Joint identification of contaminant source location, initial release time and initial solute concentration in an aquifer via ensemble Kalman filtering

Teng Xu,[1] and J. Jaime Gómez-Hernández[1]

Corresponding author: Teng Xu, Institute for Water and Environmental Engineering, Universitat Politècnica de València, Camino de Vera, s/n, 46022 Valencia, Spain. (tenxu@posgrado.upv.es)

[1]Institute for Water and Environmental Engineering, Universitat Politècnica de València, Valencia, Spain.

Abstract. When a contaminant is detected in a drinking well, source location, initial contaminant release time and initial contaminant concentration are, in many cases, unknown; the responsible party may have disappeared and the identification of when and where the contamination happened may become difficult. Although contaminant source identification has been studied extensively in the last decades, we propose —to our knowledge, for the first time— the use of the ensemble Kalman filter (EnKF), which has proven to be a powerful algorithm for inverse modeling. The EnKF is tested in a two-dimensional synthetic deterministic aquifer, identifying, satisfactorily, the source location, the release time, and the release concentration, together with an assessment of the uncertainty associated with this identification.

1. Introduction

When, where and how much contaminant was introduced in an aquifer is a question many times asked once a pollutant is found in a drinking well. How to answer these three questions using the concentrations observed in monitoring wells downgradient of the contamination event has been subject of research for many years. *Gorelick et al.* [1983] used least-squares regression and linear programming combined with contaminant transport simulation to identify a pollutant source location by matching simulated and observed concentrations. *Woodbury and Ulrych* [1996] used minimum relative entropy to recover the release history of a plume in a one-dimensional system and then extended it to three dimensions [*Woodbury et al.*, 1998]. *Michalak and Kitanidis* [2003] proposed a Bayesian stochastic inverse modeling framework to estimate contamination history and extended it to the estimation of the antecedent distribution of a contaminant at a given point back in time [*Michalak and Kitanidis*, 2004]. *Neupauer and Wilson* [1999] used a backward probability model to derive travel time probability density functions, and later extended this method by conditioning those backward probability density functions on measured concentrations [*Neupauer and Lin*, 2006]; however, none of these methods could identify simultaneously source location and release time—the premise was that one of them had to be known to identify the other one. *Mahar and Datta* [2000] developed a non-linear optimization model for the estimation of the magnitude, location and duration of a groundwater pollution event under transient flow, where the governing equations of flow and transport where included in the optimization model as binding constraints. *Aral et al.* [2001] proposed a progressive genetic algorithm to solve an iterative non-linear opti-

mization problem in which contaminant source locations and release histories were defined as explicit unknown variables. *Yeh et al.* [2007, 2014] developed an approach combining simulated annealing and tabu search with a three-dimensional groundwater flow and solute transport model to estimate the source location, release concentration and release period, where the source location was determined by tabu search within a suspect source area, and trials for release concentrations and release periods were generated by simulated annealing. *Butera et al.* [2013]; *Cupola et al.* [2015] proposed a stochastic procedure (simultaneous release function and source location identification, SRSI), which estimates the source location and the release history through a Bayesian geostatistical approach; the method starts with a preliminary delineation of a probable source area and ends with a sub-area where the pollutant injection has most likely originated. *Gzyl et al.* [2014] has developed a multi-step approach to identify the source and its release history; this approach consists of three steps: performing integral pumping tests, identifying sources, and recovering the release history by means of a geostatistical approach.

In this paper, we will focus on the problem of simultaneously identifying the location of a continuous point source, its initial release time and its release concentration. The approach proposed uses the normal-score ensemble Kalman filter (NS-EnKF) [*Zhou et al.*, 2011], a variant of the ensemble Kalman filter (EnKF) which has proven to be very effective for the identification of highly heterogeneous, non-Gaussian hydraulic conductivities or porosities ([e.g., *Hendricks Franssen and Kinzelbach*, 2009; *Xu et al.*, 2013; *Xu and Gómez-Hernández*, 2015]), but which has never been applied, to the best of our knowledge, for contaminant source identification.

The EnKF is an assimilation algorithm that sequentially refines the estimates of the parameters of interest as information about the state of the system is collected. Contrary to most of the methods described above, it is not an optimization approach.

The paper proceeds with a description of the algorithm, followed by the analysis of the performance of the algorithm on a deterministic synthetic confined aquifer, and ends with the presentation and discussion of the results.

2. Methodology

The objective of this paper is to explore the applicability of the ensemble Kalman filter (more precisely of its variant the NS-EnKF) for the simultaneous identification of the parameters that define a continuous point injection of a solute into a confined aquifer, i.e., the injection location, the injection time and the the injection concentration. For the purpose of this exploration, we will assume that all other information necessary to build a flow and transport model in the aquifer is deterministically known. We realize that this is an unrealistic situation, but, for now, we wish to explore the potential of the NS-EnKF to identify the parameters defining the source —parameters quite different from those generally considered in the application of the EnKF for inverse modeling in groundwater.

As already mentioned above, the EnKF is an assimilation algorithm that re-estimates the parameters subject to identification as observations about the state of the system are collected. For this reason, the filter needs to know the relationship between parameters and state; in our case this relation is given by the groundwater flow and solute transport equations.

2.1. Groundwater flow and transport

In this work, and without loss of generality, we will assume that the flow field is at steady state; the governing equation is:

$$\nabla \cdot (K \nabla H) + q = 0 \tag{1}$$

where $\nabla\cdot$ is the divergence operator; ∇ is the gradient operator; H is the hydraulic head; K is the hydraulic conductivity [LT^{-1}], and q is the volumetric injection flow rate per unit volume of aquifer [LT^{-1}].

Also, we will assume, without loss of generality, that the solute travels in the aquifer subject to advection and dispersion; the governing equation is [*Zheng*, 2010]:

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

where C is the aqueous concentration [ML^{-3}], t is time [T], θ is the effective porosity [-], D_m is the molecular diffusion coefficient [L^2T^{-1}], α is the dispersivity tensor [L], v is the flow velocity vector related to the hydraulic head through, $v = (-K\nabla H)/\theta$ [LT^{-1}], q_s is the volumetric flow rate per unit volume of the aquifer representing fluid sources or sinks [T^{-1}], and C_s is the concentration of the source or sink flux [ML^{-3}].

Equation (1) is solved by finite differences using the numerical model MODFLOW [*McDonald and Harbaugh*, 1988] and the resulting piezometric heads are used to compute the flow velocities (v) in Eq. (2). The transport equation is solved with the numerical code MT3DMS [e.g., *Zheng*, 2010; *Ma et al.*, 2012]; of the many options available in MT3D to solve the advective and dispersive components of 2, we have chosen the implicit finite-different method for both of them (Group D on page 36 of the manual).

2.2. The ensemble Kalman filter

The Kalman filter is optimal when parameters and state variables are multiGaussian distributed and linearly related [*Aanonsen et al.*, 2009]. In our case, neither the parameters (identifying the source locations and release history) nor the state variables (solute concentrations) are multiGaussian distributed or linearly related. The ensemble Kalman filter [*Evensen*, 2003] was developed precisely to circumvent the problem provoked by the non-linearity of the state-transfer function [e.g., *Chen et al.*, 2009; *Xu and Gómez-Hernández*, 2015]. And later, the normal-score ensemble Kalman filter [*Zhou et al.*, 2011] was developed to address the issue of non-Gaussianity. The main difference on the implementation of the NS-EnKF for the purpose of contaminant source identification with respect to more standard applications is the need to restart the simulation of the transient evolution of the state variable from time zero after each assimilation step, given the very strong dependence of concentrations on the source location and release time, and the impossibility to introduce the updated source location into the state-transfer equation, once the simulated has started.

The NS-EnKF algorithm for source identification is described next. Before state data observations are collected, an ensemble of N_e realizations of the parameters are generated. In this particular case, we have four independent parameters, X for the x-coordinate, Y for the y-coordinate, T, for the initial release time and P for the concentration at the source; their initial values are drawn from wide-enough uniform distributions, making up N_e quadruplets (X_0, Y_0, T_0, P_0), where the subindices indicate that these are the parameter estimates at time $t = 0$. Then, the algorithm enters in a loop of forecast and updating; during the forecast, state variables are predicted for a given time in the future; time

at which state data are measured and the discrepancy between the forecasted values at observation locations and the actually observed values is used to update the parameters (optionally, it could update the state of the system, but since each forecast step has to be run from time zero, the update of the state is of no use). The loop proceeds as follows:

1. Forecast. With the last parameter update at observation time $t-1$, rerun the flow and transport code from time 0 until observation time t. As already mentioned above, parameter update implies the changing of the source location coordinates, and the only way to properly account for this coordinate change is by rerunning the code from the beginning:

$$C_t = \psi(C_0, X_{t-1}, Y_{t-1}, T_{t-1}, P_{t-1}) \tag{3}$$

where ψ represents the state transfer equation, in our case numerically approximated by the MODFLOW and MT3D codes. Parameter values are also "forecasted" into time t by simply keeping the updated values at time $t-1$

$$(X_t, Y_t, T_t, P_t) = (X_{t-1}, Y_{t-1}, T_{t-1}, P_{t-1}) \tag{4}$$

Contrary to the standard application of Kalman filtering, the state at $t+1$ is not obtained by forecasting an updated state computed at time t; because, when the source parameters are updated, the only way to account for this update is by rerunning the model from time 0. For this reason, the mass balance problem that appears in the standard application of the Kalman filter disappears here.

2. Observation. Concentrations are observed at sampling locations, C_t^o. To simplify the formulation we will assume that observations occur at the nodes of the numerical code; otherwise, there is the need to introduce some kind of interpolation to map forecasted values onto observation locations.

3. Assimilation. Forecasted concentrations at sampling locations will not match the observed ones, parameter values are updated according to the NS-EnKF formulation:

(i) Normal-score transform of parameters. Since the parameters are drawn from a uniform distribution, and the EnKF works best if the parameters are multiGaussian distributed, the quadruplets $S_t = (X_t, Y_t, T_t, P_t)$ are transformed into Gaussian deviates by their corresponding normal-score transform functions

$$\tilde{S}_t = \begin{pmatrix} \tilde{X}_t \\ \tilde{Y}_t \\ \tilde{T}_t \\ \tilde{P}_t \end{pmatrix} = \begin{pmatrix} \phi_{X,t}(X_t) \\ \phi_{Y,t}(Y_t) \\ \phi_{T,t}(T_t) \\ \phi_{P,t}(P_t) \end{pmatrix} \tag{5}$$

In the original formulation of the NS-EnKF, both parameters and state-variables are normal-score transformed, but our experience in this study as well as in previous studies shows that the method is equally capable of identifying non-Gaussian parameters if the state-variables remain untransformed. This is the reason why only parameters are normal-score transformed in this study.

(ii) Covariance calculation. From the ensemble of normal-score transform of the parameters and the ensemble of forecast values at observation locations, compute the cross-covariance between parameters and concentrations $D_{\tilde{S}C^{f,o}}$, and the auto-covariance of concentrations $D_{C^{f,o}C^{f,o}}$, where the superindices f, o are used to clarify that only the forecasted concentrations at observation locations are used for this computation.

(iii) Update. Parameters are updated, for each quadruplet of the ensemble, according to the equation

$$\tilde{S}_t^a = \tilde{S}_t + G_t(C_t^o + e_t - C_t^{f,o}) \tag{6}$$

where $\tilde{S}_t^a$ is the updated normal-scored parameter quadruplet, $\tilde{S}_t$ is the forecasted normal-scored parameter quadruplet, C_t^o is the observed concentration vector, $C_t^{f,o}$ is the fore-

casted concentration vector at observation locations, e_t is an observation error vector (different for each member of the ensemble and drawn from a normal distribution with mean zero and covariance matrix R_t), and G_t is the Kalman gain matrix, given by

$$G_t = D_{\tilde{S}C^{f,o}}(D_{C^{f,o}C^{f,o}} + R_t)^{-1} \tag{7}$$

Concentrations are not updated, since the forecast for time t with the updated parameter values S_t^a has to be done from time $t = 0$.

(iv) Backtransform. The updated parameter quadruplets are back transformed into parameter space for each member of the ensemble using the inverse of the transformation function used in (i)

$$\begin{pmatrix} X_t^a \\ Y_t^a \\ T_t^a \\ P_t^a \end{pmatrix} = \begin{pmatrix} \phi_{X,t}^{-1}(\tilde{X}_t^a) \\ \phi_{Y,t}^{-1}(\tilde{Y}_t^a) \\ \phi_{T,t}^{-1}(\tilde{T}_t^a) \\ \phi_{P,t}^{-1}(\tilde{P}_t^a) \end{pmatrix} \tag{8}$$

4. Go back to step 1 and repeat until all concentration observations have been assimilated.

3. Application

The algorithm is demonstrated in a two-dimensional aquifer of 50 m by 50 m by 5 m, discretized into 50 by 50 by 1 cells. Hydraulic conductivities are heterogeneous in space; their logarithm follows a multiGaussian distribution of mean 2.5 ln(m/s), standard deviation 1.5 ln(m/s), and isotropic exponential variogram with a range of 20 m (see Figure 1). The four boundaries of the aquifer are impermeable, and there are 2 injection wells with an injection rate of 2 m^3/d at coordinates (4.5 m, 9.5 m) and (4.5 m, 39.5 m), and 2 pumping wells with a pumping rate of 2 m^3/d at coordinates (44.5 m, 9.5 m) and (44.5 m, 39.5 m) (see Figure 2). The rest of the parameters are homogeneous,

155 porosity is 0.3, longitudinal dispersivity is 1 m, and transversal dispersivity is 0.01 m.
156 Steady-state flow is solved, and then contaminant transport for a continuous punctual
157 release from location (11.5 m, 19.5 m) at time t= 80 days and with a concentration of
158 60 mg/l is modeled for 1000 days. The simulation time is discretized into 100 time steps.
159 Observations are taken at the end of each time step at 25 observation locations, and they
160 will be assimilated by the NS-EnKF. The locations of the observation locations is shown
161 in Figure 2, and their coordinates are listed in Table **??**

3.1. Analysis

162 An initial ensemble of 1000 parameter quadruplets are generated from uniform distribu-
163 tions with a wide range around the truth; more precisely, $X \sim U[5, 15]$ m, $Y \sim U[15, 25]$
164 m, $T \sim U[50, 150]$ d, and $P \sim U[10, 180]$ mg/l. Figure 3 shows the histograms of the
165 initial distributions of each parameter. For each quadruplet, flow and transport are sim-
166 ulated with the rest of the parameters exactly equal to those of the reference case; then,
167 during the next 50 time steps (up to 623 days) concentrations are observed in the refer-
168 ence and are assimilated by the NS-EnKF producing, at each time step, a new ensemble
169 of parameter quadruplets. These quadruplets will converge, as shown below, to the true
170 release location, time and concentration we wish to identify.

171 We have produced a number of figures to illustrate the time evolution of the source
172 identification and how, eventually, the source is correctly identified within the resolution
173 allowed by the transport simulation code. Figure 4 shows the histograms of the updated
174 quadruplets at the end of time step 50; considering that for the specific parameters used
175 to simulated transport by MT3DMS we can only specify the cell at which the solute is
176 introduced into the aquifer, the identification of the source location is exact up to the

cell discretization. Similarly happens with the identification of the time release, which is found to be in the interval $[79.5, 80.5]$ d. Finally, the release concentration is identified to be in the interval $[59.5, 61.0]$ mg/l, which is a very good approximation of the initial release value of 60 mg/l.

Figure 5 shows the time evolution of the means and variances of the four parameters as new observations are assimilated. Around time step 15, the mean values of location and time get very close to the final estimates and stay there, whereas the concentration needs 25 time steps to reach a similar stabilization; the variances decrease with time and remain close to zero around time step 25. A behavior that we have not been able to explain is the oscillation of the variance of the release concentration after time step 30, the mean estimate is stable around 60 mg/l, but the variance oscillates. The final mean values can be read in the histograms of Figure 4, and provide an estimate of the release quadruplet of (11.4 m, 19.6 m, 80 d, 60.4 mg/l), with an uncertainty characterized by a standard deviation of (0.3 m, 0.3 m, 0.7 d, 2.7 mg/l).

To appreciate the evolution in time of the identification of the source location, Figure 6 shows the identified locations before assimilation, and at the end of times steps 8, 20 and 50. The time evolution of the plume in the reference is shown in Figure 7, the observations used in the assimilation phase are sampled from these maps. We can appreciate how during the first 8 time steps virtually no improvement is produced on the identification of the source location, this is because the plume has no migrated much downstream from the source, and therefore, the plume has not been detected in the observation wells yet. At time step 20, the cell where the solute enters the aquifer starts to be noticed, and at time step 50 virtually all 1000 estimated source locations are within the release cell. To

complement this figure, Figure 8a shows the evolution of the source location estimate in one of the ensemble members; here, we can observe how the source location is updated as a function of time and Figure 8b shows how the ensemble mean of all 1000 locations move in time from the center of the suspect release area to the center of the real release cell.

4. Summary and discussion

With this work, we have proven the ability of the NS-EnKF to identify a contaminant source location, release time and concentration. The conditions under which this identification has been performed are still unrealistic from a practical point of view. The perfect knowledge of aquifer parameters, stresses and boundary and initial conditions will never happen; but our interest was to show the potential of this assimilation algorithm that has been successfully applied in hydrogeology for hydraulic conductivity characterization. The results are very satisfactory, and open a new avenue of research aimed at using this approach for source identification in more realistic cases. Our purpose is to continue research in this avenue, including, next, the simultaneous identification of a heterogeneous conductivity field.

Acknowledgments. Financial support to carry out this work was received from the Spanish Ministry of Economy and Competitiveness through project CGL2014-59841-P. All data used in this analysis are available from the authors.

References

Aanonsen, S., G. Nævdal, D. Oliver, A. Reynolds, and B. Vallès (2009), The ensemble kalman filter in reservoir engineering–a review, *SPE Journal*, *14*(3), 393–412.

Aral, M. M., J. Guan, and M. L. Maslia (2001), Identification of contaminant source location and release history in aquifers, *Journal of hydrologic engineering*, *6*(3), 225–234.

Butera, I., M. G. Tanda, and A. Zanini (2013), 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*(5), 1269–1280.

Chen, Y., D. Oliver, and D. Zhang (2009), Data assimilation for nonlinear problems by ensemble kalman filter with reparameterization, *Journal of Petroleum Science and Engineering*, *66*(1), 1–14.

Cupola, F., M. G. Tanda, and A. Zanini (2015), Laboratory sandbox validation of pollutant source location methods, *Stochastic Environmental Research and Risk Assessment*, *29*(1), 169–182.

Evensen, G. (2003), The ensemble kalman filter: Theoretical formulation and practical implementation, *Ocean Dynamics*, *53*(4), 343–367.

Gorelick, S. M., B. Evans, and I. Remson (1983), Identifying sources of groundwater pollution: an optimization approach, *Water Resources Research*, *19*(3), 779–790.

Gzyl, G., A. Zanini, R. Fraczek, and K. Kura (2014), Contaminant source and release history identification in groundwater: A multi-step approach, *Journal of contaminant hydrology*, *157*, 59–72.

Hendricks Franssen, H. J., and W. Kinzelbach (2009), Ensemble kalman filtering versus sequential self-calibration for inverse modelling of dynamic groundwater flow systems, *Journal of Hydrology*, *365*(3-4), 261–274.

Ma, R., C. Zheng, J. M. Zachara, and M. Tonkin (2012), Utility of bromide and heat tracers for aquifer characterization affected by highly transient flow conditions, *Water Resources Research*, *48*(8).

Mahar, P. S., and B. Datta (2000), Identification of pollution sources in transient groundwater systems, *Water Resources Management*, *14*(3), 209–227.

McDonald, M., and A. Harbaugh (1988), A modular three-dimensional finite-difference ground-water flow model.

Michalak, A. M., and P. K. Kitanidis (2003), A method for enforcing parameter nonnegativity in bayesian inverse problems with an application to contaminant source identification, *Water Resources Research*, *39*(2).

Michalak, A. M., and P. K. Kitanidis (2004), Application of geostatistical inverse modeling to contaminant source identification at dover afb, delaware, *Journal of Hydraulic Research*, *42*(S1), 9–18.

Neupauer, R. M., and R. Lin (2006), Identifying sources of a conservative groundwater contaminant using backward probabilities conditioned on measured concentrations, *Water resources research*, *42*(3).

Neupauer, R. M., and J. L. Wilson (1999), Adjoint method for obtaining backward-in-time location and travel time probabilities of a conservative groundwater contaminant, *Water Resources Research*, *35*(11), 3389–3398.

Woodbury, A., E. Sudicky, T. J. Ulrych, and R. Ludwig (1998), Three-dimensional plume source reconstruction using minimum relative entropy inversion, *Journal of Contaminant Hydrology*, *32*(1), 131–158.

Woodbury, A. D., and T. J. Ulrych (1996), Minimum relative entropy inversion: Theory and application to recovering the release history of a groundwater contaminant, *Water Resources Research*, *32*(9), 2671–2681.

Xu, T., and J. J. Gómez-Hernández (2015), Probability fields revisited in the context of ensemble kalman filtering, *Journal of Hydrology*, *531*, 40–52.

Xu, T., J. J. Gómez-Hernández, H. Zhou, and L. Li (2013), The power of transient piezometric head data in inverse modeling: an application of the localized normal-score enkf with covariance inflation in a heterogenous bimodal hydraulic conductivity field, *Advances in Water Resources*, *54*, 100–118.

Yeh, H.-D., T.-H. Chang, and Y.-C. Lin (2007), Groundwater contaminant source identification by a hybrid heuristic approach, *Water Resources Research*, *43*(9).

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

Zheng, C. (2010), Mt3dms v5. 3supplemental users guide: Tuscaloosa, ala., university of alabama department of geological sciences, *Tech. rep.*, Technical Report to the US Army Engineer Research and Development Center.

Zhou, H., J. Gómez-Hernández, H. Hendricks Franssen, and L. Li (2011), An approach to handling non-gaussianity of parameters and state variables in ensemble kalman filtering, *Advances in Water Resources*, *34*(7), 844–864.

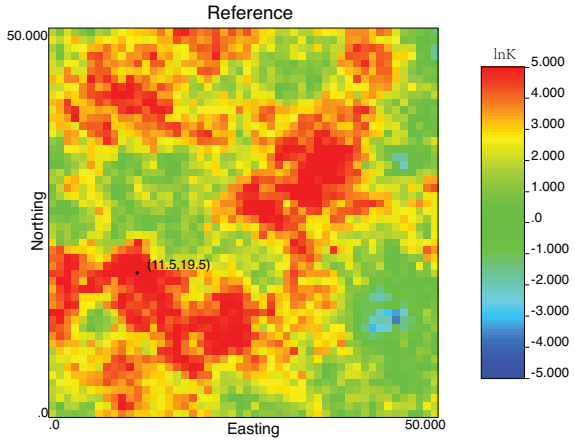

Figure 1. Reference ln$K$. The real source location (dark dot) is at coordinates (11.5 m, 19.5 m).

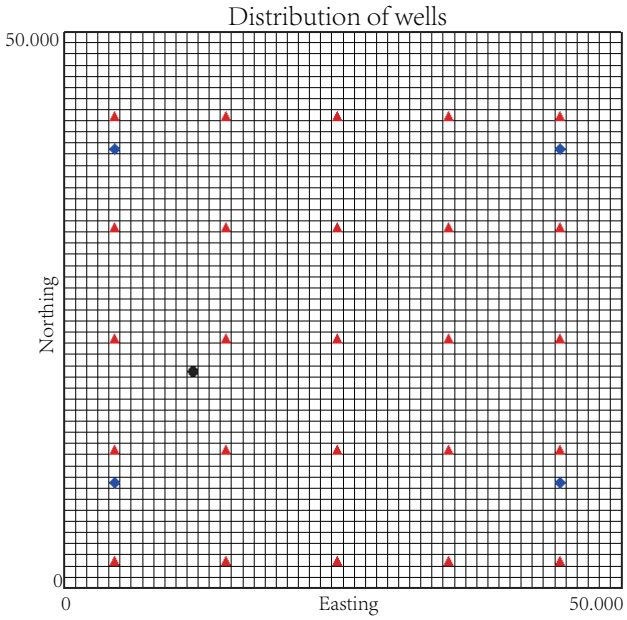

Figure 2. Well locations. Red triangles denote observation wells; blue diamonds denote injection wells (near the west boundary) and pumping wells (near the east boundary). The black circle is the contaminant source location.

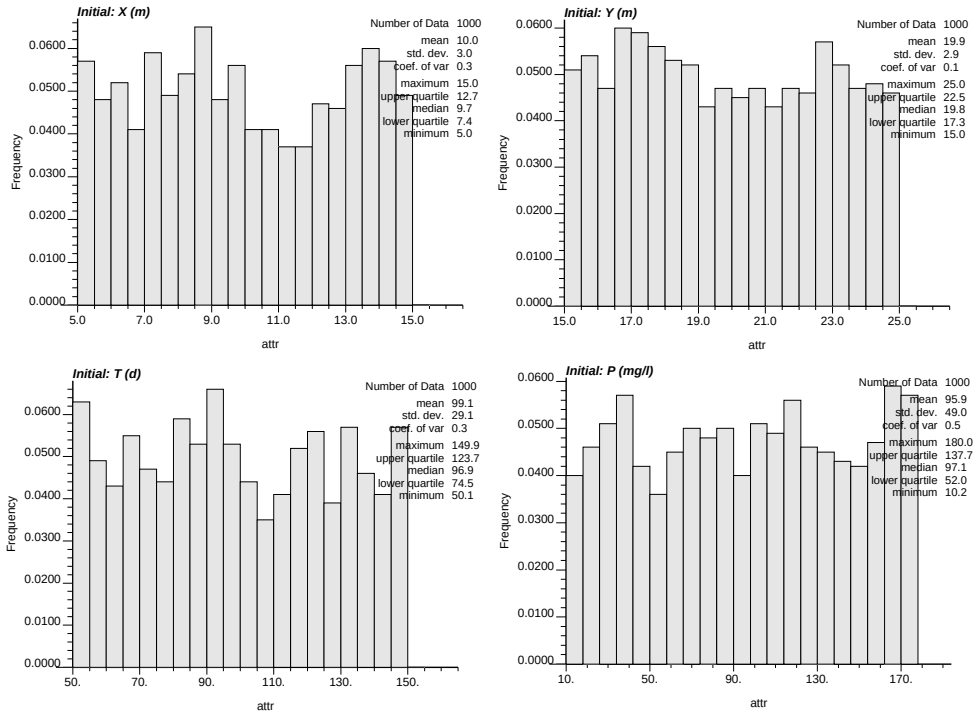

Figure 3. Histogram of initial realizations of contaminant source parameters: X and Y coordinates, release time T and release concentration P.

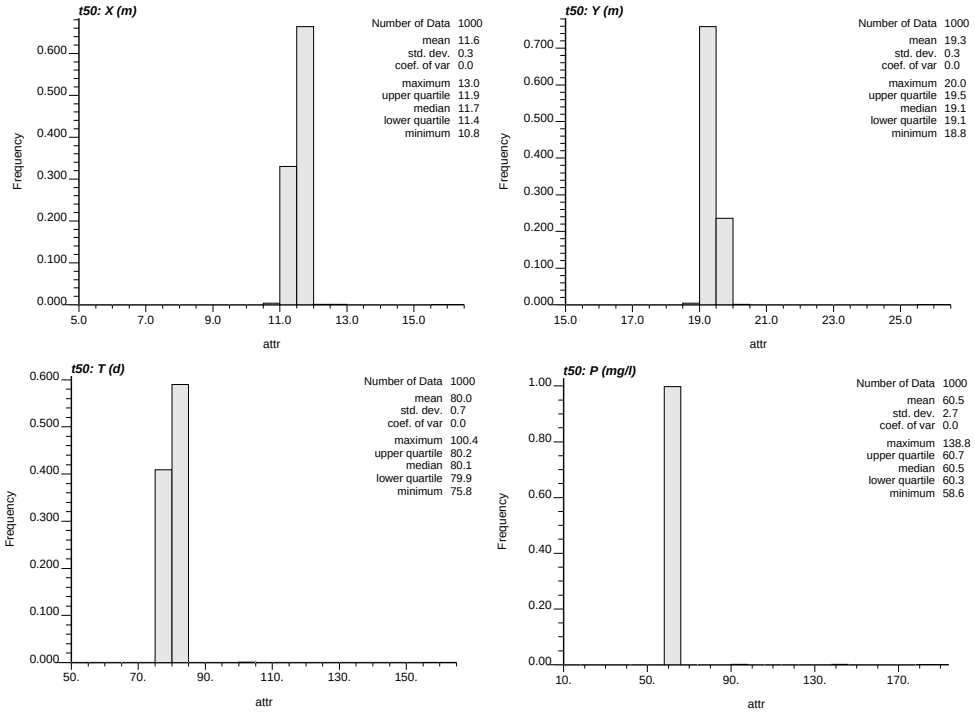

Figure 4. Histogram of the contaminant source parameters at the end of the assimilation period (50 time steps).

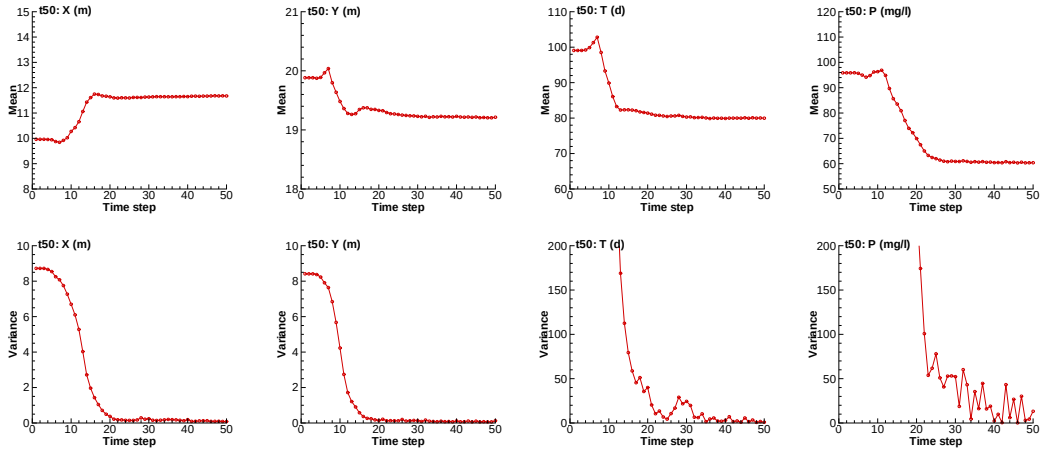

Figure 5. Top row, ensemble mean. Bottom row, ensemble variance. Time evolution of the ensemble mean and variance of the four parameters under identification.

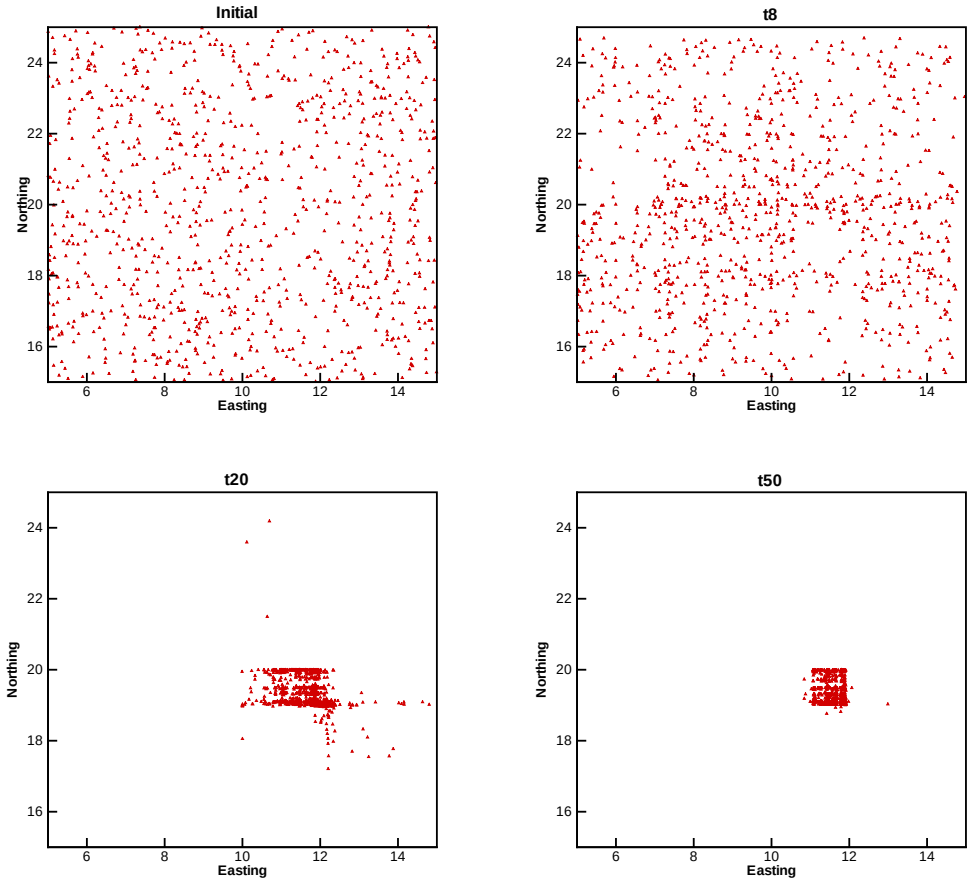

Figure 6. Ensemble source location spatial coordinates estimates at the initial time step, and at the end of time steps 8, 20 and 50.

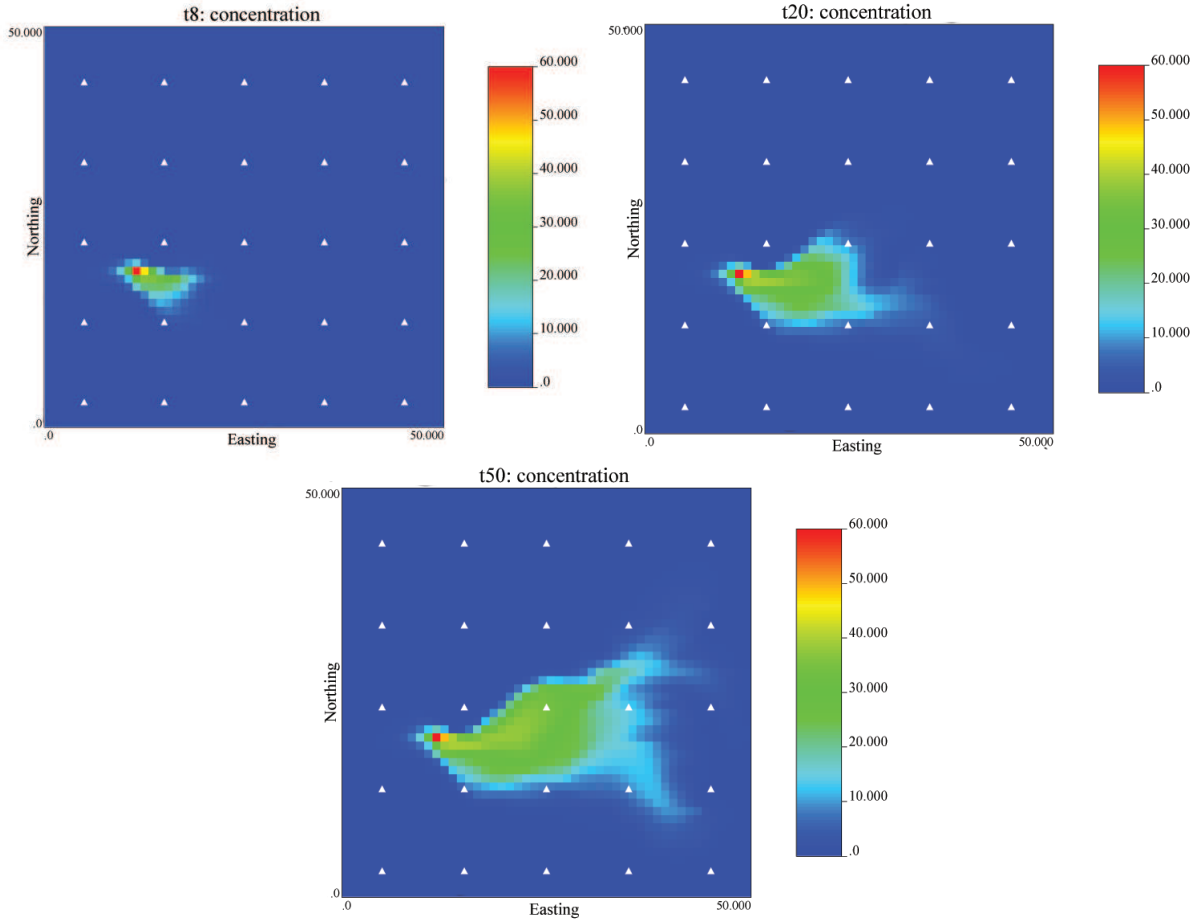

Figure 7. Solute plume evolution in time in the reference aquifer. Data are observed at well locations and assimilated via Kalman filter to update the source parameters

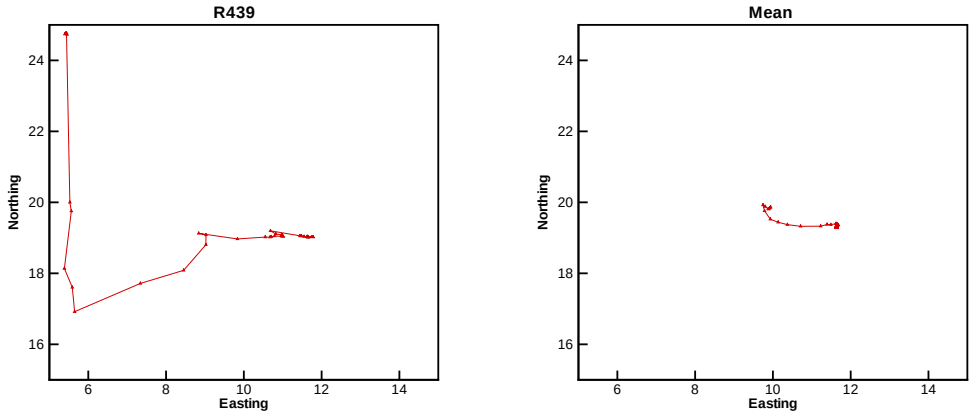

Figure 8. Evolution in time of the estimate of the source spatial coordinates in ensemble quadruplet number 439 (left) and of the ensemble mean (right)

Table 1. Coordinates in m of the observation wells.

x	y	x	y	x	y
4.5	2.5	4.5	22.5	4.5	42.5
14.5	2.5	14.5	22.5	14.5	42.5
24.5	2.5	24.5	22.5	24.5	42.5
34.5	2.5	34.5	22.5	34.5	42.5
44.5	2.5	44.5	22.5	44.5	42.5
4.5	12.5	4.5	32.5		
14.5	12.5	14.5	32.5		
24.5	12.5	24.5	32.5		
34.5	12.5	34.5	32.5		
44.5	12.5	44.5	32.5		