Simultaneous identification of a contaminant source and hydraulic conductivity via the restart normal-score ensemble Kalman filter

Teng Xu \(^1\)

J. Jaime Gómez-Hernández \(^2\)

\(^1\)corresponding author. E-mail: tenxu@posgrado.upv.es

\(^2\)E-mail: jaime@dihma.upv.es
Simultaneous identification of a contaminant source and hydraulic conductivity via the restart normal-score ensemble Kalman filter

Teng Xu\textsuperscript{a,*}, J. Jaime Gómez-Hernández\textsuperscript{a}

\textsuperscript{a}Research Institute of Water and Environmental Engineering, Universitat Politècnica de València, 46022, Valencia, Spain

Abstract

Detecting where and when a contaminant entered an aquifer from observations downgradient of the source is a difficult task; this identification becomes more challenging when the uncertainty about the spatial distribution of hydraulic conductivity is accounted for. In this paper, we have implemented an application of the restart normal-score ensemble Kalman filter (NS-EnKF) for the simultaneous identification of a contaminant source and the spatially variable hydraulic conductivity in an aquifer. The method is capable of providing estimates of the spatial location, initial release time, the duration of the release and the mass load of a point-contamination event, plus the spatial distribution of hydraulic conductivity together with an assessment of the estimation uncertainty of all the parameters. The method has been applied in synthetic aquifers exhibiting both Gaussian and non-Gaussian patterns. The identification is made possible by assimilating in time both piezometric head and concentration observations from an array of observation wells. The method is demonstrated in three different synthetic scenarios that combine hydraulic conductivities with unimodal and bimodal histograms, and releases in high and low conductivity zones. The results prove that the specific implementation of the EnKF is capable of recovering the source parameters with some uncertainty and of recovering the main patterns of heterogeneity of the hydraulic conductivity fields by assimilating a sufficient number of state variable observations. The proposed approach is an important step towards contaminant source identification in real aquifers, which may have logconductivity spatial distributions with either Gaussian or non-Gaussian features, yet, it is still far from practical applications since the transport parameters, the external sinks and sources and the initial and boundary conditions are assumed known.

Keywords: Contaminant source identification, restart Ensemble Kalman filter, Heterogeneity, Normal-score transform

\textsuperscript{*}Corresponding author: Teng Xu Tel: +34 963879615 Fax: +34 963879492

Email addresses: tenxuposgrado.upv.es (Teng Xu), jaime@dihma.upv.es (J. Jaime Gómez-Hernández)
1. Introduction

Groundwater movement is slow and often a contamination plume may reach a water supply well when
the source has disappeared; or a contaminant may enter an aquifer unnoticed, i.e., from a leaky underground
storage tank, and when the contaminant is detected, nobody knows exactly its origin. Could the source
be identified from the time series of concentrations observed in one or several downstream wells from the
release point? This is a question that many researchers have posed and which has been studied in the
past decades. The vast majority of all previous approaches are based on some sort of optimization of an
objective function that measures deviations between model predictions and observations, with the source
location and release time as the parameters to identify. For example, Gorelick et al. [1] used least-squares
regression and linear programming combined with contaminant transport simulation to identify a pollutant
source location; Sun et al. [2] proposed a constrained robust least squares approach (CRLS) for contami-
nant release history identification and then used the CRLS estimator combined with a branch-and-bound
global optimization for iteratively identifying source release histories and source locations [3]; Aral et al.
[4] proposed a progressive genetic algorithm in the context of nonlinear optimization; Yeh et al. [5, 6] com-
bined simulated annealing and tabu search; Mirghani et al. [7] employed a simulation-optimization approach
that uses an evolutionary search algorithm; Dokou and Pinder [8] developed an optimal search strategy
for source location identification; Amirabdollahian and Datta [9] designed an optimal source identification
model using Adaptive Simulated Annealing to identify the contaminant source; Ayvaz [10] proposed a hybrid
simulation-optimization approach for solving the areal groundwater pollution source identification problem.
Some approaches are not based on optimization; for instance, Butera et al. [11], Cupola et al. [12] employed a
Bayesian geostatistical approach to identify the source location, after a preliminary delineation of a probable
source area; and Gzyl et al. [13] used a quasilinear geostatistical methodology to identify the source location
from some suspect contamination sources.

All the studies mentioned above mainly focus on the identification of source location information in a de-
terministic aquifer where the aquifer properties are assumed perfectly known. There are a few studies carried
out for the simultaneous identification of contaminant source information and aquifer properties. Wagner [14]
used non-linear maximum likelihood combined with groundwater flow and contaminant transport simulation
for the estimation of flow parameters in a two-zone aquifer and the source parameters. Sidauruk et al. [15]
developed inverse procedures based on optimizing correlation coefficients to locate ground water contaminant
sources and to identify transport parameters of a homogeneous aquifer. Mahar and Datta [16] used a nonlin-
ear optimization model where the flow and transport equations are embedded as constraints to estimate the
magnitude, location and duration of a groundwater pollution source and later [17] they extended the method
to include the simultaneous estimation of aquifer parameters in a homogeneous and isotropic aquifer system.

Singh and Datta [18] utilized a trained artificial neural network to simultaneously estimate a groundwater
contaminant source and to estimate the hydraulic conductivity, porosity, and dispersivity of a homogeneous
aquifer. Datta et al. [19] presented a methodology using a combined optimization-simulation approach for
simultaneously identifying a contaminant source and estimating aquifer homogeneous parameters, where the
methodology links an optimization method to a groundwater flow and transport simulator as an external
module. Koch and Nowak [20] combined a stochastic multiphase model and a reverse transport formulation
within an inverse Bayesian methodology for the joint inversion of contaminated source zone architectures
and aquifer parameters.

However, with the exception of the work by Koch and Nowak Koch and Nowak [20], in all of those
studies, the aquifers analyzed are either homogeneous, or, at most divided in a few homogeneous subzones.
In a previous paper [21], we proposed the use of the ensemble Kalman filter (EnKF) —more precisely, its
variant, the restart normal-score EnKF (NS-EnKF), which has proven to be a very efficient inverse modeling
algorithm [e.g., 22, 23, 24, 25]— for contaminant source identification in a heterogeneous but deterministic
aquifer, and we ended the paper with the conclusion that although the NS-EnKF performed very well for
contaminant source identification, its application to a deterministic aquifer was unrealistic since detailed
aquifer heterogeneity will never be available in a practical case. In this paper, recognizing the importance of
proper characterization of conductivity for solute transport prediction [e.g., 26, 27], we move one step further
and we demonstrate the applicability of the NS-EnKF for the simultaneous identification of a heterogeneous
conductivity field and the parameters defining the source of a point contamination by assimilating in time
piezometric heads and solute concentrations. We must recognize upfront, that the rest of parameters and
variables controlling flow are assumed known, i.e., initial conditions, boundary conditions, and external
stresses, as well as the parameters controlling transport, which are considered homogeneous and known;
yet, the proposed methodology brings us closer to its application in a realistic setting from what had been
proposed in the past in the literature.

Next, the paper describes the proposed algorithm and continues with an application of the restart NS-
EnKF for three scenarios in two different heterogenous synthetic aquifers. The paper ends with a summary
and discussion, followed by an Appendix in which an extra scenario is discussed.
2. Ensemble Kalman filter

The EnKF proposed by Evensen [28] is the evolution of the Kalman filter to handle nonlinear transfer functions. It is based on using an ensemble of realizations to approximate the covariances and cross-covariances of parameters and state variables needed during the updating step of the Kalman filter. In recent years, the EnKF has received much attention for its efficiency and effectiveness in dealing with large dimensions and it has been widely applied in many fields, such as oceanography, meteorology, petroleum engineering, or hydrology [e.g., 29, 30, 31, 32, 23, 33].

Briefly, recall that the Kalman filter is an assimilation technique that updates the state and the parameters of the system sequentially in time as new state data are collected. The update is proportional to the deviations between the state forecast and the state observations at a few locations; in the ensemble Kalman filter, this update is heterogeneous in space and different for each member of the ensemble of realizations. In our case, the state variables are piezometric head ($H$) and solute concentration ($C$), and the parameters to update are hydraulic conductivity ($K$) and contaminant source information including location ($X$ for the x-coordinate, $Y$ for the y-coordinate), initial release time ($T$), release duration ($\Delta T$), and solute mass-loading rate ($M$).

Next, we detail the specific implementation of the restart NS-EnKF, which, like all the Kalman filter-based approaches, consists of two steps: forecast and analysis.

In the forecast step, state variables are forecasted into the next time step. Commonly, this forecast is done from the state estimate after the last updating time step; however, for the contaminant source identification, as was explained by Xu and Gómez-Hernández [21], Camporese et al. [34, 35], Crestani et al. [36, 37], it is necessary to make the forecast from time zero, since the contaminant source parameters refer to the source at time zero. This restart of the simulation after each time step, which is mandatory for the purpose of identifying the contaminant source, gives the name of “restart” NS-EnKF to the algorithm. Therefore, at each time step, the forecast of piezometric heads and concentrations is done from time zero until the end of the $t^{th}$ time step ($C_t, H_t$) based on the concentrations and heads at time zero ($C_0, H_0$) and on the last updates of the source information and of the conductivity field at the $(t - 1)^{th}$ time step ($X_{t-1}, Y_{t-1}, T_{t-1}, \Delta T_{t-1}, M_{t-1}, K_{t-1}$). The forecast uses a state transfer equation $\psi$, which is non other than a combined groundwater flow and contaminant transport model,

$$(C_t, H_t) = \psi(C_0, H_0, X_{t-1}, Y_{t-1}, T_{t-1}, \Delta T_{t-1}, M_{t-1}, K_{t-1}).$$ (1)

In the demonstration examples, we assume that the system is under transient conditions (both for ground-
water flow and for contaminant transport) and that only advection and dispersion are considered as transport mechanisms. The governing equations of the transient groundwater flow and contaminant transport are given in Eq. (2) [38] and Eq. (3) [39], respectively:

\[
S_s \frac{\partial H}{\partial t} - \nabla \cdot (K \nabla H) = W, \tag{2}
\]

\[
\frac{\partial (\theta C)}{\partial t} = \nabla \cdot [\theta (D_m + \alpha v) \cdot \nabla C] - \nabla \cdot (\theta v C) - q_s C_s. \tag{3}
\]

where \( S_s \) denotes specific storage \([L^{-1}]\), \( H \) is the hydraulic head \([L]\); \( t \) is time \([T]\); \( \nabla \cdot \) is the divergence operator; \( \nabla \) is the gradient operator; \( K \) is the hydraulic conductivity \([LT^{-1}]\); \( W \) denotes sources and sinks per unit volume \([T^{-1}]\). \( C \) is aqueous concentration \([ML^{-1}]\), \( \theta \) is the effective porosity of the aquifer sediment \([-]\). \( D_m \) is the molecular diffusion coefficient \([L^2T^{-1}]\), \( \alpha \) is the dispersivity tensor \([L]\), \( v \) is the flow velocity vector \([LT^{-1}]\), which is related to the specific discharge through, \( v = -\frac{K}{\theta} \nabla H \); \( q_s \) 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 \([ML^{-3}]\).

The groundwater flow simulator MODFLOW [40] is used to solve Eq. (2), and the resulting flow velocity field \( v \) is used as input in Eq. (3), which is solved with the transport simulator MT3DMS [e.g., 39, 41].

Note that although the EnKF can handle nonlinearities, it still fails to characterize non-Gaussian parameters. To solve this problem and take advantage of the optimality of the EnKF for parameter identification under Gaussian conditions, several techniques, such as Gaussian mixture model, reparameterization, and normal-score transform (also referred to in the literature as Gaussian anamorphosis) have been employed in combination with the EnKF [e.g., 42, 43, 44]. The normal-score EnKF (NS-EnKF), as described by Zhou et al. [44], is one of the alternatives based on the use of a Gaussian anamorphosis to transform the parameters being identified into (univariate) Gaussian deviates [e.g., 45, 46, 23]. In this paper, we use this approach to handle the non-Gaussian parameters.

The source contamination parameters \((X_{t-1}, Y_{t-1}, T_{t-1}, \Delta T_{t-1}, M_{t-1})\) and the conductivities \(K_{t-1}\) are all transformed into Gaussian deviates by using a specific normal-score transform for each parameter,
where $\phi_{X;t}, \phi_{Y;t}, \phi_{T;t}, \phi_{\Delta T;t}, \phi_{M;t}, \phi_{K;t}$ are the normal-score transform functions.

The prediction step in the NS-EnKF refers to the prediction of the normal-scored transformed variables from the normal-scored updated values at the last time step. Such a prediction involves the back-transformation of the updated normal scores followed by the forward modeling of flow and transport using Eq. (2) and (3) and then the transformation of the model results into normal scores.

In the analysis step, the EnKF will update the last estimate of the (normal-scored) parameters based on the discrepancy between the forecasted state variable values and the observed ones. The details of the normal-score EnKF can be looked up elsewhere [e.g., 44]. We simply summarize here that there is an ensemble of augmented vectors containing parameters and state variables, each vector of the ensemble can be split in two, a vector of parameters $S$ (in our case, the normal-scored values in Eq. (5)), and a vector of states $V$ (in our case, the piezometric heads and solute concentrations in Eq. (6)),

\[
S = \begin{bmatrix}
\tilde{X} \\
\tilde{Y} \\
\tilde{T} \\
\Delta T \\
\tilde{M} \\
\tilde{K}
\end{bmatrix},
\]

\[
V = \begin{bmatrix}
C \\
H
\end{bmatrix}.
\]

From the ensembles of both vectors, compute the experimental state covariance ($D_{VV}$) for all pairs of locations with observational data, and the experimental state-parameter cross-covariance for all pairs of normal-scored parameter values and observational data ($D_{SV}$). Once the experimental covariances are
computed, the normal-scored parameters are updated using the following equation:

\[ S_t = S_{t-1} + G_t (V_0^t + e_t - V_f^t) \] (7)

with

\[ G_t = D_{SV} (D_{VV} + R_t)^{-1}, \] (8)

where \( S_t \) contains the updated normal-scored parameters, \( G_t \) is the Kalman gain matrix, \( V_0^t \) is the vector of observed concentrations and observed heads, \( V_f^t \) is the vector of forecasted concentrations and forecasted heads at the observation locations, and \( e_t \) is an observation error with mean zero, standard deviation of 0.05, and diagonal covariance \( R_t \). Notice that the update is proportional to the departure between observed \( V_0^t \) and forecasted state variables \( V_f^t \) (at observation locations), with the Kalman gain \( G_t \) being the proportionality factor. The Kalman gain (Eq. (8)) is proportional to the degree of correlation between state data and parameters as measured by the cross-covariance \( D_{SV} \), and inverse proportional to the redundancy of the observed states as measured by the auto-covariance \( D_{VV} \).

After all the normal-scored parameters are updated, back transform them into parameter space using the inverse of the normal-score functions (see Equation Eq. (9)),

\[
\begin{pmatrix}
    X_t \\
    Y_t \\
    T_t \\
    \Delta T_t \\
    M_t \\
    K_t
\end{pmatrix}
= \begin{pmatrix}
    \phi^{-1}_{X,t}(\tilde{X}_t) \\
    \phi^{-1}_{Y,t}(\tilde{Y}_t) \\
    \phi^{-1}_{T,t}(\tilde{T}_t) \\
    \phi^{-1}_{\Delta T,t}(\tilde{\Delta T}_t) \\
    \phi^{-1}_{M,t}(\tilde{M}_t) \\
    \phi^{-1}_{K,t}(\tilde{K}_t)
\end{pmatrix}
\] (9)

3. Application

Two synthetic confined aquifers are built: one has a unimodal conductivity distribution; the other has a bimodal conductivity distribution with strong continuity of the values around the larger mode of the two, mimicking a channelized geology. Both aquifers are modeled as confined and discretized into 50 by 50 by 1 cells. (No units are reported throughout the paper since the results are insensitive to the units used as long as all quantities have coherent units. Absolute magnitudes are not important, what is important are relative values, such as the number of orders of magnitude of variability of the conductivity, the relative size
of the heterogeneity features with respect to the aquifer size, or the size of the plume within the aquifer for
the length of the simulation period. Physical dimensions are given in square brackets when a variable or
parameter appears for the first time.) Each cell is 1 [L] by 1 [L] by 5 [L].

Conductivities in both aquifers are heterogeneous. The reference log-conductivity field of the unimodal
aquifer (see Figure 1) is generated with GCOSIM3D [47] from a Gaussian distribution with the parameters
shown in the first row of Table 1. The channelized aquifer consists of 35% high conductivity channel
values and 65% low conductivity non-channel medium. The procedure for the generation of the reference
log-conductivity field is as follows: first, generate a binary facies map using the SNESIM code [48] with the
training image in Figure 2, and then populate each facies with log-conductivity values from two multiGaussian
distributions with the parameters shown in the last two rows of Table 1, each facies is generated independently
with the code GCOSIM3D. The resulting reference log-conductivity field of the channelized aquifer is shown
in Figure 1 with a global mean of -1.2 [ln(LT^{-1})] and a global standard deviation of 1.9 [ln(LT^{-1})].

In both aquifers, boundaries are impermeable, and there are 2 injection wells, labeled #1 and #2 (near
the west boundary), with injecting rates of 10.2, and 9.5 [L^3T^{-1}], respectively, and 3 pumping wells, labeled
#3, #4, and #5 (near the east boundary), with pumping rates of 5.7, 5.5 and 8.8 [L^3T^{-1}], respectively.
This setup induces a flow from left to right with the presence of three sinks which will act as attractors of
the contaminant plume. The initial piezometric heads are set to 8 [L].

There are also 25 observation wells (see Figure 3), which will be used during the updating step of the
restart NS-EnKF. The observation wells labeled #6, #7 and #8 will be used to display the performance
of the updated parameters in the forecast of piezometric heads and contaminant concentrations with the
final updated parameters. The number of observation wells could seem large; but, at this point, we wish to
test the feasibility of using the EnKF for the simultaneous source identification and conductivity estimation.
In practice, the density of observation wells may not be as large prior to contaminant detection; but, after
contamination has been detected, if the solute is dangerous, the post-detection network could become
as dense, if not denser, as the one we have used.

Other parameters—which are considered as homogeneous for both aquifers—are: porosity of 0.3 [-],
longitudinal dispersivity of 0.2 [L], transverse to longitudinal dispersivity ratio of 0.01, no molecular diffusion,
and initial contaminant concentration of zero [ML^{-3}]. Considering the transport controlling parameters as
well as the initial conditions homogeneous, not subject to identification, and perfectly known makes the
experiment less realistic than we would like, but introducing the identification of the spatially variable
conductivity together with the contaminant source information is a big leap with respect to previous works
on this subject and brings us closer to a potential contamination identification in more realistic cases.

The total simulation time is 500 [T], discretized into 100 time steps of increasing size following a geometric series with ratio 1.01. For this geometric series, the initial step is 2.93 [T]. Concentrations and piezometric heads are observed during the first 60 time steps (for a total time of 239.5 [T]).

The release of the contaminant always starts at time 30 (around the 10th time step) and ends at time 100 (around the 30th time step). In the unimodal aquifer, the release is at location \((X,Y) = (11.5, 30.5)\) [L] with a mass-loading rate of 100 [MT\(^{-1}\)]. We will refer to this case as scenario S1. In the channelized aquifer we consider two different situations: in the first one, the release happens at location \((X,Y) = (11.5, 30.5)\) in a high permeability zone; and in the second one, the release happens at location \((X,Y) = (11.5, 19.5)\) in a low permeability zone. The release load and timing is the same as for the other aquifer. We will refer to these cases as scenarios S2 and S3, respectively. (Scenarios are summarized in Table 2.)

Two ensembles of 800 logconductivity realizations are generated, one for scenario S1 and another one for scenarios S2 and S3. The random function models used for the generation of these ensembles are the same as the ones used for the generation of the corresponding reference fields. They are unconditional realizations based on their marginal distributions and the training image for S2 and S3.

Although in previous works we have shown that the impact of prior model selection is very limited [49, 50], the knowledge of the marginal distribution of logconductivity could be considered as too advantageous. For this reason, an additional scenario, referred to as S1b, is included in the Appendix. Scenario S1b is a duplicate of scenario S1 but using a non-informative uniform marginal distribution for logconductivity between -1 [ln(LT\(^{-1}\))] and 5 [ln(LT\(^{-1}\))] for the generation of the initial ensemble of 800 realizations. This scenario is not discussed within the main body of the paper. The reader is referred to the Appendix to understand the impact that a non-informative prior has in the performance of the approach.

An ensemble of 800 7-tuplets for the source parameters is also generated, which will be used for the three scenarios. Each member of this ensemble is attached to the realization with the same sequence order in the initial ensemble of logconductivity fields. Employing 800 realizations in the ensemble helps in reducing filter inbreeding and spurious correlations generally associated with small ensemble sizes [22, 23]. The values of each 7-tuplet are generated, independently, from the following uniform distributions: initial release time \(T \in [10, 90]\), release duration \(\Delta T \in [10, 110]\), mass-loading rate \(M \in [50, 140]\), and source location \((X,Y) \in [5, 15] \times [25, 35]\) for scenarios S1, and \((X,Y) \in [5, 15] \times [17, 35]\) for scenarios S2 and S3 (see Figure 1 for the areas of the prior guesses of the release locations). Notice that we have chosen intervals not centered about the reference values, so that the means of the initial ensemble are biased with respect to the reference...
values. Also notice that for the bimodal aquifer we initially allow that the source location be either in a low or high permeability area.

4. Results

Figure 4 shows the contaminant plume evolution at the 10th (30.68 [T]), 30th (102.02 [T]), 40th (143.38 [T]), and 60th (239.53 [T]) time steps in the three reference scenarios. Figure 5 shows the evolution in time of the total contaminant mass in the reference aquifers. Recall that the injection wells are near the western boundary and pumping wells are near the eastern boundary, and that the contaminant release starts around the 10th time step and stops around the 30th time step. The contaminant plume spreads from west to east at different speeds and with different spreading rates depending on the scenario: in the unimodal aquifer the solute travels more or less horizontally following the mean head gradient, and the plume does not reach the pumping wells until about time 50; in the multimodal aquifer, the plume travels quickly in the high conductivity zone and reaches the pumping wells at about the time the contaminant injections ceases, while the plume moves very little in the low conductivity zone and has not reached the pumping wells at the end of the simulation.

Figure 6 shows the ensemble mean and the ensemble variance of the updated log-conductivity fields after the 60th time step for the three scenarios. Recall that there are no conditioning data used for the generation of the initial log-conductivity fields, hence, the initial ensemble mean and initial ensemble variance are flat and the values are equal to their prior global mean and variance. After observed piezometric head and concentrations are assimilated, the ensemble mean of the updated log-conductivity fields can retrieve the main patterns of the references. The ensemble variance of scenarios S1 gets close to zero for the most part of the aquifer, whereas for scenario S2 and S3 the reduction is not as large, with still some uncertainties corresponding with the boundaries of the channels. These results, which are consistent with previous findings [23], indicate that the piezometric head information (the same for scenarios S2 and S3) is not enough to fully identify the strong channels; concentration data also carries information about the conductivity heterogeneity. In scenario S3 the plume has not traveled much and therefore the conductivity information on the concentration data is smaller than for scenario S3 and, as a result, the ensemble variances are, overall, larger than for S2.

Before the analysis of the results regarding the source parameters, we wish to analyze the reproduction of the piezometric and concentration data at well locations #6, #7 and #8 (see Fig. 3) as well as of the spatial distribution of piezometric head at the end of time step 60, and the spatiotemporal evolution of concentration.
Figures 7 and 8 show the evolution of piezometric heads and concentrations, respectively, at the three wells as computed in each member of the initial ensemble for a simulation time of 500 (100 time steps), which goes beyond the assimilation time of 240 (60 time steps). The initial realizations of conductivity are unconditional and, consequently, the spread of the evolution of heads and concentrations through the ensemble is very large.

Figures 9 and 10 show the evolution of piezometric heads and concentrations, respectively, at the same three wells, with the parameters updated after the 60th time step. We can observe that the reproduction of the piezometric heads in the updated ensemble is almost perfect, while the reproduction of the concentrations has improved considerably with respect to the predictions on the unconditional ensemble.

Figure 11 shows piezometric information at the end of time step 60. We have chosen to show the spatial distribution of heads in the reference field, in a randomly selected realization (#29), the ensemble mean of the 800 realizations and the ensemble variance. The good reproduction of the observed piezometric heads at wells #6, #7 and #8, which were used in the assimilation steps, extends for the entire aquifer as it is indicated by the almost zero variance in most of the aquifer for all scenarios, and by looking at the similarity between the heads in the reference and the heads in a given realization or the ensemble mean. We have chosen to show the results only for time step 60 and for a single realization because the results for other time steps or other realizations are virtually the same.

A similar analysis is performed for concentrations. Figures 12, 13 and 14 show the time evolution of concentration in three randomly selected realizations (#29, #537, and #695) for the three scenarios. When looking at any of the columns in these figures we are observing the variability across realizations for different time steps. While the plumes are similar, we can observe some differences for all scenarios and all time steps; yet, the shapes and topologies of the plumes when compared to the reference ones in Figure 4 are well reproduced. With the same observation wells for piezometric heads and concentrations, the reproduction of the concentrations is harder to achieve since the reference plumes are only "seen" in a small number of observation wells.

Notice also that each realization has associated a specific updated 7-tuple of contaminant source parameters, which explains why, for instance, the plume appears earlier in realization #29 than in the other realizations because the starting time associated with that realization is smaller than 10.

The concentration time evolution derived from the updated realizations after 60 time steps of data assimilation can be summarized by their ensemble mean and their ensemble variance displayed in Figures 15 and 16, respectively. As with the individual realizations, the mean map captures well the sizes and topologies of the plumes in the reference, although it misses the reproduction of the maximum values.
The largest variances appear at the locations of the reference peak concentrations. We did not show the corresponding maps computed with the initial ensemble of realizations, prior to any assimilation of data, because in that case the mean plumes spread over the entire aquifer and the variances are very large. In our examples, the ensemble Kalman filter shows again its potential to characterize the spatial heterogeneity of hydraulic conductivity with the additional complication of not knowing, at the beginning of the simulation, the time or the location of the contaminant source.

Finally, Figure 17 shows boxplots for the parameters that describe the contaminant source, i.e., \(X, Y, T, \Delta T\) and \(M\) computed at different times during the assimilation process. Starting from an unknown spatial distribution of hydraulic conductivities and an unknown contaminant source, the use of a restart NS-EnKF and the assimilation of piezometric heads and solute concentrations from a network of observation wells, we end with a reasonable characterization of the spatial variability of hydraulic conductivities (independently of whether the underlying reference follows a Gaussian —unimodal— distribution or not) and a reasonable identification of when, where and how much solute entered in the aquifer. Characterization that is duly quantified with an uncertainty derived from the analysis of the ensemble of realizations. Regarding the parameters describing the source, we can see that for almost all of them, at time step 60, the median values (red bars in the boxplots) are quite close to the reference ones. The exceptions are the initial time for S3 (release in the channeled aquifer in a low conductivity zone) and the mass-loading rate for S1 (release in the unimodal aquifer). The result for the initial time in S3 can only be interpreted in view of the overall uncertainties of all source parameters for this scenario; while they have reduced with respect to their initial uncertainties at time zero, the reduction is not very large for many of them, and this is due to the little concentration information collected (refer to Figure 4, third row, to see the evolution of the plume in the reference field with respect to the observation locations), therefore the conclusion is that more concentration data would be needed for a better characterization of the source parameters in S3. In the cases of S1, the median values of all parameters converge very quickly to the reference values and after time step 30 they do not change much except for the mass loading, for which the median value keeps increasing as time passes and the remaining uncertainty is still large. The mass loading is, for all scenarios, the parameter that consistently gets the larger uncertainty at the end of the assimilation period for all parameters and this must be attributed to the dilution of the plume concentration, which makes that the absolute differences between the observed concentrations at the wells and the predicted values are small even for relatively large changes in the mass loading at the injection point.

Each scenario behaves differently regarding the identification of the source parameters. In S1, the median
value for the \(X\) coordinate is close to the reference one, but the uncertainty is large; this large uncertainty in
\(X\) could be explained by the same reason as the large uncertainty in the mass-loading rate: the dilution of
the plume together with its longitudinal spreading makes that changes in \(X\) do not introduce large absolute
changes of concentrations at the observation locations. The contrary happens with \(Y\), its median value
converges almost exactly to the reference value at time step 30 and remains there for the rest of the simulation;
considering that transversal dispersivity is small, changes in \(Y\) will introduce noticeable modifications in the
concentrations at observation wells, making it a more sensitive parameter for identification purposes. The
median starting time \(T\) and the median injection time \(\Delta T\) converge also quickly to the reference values,
with residual uncertainties considerably smaller than their starting ones. In S2, the identification of \(Y\), \(T\)
and \(\Delta T\) by their median values is remarkable, with very little uncertainty left; the value of \(X\) is also well
identified but with larger uncertainty than the previous parameters; again, the explanation could be the
same as for S1, a small sensitivity of concentrations to fluctuations of \(X\) about its reference value. The
mass-loading rate is correctly identified by its median value, and the interquartile range is one third of its
value at the beginning of the simulation, yet, with regard to the other parameters its uncertainty is still
large; the explanation for this large uncertainty is again related to the small sensitivity of the concentrations
at the observation locations to changes of the mass loading within its uncertainty interval. Finally, in S3,
as it was already pointed out, the plume displaces very little during the 60 time steps and the amount of
information carried by the solute concentrations at the observation locations is small; yet, the algorithm
is able to identify the source coordinates \(X\) and \(Y\) by their median values, and their final uncertainty is
not very large. Notice that for S3, contrary to S2, displacing vertically the release location does not have
such a large impact in the plume movement for the duration of the assimilation period, therefore, the final
uncertainty is larger than for S2. The release duration ends also with a large uncertainty, but this is due to
the large uncertainty associated to the release initial time, and the mass-loading rate also ends with a large
uncertainty for the same reasons discussed above for S1 and S2.

5. Summary and conclusions

In this work, we have removed one of the most critical limitations of our previous work [21], in which we
demonstrated the application of the restart NS-EnKF for contaminant source identification but assuming
that the spatial heterogeneity of hydraulic conductivities was perfectly known. The joint identification of
hydraulic conductivities and contaminant source information has been performed for three different scenarios
combining unimodal and bimodal conductivity distributions (with and without channeling) and releases in
high and low conductivity zones. The results show that such an identification is possible but that the uncertainty associated with the final ensembles of realizations is still significant, especially regarding concentration spatiotemporal evolution and some of the parameters identifying the source.

The analysis was performed assuming that no information on conductivities is available except for their marginal distributions. (An additional scenario in which the marginal distribution is also unknown is included in the appendix showing the small impact that this knowledge has in the final results.) In practice, there are always some measurements of conductivity, which, if included as conditioning data, would help in improving the overall characterization as already shown elsewhere [23]. The analysis was also performed assuming that initial and boundary conditions are perfectly known, and that the transport parameters, such as porosity or dispersivity, are also perfectly known. In practice this will never be the case, and either they should be included in the identification process or they should be modeled accounting for their uncertainty.

The final uncertainties on the conductivity spatial distribution and on the contaminant source parameters should not be considered as a failure of the method to solve the problem but as a measure of the information content of the available observations. Careful analysis of these uncertainties should be used to guide a possible redesign of the sampling network.

In summary, we believe that the restart normal-score ensemble Kalman filter has the potential of helping in the joint identification of a contaminant source and the spatial heterogeneity of conductivity, and that we have advanced substantially from previous works in which the spatial heterogeneity of conductivity was not considered.

Acknowledgements Financial support to carry out this work was received from the Spanish Ministry of Economy and Competitiveness through project CGL2014-59841-P. The authors acknowledge the Associate Editor, and the anonymous reviewers for their thoughtful and constructive comments.

Appendix. Non-informative prior marginal distribution

All scenarios in the main body of the paper have been generated assuming that the marginal distribution of the logconductivities is known. However, in reality, this will not be so and there always be some uncertainty about this distribution. To evaluate the impact of this uncertainty we have run an additional scenario, which we refer as S1b, as a duplicate of scenario S1 but using a non-informative prior marginal distribution of logconductivity: a uniform distribution between -1 [$\ln(LT^{-1})$] and 5 [$\ln(LT^{-1})$]. Such an analysis in the context of just logconductivity identification has already been performed by us in the past [49, 50] showing
that the prior model selection has little impact in the performance of the NS-EnKF. A conclusion that
could be already anticipated by looking at the figures associated with the initial ensemble of unconditional
realizations in the main body of the paper.

Except for the marginal distribution used for the generation of the initial ensemble of realizations, all
remaining parameters and the different steps of the analysis are exactly the same as for S1. Next, we show
the same sets of figures as shown for scenario S1 in the body of the paper: in Figure 18 the ensemble mean
and ensemble variance of the updated lnK, in Figure 19 the time evolution of the piezometric heads and
concentrations at three well locations #6, #7 and #8 for the initial ensemble of parameters, in Figure 20
the time evolution of the piezometric heads and concentrations at the same three well locations for the
parameters updated after the 60th time step, in Figure 21 the piezometric heads at the end of the 60th time
step, in Figure 22 the evolution of the contaminant plume in the 29th, 537th, and 695th realizations with
the parameters updated after the 60th time step in Figure 23 the ensemble mean and ensemble variance of
the contaminant plume evolution with the parameters updated after the 60th time step, and in 24 the box
plots of the source parameters.

A quick comparison of any of these figures with the corresponding ones for S1 shows small changes with
a final performance of S1b very similar to S1. The final updated parameters are virtually the same in S1b
and S1. The variance of the updated logconductivities is slightly larger towards the edge of the fields for
S1b than for S1. The predicted piezometric heads and concentrations on the updated fields are almost the
same in S1 and S1b. And the boxplots for the contaminant source parameters are also very similar when
comparing S1 and S1b.

In summary, the NS-EnKF is capable to update an ensemble of initial realizations with a prior distri-
bution probability far from the true one into an ensemble of realizations that will be conditioned to the
observation data and converge towards the reproduction of the reference logconductivity field as well as
capture the parameters defining the unknown contamination source. These results are in line of previous
analysis of the NS-EnKF that proved its updating power of an initial ensemble of realizations even when
these initial ensemble is far from the reference field due to lack of information about the spatial variability
of logconductivity [49, 50].


[2] Sun, A.Y., Painter, S.L., Wittmeyer, G.W.. A constrained robust least squares approach for contam-
iniant release history identification. Water resources research 2006;42(4).


[40] McDonald, M., Harbaugh, A.. A modular three-dimensional finite-difference ground-water flow model 1988;.


Table 1: Parameters of the random functions describing the heterogeneity of $\ln K$ for the two aquifers. $\lambda_x$ and $\lambda_y$ are the ranges in the $x$ and $y$ directions of an exponential covariance.

<table>
<thead>
<tr>
<th>$\ln K$</th>
<th>Facies</th>
<th>Proportion</th>
<th>Mean</th>
<th>Std. dev</th>
<th>Variogram type</th>
<th>$\lambda_x$</th>
<th>$\lambda_y$</th>
<th>sill</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unimodal aquifer</td>
<td></td>
<td>2.0</td>
<td>1.5</td>
<td></td>
<td>exponential</td>
<td>20</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>Channelized aquifer</td>
<td>Channel</td>
<td>0.35</td>
<td>1.0</td>
<td>1.0</td>
<td>exponential</td>
<td>20</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Non-channel</td>
<td>0.65</td>
<td>-2.5</td>
<td>0.6</td>
<td>exponential</td>
<td>20</td>
<td>20</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Table 2: Definition of scenarios.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unimodal aquifer</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Channelized aquifer</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Source location (11.5, 30.5)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source location (11.5, 19.5)</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: Scenarios S1-S3. Reference fields of $\ln K$. The black circle denotes the source location. The black quadrilateral is the suspect area to be used in the identification step.
Figure 2: Training image.

Figure 3: Distribution of wells. Red triangles are measurement wells; blue diamonds are injection wells near the west boundary and pumping wells near the east boundary.
Figure 4: Scenarios S1-S3. Contaminant plume evolution at the 10th (beginning of solute injection), 30th (end of solute injection), 40th and 60th time steps in the reference aquifers. White triangles mark the measurement wells.

Figure 5: Scenarios S1-S3. Time evolution of the total solute mass in the aquifer.
Figure 6: Scenarios S1-S3. Updated ensemble mean (top row) and updated ensemble variance (bottom row) of lnK after the 60th time step.

Figure 7: Scenarios S1-S3. Time evolution of the piezometric heads at the three wells #6, #7 and #8 for the initial ensemble of source information parameters and lnK. The red line is the evolution of the piezometric head in the reference. The black lines correspond to the 5 and 95 percentiles of all realizations, and the green line corresponds to the median. The vertical dashed lines mark the end of the assimilation period.
Figure 8: Scenarios S1-S3. Same caption as previous figure but regarding solute concentration.
Figure 9: Scenarios S1-S3. Same caption as previous figure but regarding piezometric heads computed on the updated ensembles after the 60th assimilation time step.
Figure 10: Scenarios S1-S3. Same caption as previous figure but regarding solute concentrations computed on the updated ensembles after the 60th assimilation time step.
Figure 11: Piezometric heads at the end of the 60th time step for the three scenarios. From top to bottom, scenarios S1, S2 and S3. From left to right, heads in the reference aquifer; heads in realization #29; ensemble mean, and ensemble variance. White triangles mark the measurement wells. Note that the scale bar for S1 is different from that for S2 and S3.
at the 10th, 30th, 40th and 60th time steps with the parameters updated after the 60th time step.

Figure 12: Scenario S1. Contaminant plume evolution of the 29th (top row), 537th (middle row), 695th (bottom row) realizations at the 10th, 30th, 40th and 60th time steps with the parameters updated after the 60th time step.

Figure 13: Scenario S2. Contaminant plume evolution of the 29th (top row), 537th (middle row), 695th (bottom row) realizations at the 10th, 30th, 40th and 60th time steps with the parameters updated after the 60th time step.
Figure 14: Scenario S3. Contaminant plume evolution of the 29th (top row), 537th (middle row), 695th (bottom row) realizations at the 10th, 30th, 40th and 60th time steps with the parameters updated after the 60th time step.

Figure 15: Scenarios S1-S3. Ensemble mean of contaminant plume evolution at the 10th, 30th, 40th and 60th time steps with the parameters updated after the 60th time step.
Figure 16: Scenarios S1-S3. Ensemble variance of contaminant plume evolution at the 10th, 30th, 40th and 60th time steps with the parameters updated after the 60th time step.

Figure 17: Scenarios S1-S3. Box plots of the source location coordinates (X and Y), initial release time (T), release duration (ΔT), and mass-loading rate (M) at the initial, 10th (30.68 [T]), 20th (64.57 [T]), 30th (102.02 [T]), 40th (143.38 [T]), 50th (189.06 [T]) and 60th (239.53 [T]) time steps. The dashed horizontal black line corresponds to the reference value.
Figure 18: Scenario S1b. Updated ensemble mean (left column) and updated ensemble variance (right column) of lnK after the 60th time step.

Figure 19: Scenario S1b. Time evolution of the piezometric heads (top row) and solute concentrations (bottom row) at the three wells #6, #7 and #8 for the initial ensemble of source information parameters and lnK. The red line is the evolution of the piezometric head in the reference. The black lines correspond to the 5 and 95 percentiles of all realizations, and the green line corresponds to the median. The vertical dashed lines mark the end of the assimilation period.
Figure 20: Scenario S1b. Same caption as previous figure but regarding piezometric heads and concentrations computed on the updated ensembles after the 60th assimilation time step.

Figure 21: Scenario S1b. Piezometric heads at the end of the 60th time step. From left to right, heads in the reference aquifer; heads in realization #29; ensemble mean, and ensemble variance. White triangles mark the measurement wells.
Figure 22: Scenario S1b. Contaminant plume evolution of the 29th (top row), 537th (middle row), 695th (bottom row) realizations at the 10th, 30th, 40th and 60th time steps with the parameters updated after the 60th time step.

Figure 23: Scenarios S1b. Ensemble mean (top row) and ensemble variance (bottom row) of contaminant plume evolution at the 10th, 30th, 40th and 60th time steps with the parameters updated after the 60th time step.
Figure 24: Scenarios S1b. Box plots of the source location (X and Y), initial release time (T), release duration (∆T), and mass-loading rate (M) at the initial, 10th, 20th, 30th, 40th, 50th and 60th time steps. The dashed horizontal black line corresponds to the reference value.