786 score ensemble kalman filter. *Advances in Water Resources* 112, 106–123.
787 doi:10.1016/j.advwatres.2017.12.011.
- 788 Xu, T., Gómez-Hernández, J.J., Chen, Z., Lu, C., 2020. A comparison be-
789 tween ES-MDA and restart EnKF for the purpose of the simultaneous
790 identification of a contaminant source and hydraulic conductivity. *Journal*
791 *of Hydrology* , 125681doi:10.1016/j.jhydrol.2020.125681.
- 792 Zanini, A., Woodbury, A.D., 2016. Contaminant source reconstruction by
793 empirical bayes and akaikes bayesian information criterion. *Journal of Con-*
794 *taminant Hydrology* 185-186, 74–86. doi:10.1016/j.jconhyd.2016.01.006.
- 795 Zhao, Y., Forouzanfar, F., Reynolds, A.C., 2016. History matching of multi-
796 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.