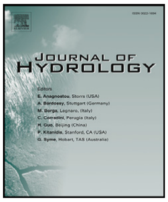
Journal of Hydrology 615 (2022) 128642

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol



Research papers

Ensemble random forest filter: An alternative to the ensemble Kalman filter for inverse modeling

Vanessa A. Godoy *, Gian F. Napa-García, J. Jaime Gómez-Hernández

Research Institute of Water and Environmental Engineering, Universitat Politècnica de València, 46022, Valencia, Spain

A R T I C L E I N F O

This manuscript was handled by Corrado Corradini, Editor-in-Chief, with the assistance of Saket Pande, Associate Editor.

Keywords:
Groundwater flow
Inverse modeling
Random forest
Bayesian methods

A B S T R A C T

The ensemble random forest filter (ERFF) is presented as an alternative to the ensemble Kalman filter (EnKF) for inverse modeling. The EnKF is a data assimilation approach that forecasts and updates parameter estimates sequentially in time as observations are collected. The updating step is based on the experimental covariances computed from an ensemble of realizations, and the updates are given as linear combinations of the differences between observations and forecasted system state values. The ERFF replaces the linear combination in the update step with a non-linear function represented by a random forest. This way, the non-linear relationships between the parameters to be updated and the observations can be captured, and a better update produced. The ERFF is demonstrated for log-conductivity identification from piezometric head observations in several scenarios with varying degrees of heterogeneity (log-conductivity variances going from 1 up to 6.25 (ln m/d)^2), number of realizations in the ensemble (50 or 100), and number of piezometric head observations (18 or 36). In all scenarios, the ERFF works well, reconstructing the log-conductivity spatial heterogeneity while matching the observed piezometric heads at selected control points. For benchmarking purposes, the ERFF is compared to the restart EnKF to find that the ERFF is superior to the EnKF for the number of ensemble realizations used (small in typical EnKF applications). Only when the number of realizations grows to 500 the restart EnKF can match the performance of the ERFF, albeit at more than double the computational cost.

1. Introduction

Characterization of the subsurface heterogeneity is of critical concern for modeling groundwater flow (i.e., Capilla et al., 1999; Li et al., 2011; Feyen et al., 2003; Fernàndez-Garcia and Gómez-Hernández, 2007) since it requires heterogeneous values of hydrogeologic parameters, which commonly are only sparsely available, if at all. To overcome the incomplete knowledge of the system and obtain better predictions with numerical models, state variables such as piezometric head—generally more extensively sampled—can be assimilated to improve the characterization of harder-to-measure parameters such as hydraulic conductivity (Carrera et al., 2005; Wen et al., 1999). Even with such an improvement, parameter heterogeneity is never completely known, and its uncertainty also needs to be characterized.

Stochastic data assimilation is an inverse modeling approach that can be used to characterize parameter heterogeneity and its uncertainty by assimilating state data sequentially in time (Zhou et al., 2014). The ensemble Kalman filter (EnKF) proposed by Evensen (1994) is a very popular data assimilation method for stochastic inverse modeling that has been proven very efficient in numerous applications in fields as varied as atmospheric science, oceanography, geophysics, geotechnical and petroleum engineering, hydrology, or hydrogeology (Yin et al., 2015; Xu and Gómez-Hernández, 2016; Shuai et al., 2016; Zhu et al., 2017; Chen et al., 2018; Liu et al., 2018; Gelsinari et al., 2020; Kim et al., 2020; He et al., 2021).

Data assimilation for inverse modeling, as implemented by the EnKF and its variants, is based on two main steps, a forecast of system evolution followed by an update (or correction) of the parameters describing the system based on the discrepancy, at a few locations, between predictions and observations. The updates are computed using linear combinations with the weights calculated using covariance functions in a manner very similar to the geostatistical interpolation technique of cokriging. Such a linear scheme is a drawback of the Kalman-based data assimilation methods since it is optimal only when the system evolves in time following a linear state equation. Still, when the system evolves non-linearly, the model is suboptimal, although its performance may be very good, as demonstrated by its successful applications. A typical example of an EnKF implementation in which the relationship between the parameters and the state is non-linear is for inverse groundwater modeling (Evensen, 1994; Xu et al., 2013).

* Corresponding author.
E-mail address: vaalde1@iiama.upv.es (V.A. Godoy).

https://doi.org/10.1016/j.jhydrol.2022.128642
Received 29 May 2022; Received in revised form 5 September 2022; Accepted 12 September 2022
Available online 5 November 2022

One of the reasons for the success of the EnKF is that the experimental covariances are computed from ensembles of realizations that contain parameter values and their corresponding predictions. The ensemble size is critical; it should be as small as possible to save CPU time, but it should be as large as possible to obtain good experimental covariance estimates that will prevent filter inbreeding and the appearance of spurious correlations and avoid filter divergence. (These problems could be mitigated for small ensemble sizes with covariance localization techniques (Chen and Zhang, 2006; Todaro et al., 2019; Xu et al., 2013).)

Chen and Zhang (2006) studied the sensitivity of the EnKF to, among other factors, the ensemble size and the choice of the initial ensemble, and they showed that prior knowledge of the underlying field, such as the structure of the covariance function, plays an important role in data assimilation. Besides that, they found that a correct estimation of uncertainty may require large ensemble sizes. The need for large ensemble sizes and good prior knowledge of the spatial variability of the field, the linear nature of the updating step, and its big computational cost call for new strategies to improve available data assimilation ensemble methods.

In the last years, machine learning and big data are permeating all ambits of science and technology. The easiness with which large amounts of data are acquired in real-time and the new approaches to process them to build data-based predictive models have given rise to a new paradigm in the treatment of information that is starting to be used in environmental and water resources studies (Asher et al., 2015; Sit et al., 2020; Tahmasebi and Sahimi, 2021; Mariethoz and Gómez-Hernández, 2021). In groundwater modeling, machine learning algorithms have been used mainly to replace process-driven models with data-driven ones to predict piezometric heads or solute concentrations from ancillary variables. The justification is that the data-driven models are cheaper to run and may capture relationships that could escape a process-driven analysis (Knoll et al., 2019; Al-Abadi and Alsamaani, 2020; Nguyen et al., 2020; Sachdeva and Kumar, 2021; An et al., 2021). Although these algorithms have proven their ability to deal with a wide range of problems in groundwater, they are seldom used for stochastic inverse modeling purposes. To the best of the authors' knowledge, it has not yet been used as a data assimilation algorithm capable of replacing the restart EnKF (r-EnKF) (Chen et al., 2018; Xu and Gómez-Hernández, 2018, 2016). Without trying to be exhaustive, some example applications of machine learning in groundwater inverse modeling are the works by Mo et al. (2019), who combined an autoregressive neural network-based surrogate method for forward modeling with an iterative local updating ensemble smoother (ILUES) (Zhang et al., 2018) to solve high-dimensional contaminant transport inverse problems; (Bao et al., 2020, 2022), who used Generative Adversarial Networks (GAN) (Goodfellow et al., 2014) to reparameterize hydraulic conductivity, using a low dimension latent variable, and then coupling it to an ensemble smoother with multiple data assimilation (ES-MDA) Emerick and Reynolds (2013); or Zhang et al. (2020), who used deep learning to improve the ensemble smoother, although their starting ensemble was built with good prior knowledge of the underlying hydraulic conductivity spatial heterogeneity.

Since the weakest point of the Kalman-based data assimilation methods is the linear updating step, which is equivalent to cokrige the perturbations of hydraulic conductivity from the deviations between predicted and observed piezometric heads, it is proposed to replace the covariance-based updating step with a random forest-based updating. Random forest updating should be able to capture the multipoint non-linear relationships between conductivities and piezometric heads. This new method is termed ensemble random forest filter (ERFF). The idea of using random forests (Breiman, 2001) was inspired by the work by Hengl et al. (2018) in which the authors propose, as an alternative to kriging, a new framework for spatial interpolation using random forest, demonstrating that this approach is capable of capturing relationships that go beyond the linear correlation intrinsic to the covariance. The framework proposed by Hengl et al. (2018) seeks the (non-linear) interpolation of an attribute from sparsely observed attribute values. In ERFF, however, the task is to interpolate piezometric head deviations (between observed and predicted values) to provide correction increments for hydraulic conductivity over the entire aquifer model. By taking advantage of the ensemble of realizations and subtracting them two by two, a new set of realizations (an order of magnitude larger) is built to train the random forest. Finally, the ERFF replaces the calculation and inversion of covariance matrices with random forest training.

The ERFF is demonstrated in three synthetic aquifers of varying heterogeneity (variances ranging from 1.0 to 6.25 (ln (m/d))^2). A sensitivity analysis of the ensemble size and the number of observations is carried out. Differently from previous researchers (Mo et al., 2019; Goodfellow et al., 2014; Zhang et al., 2020) and in line with the work by Xu et al. (2013), it is assumed that there is no prior information about the spatial heterogeneity of hydraulic conductivity, but only information about its mean value and its variance. Xu et al. (2013) have already shown the power of transient piezometric heads in the characterization of hydraulic conductivity by the EnKF when no prior information is available. As will be shown, this power is intrinsic and can be taken advantage of by the ERFF. The concept of localization (Xu et al., 2013; Todaro et al., 2019) is also included in the implementation of the ERFF to reinforce the notion of spatial correlation by training the random forest giving more weight to the observations that are closer to the point being updated. The ERFF results are benchmarked against the r-EnKF.

The structure of this paper is as follows. First, the basics of ensemble Kalman filtering are introduced, followed by describing how the EnKF becomes the ERFF. Second, the three reference synthetic transient groundwater flow problems and the scenarios that will be analyzed are described. Third, the results for the different scenarios are shown, and one of the scenarios is compared with the r-EnKF. And fourth, the paper ends with a summary and an outlook on potential lines of continuing research.

2. Stochastic data assimilation

The EnKF algorithm (Evensen, 1994) is the evolution of the Kalman filter (Kalman, 1960) to handle nonlinear state transfer functions by using a Monte-Carlo approach. The EnKF (in the context of inverse modeling) is a sequential data assimilation method that updates model parameters based on the discrepancies between model predictions and experimental observations at a few locations. The relationship between parameters and observations must be known, and a forward model relating parameters and state variables must be available. In the original implementation of the EnKF for inverse modeling, both model parameters and system states were updated. Still, it was found that the updated states might violate constitutive relationships (such as mass conservation), and the restart EnKF was introduced, whereby only model parameters are updated, and the forecast for the next time step is always performed from time zero. The reader interested in the details of the EnKF is referred to the many papers published, particularly those by Evensen (1994, 2003). In the following, a brief description of the r-EnKF is presented to introduce the ERRF.

2.1. r-EnKF: Ensemble data assimilation with covariance-based updating

Consider a transient groundwater flow model in which piezometric heads are predicted based on the hydraulic conductivity values on a discretized aquifer (plus corresponding boundary, initial conditions, and forcing terms). The forward model relating them is

$$\mathbf{y}(t) = g(\mathbf{x}, \mathbf{y}(t - \Delta t)), \tag{1}$$

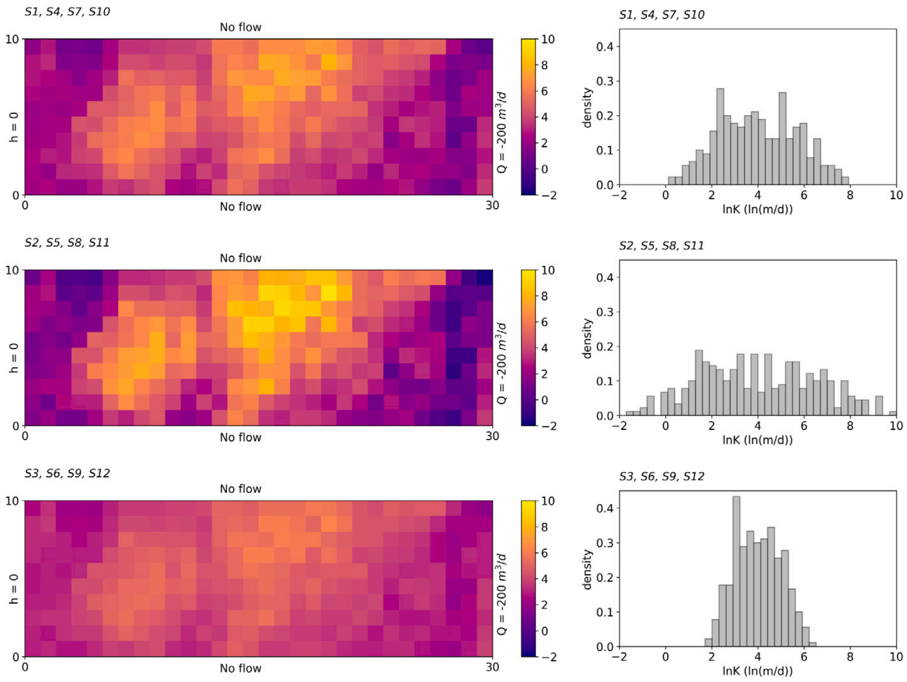


Fig. 1. Reference fields and corresponding histograms.

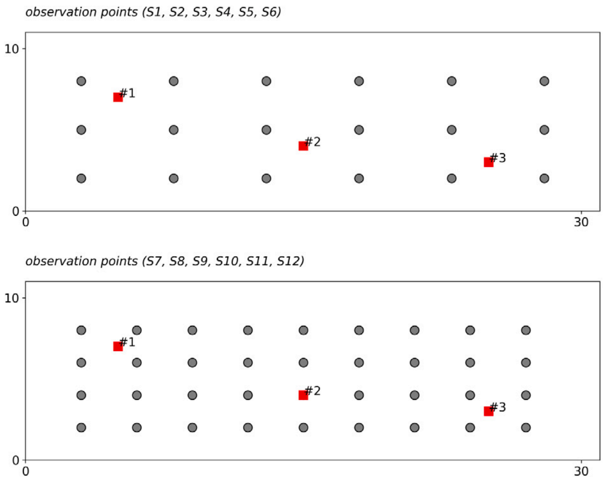

observation points (S1, S2, S3, S4, S5, S6)
observation points (S7, S8, S9, S10, S11, S12)
#1
#2
#3
0
10
30

Fig. 2. Observation (circle) and control (square) points.

where t is time, $\mathbf{x} \in \mathbb{R}^{n_p}$ is the hydraulic conductivity, $\mathbf{y} \in \mathbb{R}^{n_o}$ is the predicted system state at measurement locations, $g(\cdot)$ is a function that includes the numerical flow model plus an observation operator that extracts the predictions at observation locations, n_p is the number of cells in which the aquifer has been discretized and for which the hydraulic conductivity needs to be known to solve the numerical flow equation, and n_o are the number of piezometric head observation locations. Piezometric heads are collected sequentially in time, and the purpose of the r-EnKF is, after each data collection, to update the hydraulic conductivities so that after a sufficient number of updates, the hydraulic conductivity spatial distribution resembles the true but unknown one. The r-EnKF consists of an initialization step followed by repeated forecast and update steps as follows:

1. Initialization step. An initial ensemble of n_e realizations of hydraulic conductivity $\mathbf{X}^{ini}$ is generated using statistical or geostatistical methods and incorporating as much prior knowledge as possible. In this paper, it is assumed that no prior information about the spatial variability of hydraulic conductivity is available. The initial set of realizations is made up of homogeneous realizations, each with a value drawn from a univariate distribution.
2. Forecast step from time zero. In this step, the transient groundwater flow forward model is solved, from time zero, for each realization i, to obtain model predictions of the piezometric heads at time step t using the latest update of the conductivities (for the first update, the initial ensemble of conductivities is used). (Recall that to ensure mass conservation is not violated by the piezometric heads at time t, the simulation is always restarted from time zero.)

$$\mathbf{y}_{i,t} = g(\mathbf{x}_{i,t-1}, \mathbf{y}_{i,0}), \qquad i = 1, \dots, n_e, \tag{2}$$

where $\mathbf{y}_{i,t}$ is the vector of forecasted piezometric heads at the tth time step, and $\mathbf{x}_{i,t-1}$ is the last update of hydraulic conductivities at the previous time step $(t-1)$. For the first time step, $\mathbf{x}_{i,t-1}$ is $\mathbf{x}_i^{ini}$.

3. Update step. The vector of hydraulic conductivities is updated based on the discrepancies between forecasted and observed piezometric heads. The updated parameter vector $\mathbf{x}^u$ is given, for the ith realization at the tth time step, by

$$\mathbf{x}^u_{i,t} = \mathbf{x}^f_{i,t} + \mathbf{K}_t \left[\mathbf{y}^o_t + \varepsilon^o_{i,t} - \mathbf{y}^f_{i,t} \right], \tag{3}$$

where the subscripts i and t refer to a specific realization and time step, respectively; $\mathbf{x}^f_{i,t=1} = \mathbf{x}^{ini}_i$ and $\mathbf{x}^f_{i,t} = \mathbf{x}^u_{i,t-1}$, $\mathbf{y}^f_{i,t}$ is the vector of model predictions at observation locations; $\mathbf{y}^o_t$ is the vector of state values at observation locations; $\varepsilon^o_{i,t}$ is the vector of observation errors (the observations errors have zero mean and a covariance matrix $\mathbf{R}_t$); and $\mathbf{K}_t$ is the Kalman gain matrix, given by

$$\mathbf{K}_t = \mathbf{C}^t_{XY} \left(\mathbf{C}^t_{YY} + \mathbf{R}_t \right)^{-1}, \tag{4}$$

where $\mathbf{C}^t_{YY}$ is the auto-covariance of the state variables and $\mathbf{C}^t_{XY}$ is the cross-covariance between parameters and state variables for the tth time step, which are computed from the ensemble of

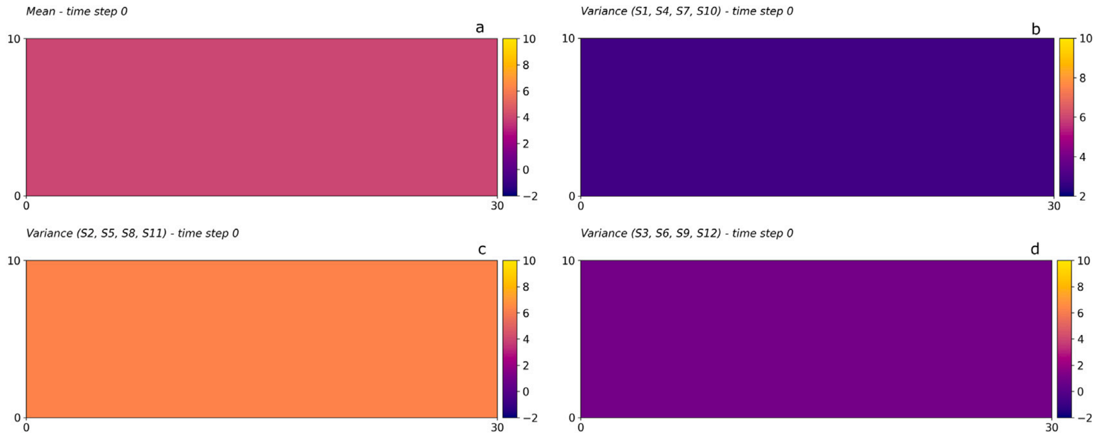


Fig. 3. Mean and variances of the initial ensembles.

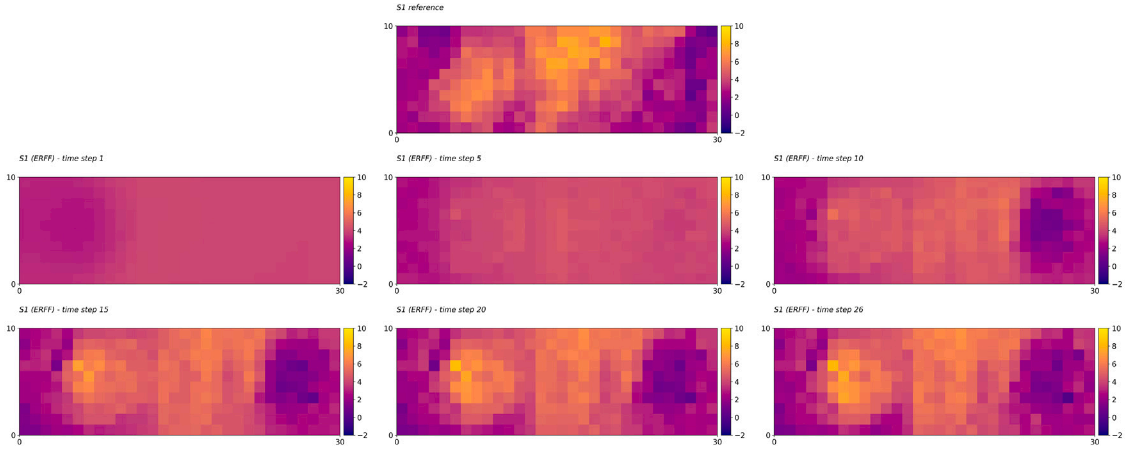


Fig. 4. Evolution in time of the ensemble mean for scenario S1.

realizations as

$$\mathbf{C}^t_{YY} = \frac{1}{n_e - 1} \sum_{i=1}^{n_e} \left(\mathbf{y}_{i,t} - \overline{\mathbf{y}}_t\right)\left(\mathbf{y}_{i,t} - \overline{\mathbf{y}}_t\right)^{\mathrm{T}}, \tag{5}$$

$$\mathbf{C}^t_{XY} = \frac{1}{n_e - 1} \sum_{i=1}^{n_e} \left(\mathbf{x}_{i,t} - \overline{\mathbf{x}}_t\right)\left(\mathbf{y}_{i,t} - \overline{\mathbf{x}}_t\right)^{\mathrm{T}}, \tag{6}$$

with $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ being the ensemble means of parameters and predictions, respectively.

Covariance localization is used to mitigate the problem of spurious correlations. It is done by element-wise multiplication of the originals covariance matrices and a distance-dependent correlation function that reduces the correlations between points as the Euclidian distance between them increases. The cross-covariance and the auto-covariance are then calculated as

$$\tilde{\mathbf{C}}^t_{XY} = \mathbf{C}^t_{XY} \circ \lambda \tag{7}$$

$$\tilde{\mathbf{C}}^t_{YY} = \mathbf{C}^t_{YY} \circ \lambda, \tag{8}$$

where ∘ represents the Schur product, and λ is a correlation function, given by:

$$\lambda(r) = \begin{cases} -\frac{1}{4}\left(\frac{r}{a}\right)^5 + \frac{1}{2}\left(\frac{r}{a}\right)^4 + \frac{5}{8}\left(\frac{r}{a}\right)^3 - \frac{5}{3}\left(\frac{r}{a}\right)^2 + 1, & 0 \leqslant \mathrm{r} \leqslant \mathrm{a}; \\ \frac{1}{12}\left(\frac{r}{a}\right)^5 - \frac{1}{2}\left(\frac{r}{a}\right)^4 + \frac{5}{8}\left(\frac{r}{a}\right)^3 & \\ \quad + \frac{5}{3}\left(\frac{d}{a}\right)^2 - 5\left(\frac{r}{a}\right) + 4 - \frac{2}{3}\left(\frac{r}{a}\right)^{-1}, & \mathrm{a} \leqslant \mathrm{r} \leqslant 2\mathrm{a}; \\ 0 & \mathrm{r} > 2\mathrm{a}, \end{cases} \tag{9}$$

where a is the distance beyond which no spatial correlation is expected, and r is the Euclidean distance between the observation and the point where log-conductivity has to be updated.

4. Back to the forecast step.

In a problem where there are n_p parameters (in our case, n_p will be the number of cells in the numerical model) and n_o observations, vectors $\mathbf{x}^u_{i,t}$ and $\mathbf{x}^f_{i,t}$ have sizes $n_p \times 1$, vectors $\mathbf{y}^o_t$, $\varepsilon^o_{i,t}$, and $\mathbf{y}^f_{i,t}$ have sizes $n_o \times 1$, the Kalman gain $\mathbf{K}_t$ and the cross-covariance $\tilde{\mathbf{C}}^t_{XY}$ are matrices of size $n_p \times n_o$, and the matrices $\tilde{\mathbf{C}}^t_{YY}$ and $\mathbf{R}$ are of size $n_o \times n_o$. When the observation errors are modeled as uncorrelated, $\mathbf{R}_t$ is a diagonal matrix. In the covariance matrix calculations, $\overline{\mathbf{x}}_t$ is a column vector of size $n_p \times 1$ with the average values of each parameter computed through the realizations, $\overline{\mathbf{x}}_t = \frac{1}{n_e}\sum_{i=1}^{n_e} \mathbf{x}_{i,t}$, and, similarly $\overline{\mathbf{y}}_t$ is a column vector of size $n_o \times 1$ with the average values of each state variable computed through the ensemble of realizations, $\overline{\mathbf{y}}_t = \frac{1}{n_e}\sum_{i=1}^{n_e} \mathbf{y}_{i,t}$.

2.2. ERFF: Ensemble data assimilation with random forest-based updating

The ERFF proposal is to replace the linear updating in Eq. (4) with a non-linear update based on a random forest prediction. Eq. (4) can be rearranged as follows

$$\mathbf{x}^u_{i,t} - \mathbf{x}^f_{i,t} = \mathbf{K}_t\left[\mathbf{y}^o_t - \mathbf{y}^f_{i,t} + \boldsymbol{\varepsilon}^o_{i,t}\right], \tag{10}$$

and rewritten as

$$\Delta\mathbf{x}_{i,t} = \varphi(\Delta\mathbf{y}_{i,t}), \tag{11}$$

where $\Delta\mathbf{x}_{i,t}$ and $\Delta\mathbf{y}_{i,t}$ are the correction (to be applied to the current estimate of the parameters) and the discrepancy (between state predictions

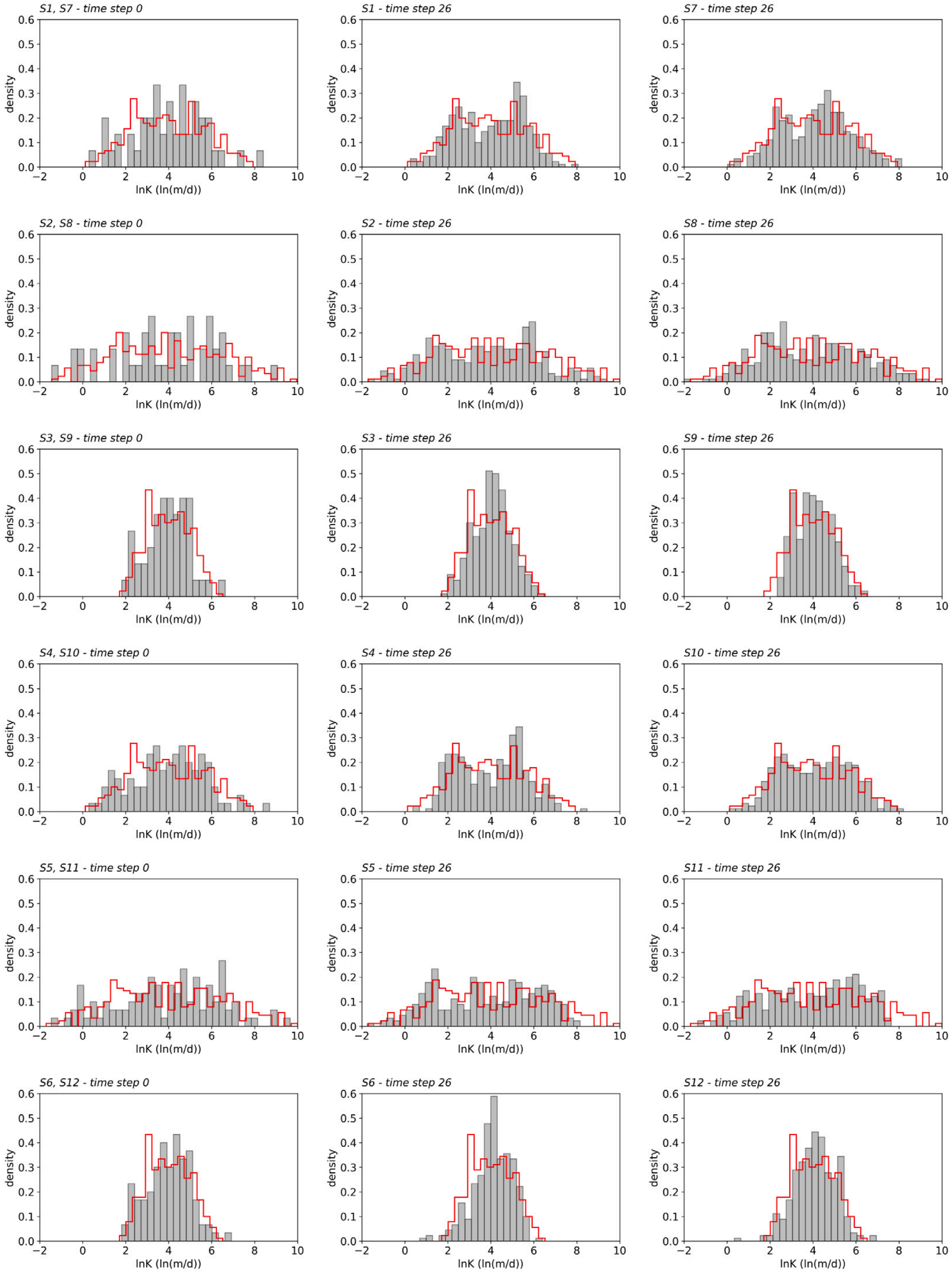


Fig. 5. Left column: Initial log-conductivity histograms. Central and right columns: Final log-conductivity histograms, for 18 and 36 observation scenarios respectively. The hollow red histograms correspond to the reference fields. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and observations) vectors, respectively. In the r-EnKF, function φ is a linear combination of discrepancies, where $\mathbf{K}_t$ (Eq. (4)) is the matrix of coefficients. In the ERFF, φ will be replaced by a random forest regressor, which should be able to capture any linear or non-linear relationship existing between $\Delta\mathbf{x}_{i,t}$ and $\Delta\mathbf{y}_{i,t}$.

Random forest regression is a supervised machine learning algorithm for building a predictor ensemble with a set of decision trees (that is, a forest) that grow in bootstrapped sub-samples of the dataset (that is, randomly selected samples with replacement). Predictions are obtained by aggregating the various predictors from each decision tree into a single average value (Breiman, 2001; Cutler et al., 2012; Biau, 2012). The bootstrap aggregation procedure used in random forest produces robust and highly accurate predictions without overfitting (Biau, 2012; Hengl et al., 2018). As the mathematical framework of the random forest itself is not the focal point of this work, interested readers are encouraged to refer to Breiman (2001), Cutler et al. (2012), and Biau (2012) for a more in-depth analysis of the technique.

A random forest has to be built for each cell in the model where log-conductivity is to be estimated. Once built, the discrepancies between forecasted piezometric heads (different for each realization) and observed values are fed to the random forest to provide an estimate of the log-conductivity perturbation to apply, at that specific cell, to each realization. The ERFF consists of the same steps as the r-EnKF: an

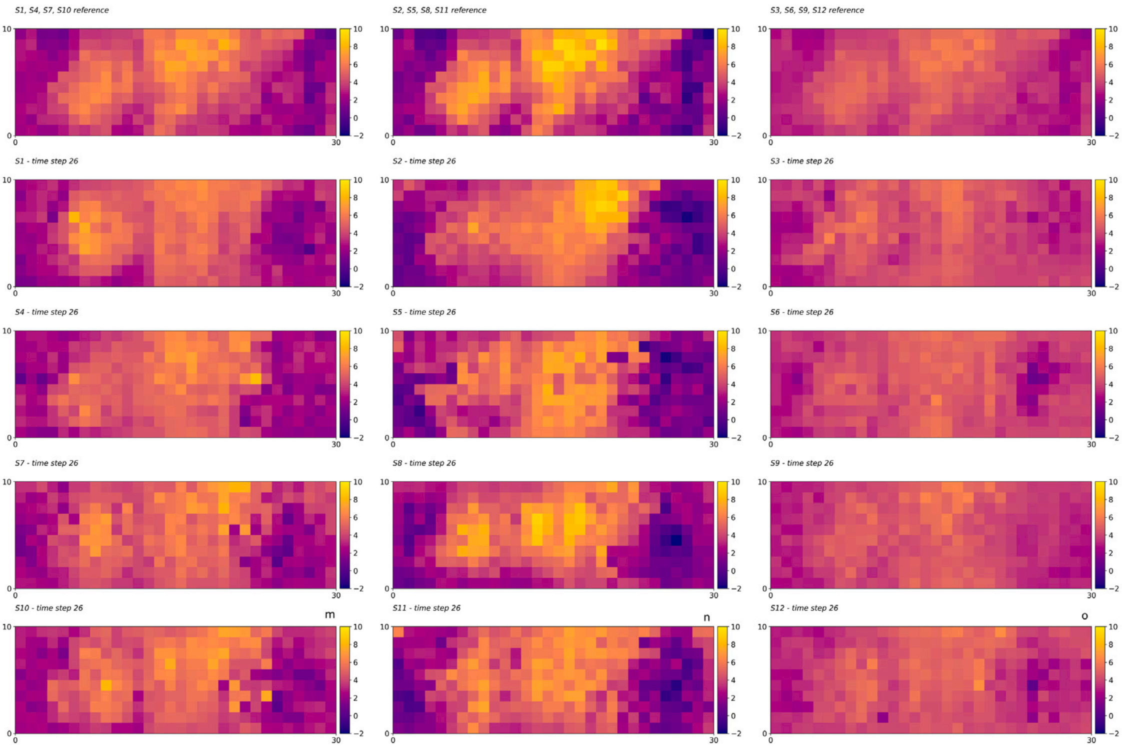


Fig. 6. Reference fields (top row) and log-conductivity ensemble mean at time step 26 for the different scenarios.

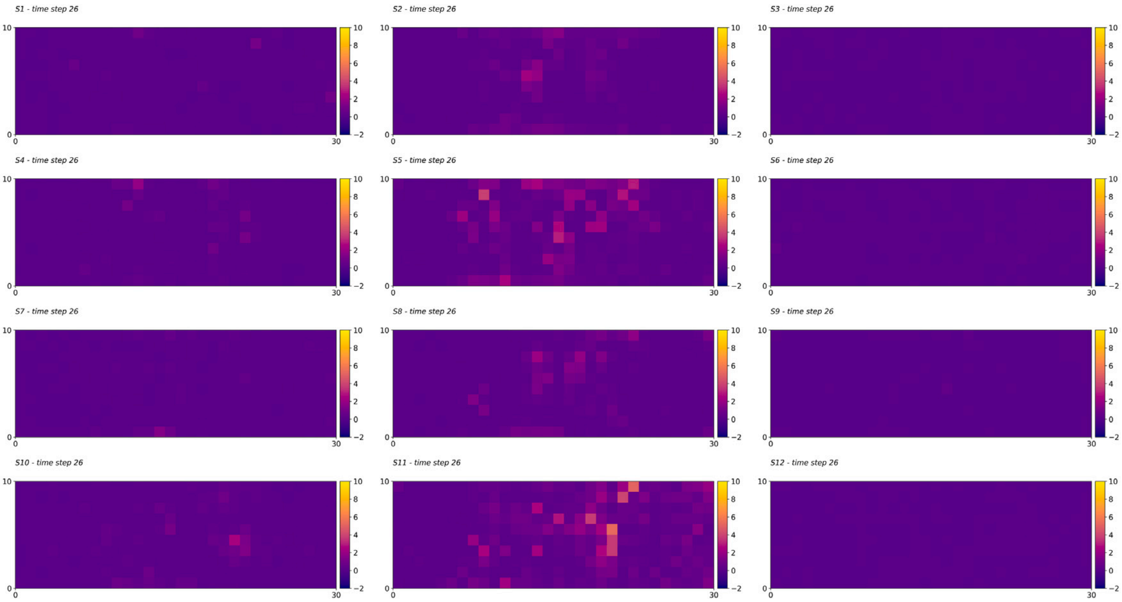


Fig. 7. Log-conductivity ensemble variance at time step 26 for the different scenarios.

initialization step followed by repeated forecast and update steps. The difference lies in the update step, which is done using random forests, as explained next. Consider a set of n_e realizations of the hydraulic conductivity and the associated n_e realizations of the piezometric heads at a given time step. With such a set, subtracting two by two each conductivity realization and its associated piezometric heads, an ensemble of $n'_e = n_e(n_e-1)/2$ realizations of differences can be built

$$\left.\begin{aligned}\Delta \ln \mathbf{K}_{i_3,t} &= \ln \mathbf{K}_{i_2,t} - \ln \mathbf{K}_{i_1,t}\\ \Delta \mathbf{h}_{i_3,t} &= \mathbf{h}_{i_2,t} - \mathbf{h}_{i_1,t}\end{aligned}\right\}$$

$$i_1 = 1,\dots,n_e-1, i_2 < i_1 \le n_e, i_3 = 1,\dots,n'_e \qquad (12)$$

where $\Delta \ln \mathbf{K}_{i_3,t}$ is a realization of log-conductivity differences at time step t, and $\Delta \mathbf{h}_{i_3,t}$ is a realization of piezometric head differences at the same time step and for the same conductivity realizations used to obtain the log-conductivity difference. Next, consider that observations have been taken at a subset of n_o locations. These observations will depart from the forecasted values, and the differences between observations and forecasts will change for each realization of log-conductivity. Consider now a specific cell in the numerical model, j; from the ensemble of differences, it is possible to build a training data set composed of

$$\left.\begin{array}{r}\Delta \ln K_{i,j,t}\\ \Delta h_{i,k,t}, k = 1,\dots,n_o\end{array}\right\} i = 1\dots,n'_e \qquad (13)$$

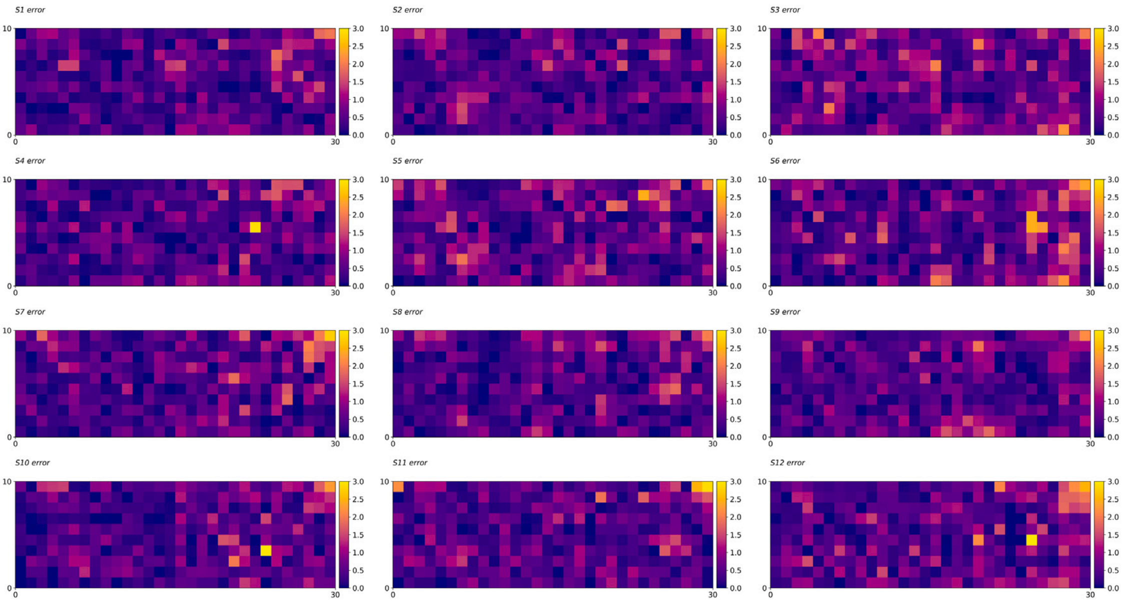


Fig. 8. Standardized deviations between the reference fields and the ensemble means at time step 26 for the different scenarios.

from which to train a random forest to predict the perturbation of log-conductivity at location j associated with perturbations of the piezometric heads at the set of n_o locations. Once this random forest is trained, the differences between the observed heads and the predicted ones in each realization are calculated. The random forest is used to predict a log-conductivity difference to apply to the current value of log-conductivity at that specific location. This procedure is repeated for each cell in the aquifer until all conductivity values are updated. This procedure could be extended for the update of multiple parameters given observations of multiple variables, such as, for instance, updating log-conductivities and porosities from piezometric head and concentration observations.

To reinforce the need to account for spatial correlation, the head differences are weighted before their use according to

$$\Delta' h_{i,k,t} = \Delta h_{i,k,t}\lambda^{-1}(r) \tag{14}$$

where i is the realization index, k is the observation index, t is the time index, r is the Euclidean distance between the observation and the point where log-conductivity has to be updated, and λ is the localization function in Eq. (9).

The rationale for using Eq. (9) here is the following: when the observation location is close to the log-conductivity location being updated, λ is close to one, and no correction is introduced, but when the head difference is far from the log-conductivity, the value of λ is close to zero, and the head difference is amplified in a way that the random forest will interpret that there is no relationship between head differences and log-conductivity differences. In this way, head differences close to the point being updated will receive larger weight in the log-conductivity update than head differences that are further apart.

The random forest was implemented using the scikit-learn library in Python (Pedregosa et al., 2011). Before running the different scenarios described in the next section, tuning the algorithm's hyperparameters was necessary. This is probably the most tedious part of the ERFF, which is always subject to some subjective decisions and is an application-dependent task. Several preliminary runs were performed, splitting the ensemble of differences into two subsets, 90% for training and 10% for validation, and a sensitivity analysis was performed to derive the best hyperparameter values. The values finally chosen for the hyperparameters were: number of trees in the forest 120, minimum number of samples required to split an internal node 2, minimum number of samples required to be at a leaf node 3, number of features to consider 0.65, and random state 10. All other hyperparameters were set at their default values as defined in scikit-learn.

It is important to stress one of the advantages of the ERFF over the EnKF is that to get the n_e ensemble of realizations is necessary to run n_e times the forward model, but then, the number of realizations to train the random forest increases to $n_e(ne-1)/2$ after a simple subtraction of the original n_e realizations. To get the same number of realizations for EnKF, $n_e(ne-1)/2$ would have to be forward modeled.

3. Synthetic examples

Three synthetic, two-dimensional, heterogeneous, and confined aquifers are built on a domain composed of 30 by 10 cells, each 1 m by 1 m. The GCOSIM3D code (Gómez-Hernández and Journel, 1993) was used to generate the three reference log-conductivity fields with standard deviations (SD) of 1.0, 1.7, and 2.5 ln (m/d), and all of them with a mean of 4.0 ln (m/d) and a spherical variogram with maximum and minimum ranges of 20 and 10 m, respectively, with the direction of maximum continuity oriented at 60° counterclockwise with respect to the east–west axis. Transient groundwater flow is simulated in all three synthetic aquifers under the following conditions: north and south boundaries are impervious; along the east boundary, a flow of -200 m^3/d is prescribed; heads of 0 m are prescribed along the west boundary, and initial hydraulic heads are set to 0 m everywhere. Fig. 1 shows the three log-conductivity reference fields with indication of the groundwater flow boundary conditions, along with their histograms. The total simulation time is five days, discretized into 100 time steps. Transient groundwater flow is numerically solved by MODFLOW 2005 (Harbaugh, 2005) in FloPy (Bakker et al., 2016).

Each transient simulation for each reference field was sampled at the locations shown in Fig. 2. The sampled values will be assimilated by the ERFF to retrieve the spatial heterogeneity of the reference fields. Only the observations for the 26 first time steps are used during the assimilation. The remaining 74 time steps are used for validation. Fig. 2 also shows three control points that will not be used during the assimilation, but that will also serve to validate the final results.

Twelve scenarios were defined to evaluate the performance of the ERFF. The scenarios were built to analyze the influence of the number of realizations in the ensemble, the number of observation points, and the standard deviation of the reference field. Table 1 summarizes the scenarios considered.

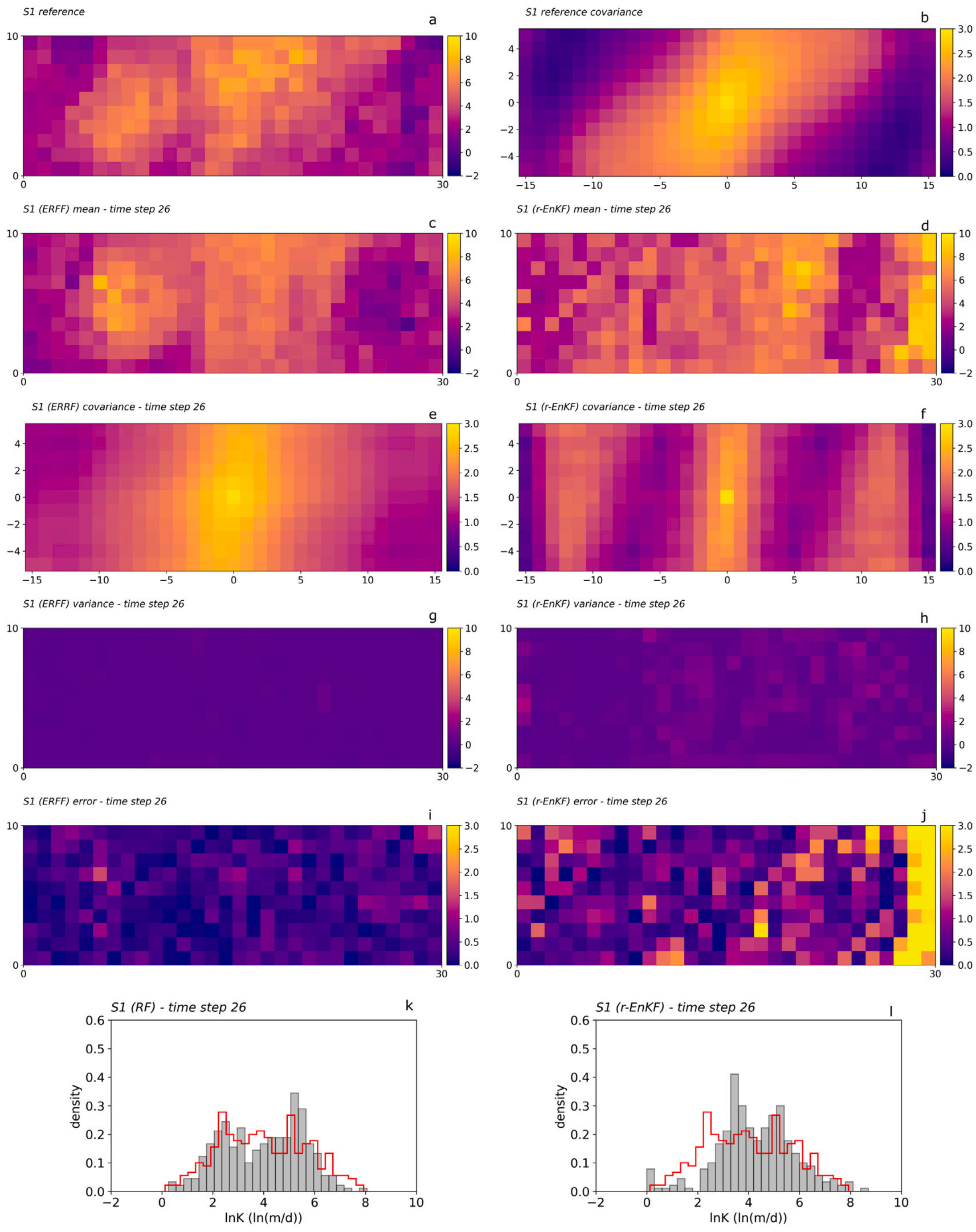


Fig. 9. Comparison between the ensemble random forest filter and the restart ensemble Kalman filter for scenario S1; (a) reference field, (b) reference field covariance map, (c) log-conductivity ensemble mean for S1, (d) log-conductivity ensemble mean for r-EnKF, (e) and (f) mean covariance maps for the final log-conductivities for S1 and r-EnKF, respectively, (g) and (h) ensemble log-conductivity variances for S1 and r-EnKF, respectively, (i) and (j) standardized log-conductivity deviations for S1 and r-EnKF, respectively, and (k) and (l) log-conductivity histograms for S1 and r-EnKF, respectively.

For the generation of the initial ensemble of realizations, it is assumed that no prior information about the spatial variability of conductivity is available. For this reason, the assimilation procedure for all scenarios starts with an ensemble of homogeneous log-conductivity realizations drawn from Gaussian probability distributions of mean 4.0 ln (m/d) and standard deviations of 1.0, 1.7, and 2.5 ln (m/d) according to the last column in Table 1. Fig. 3 displays the ensemble means (Fig. 3a) and the ensemble variances (Fig. 3b–d) for the initial log-conductivity fields for all scenarios. As expected, these values are homogeneous and equal to the prior mean (the same for all scenarios) and variance (different for the scenarios according to Table 1).

Each scenario is used to study the ERFF for identifying the reference field with the same standard deviation in the last column of Table 1. Apart from the standard deviation, the scenarios differ in the number of members of the initial ensemble, which can be 50 or 100, and the number of head observation points, which can be 18 or 36, as shown in Fig. 2. Hydraulic heads are collected and assimilated every time step for the first 26 time steps; then, the model continues running until time step 100.

In all scenarios, localization is used, with a parameter a in Eq. (9) equal to 12 m, implying that virtually no spatial correlation between

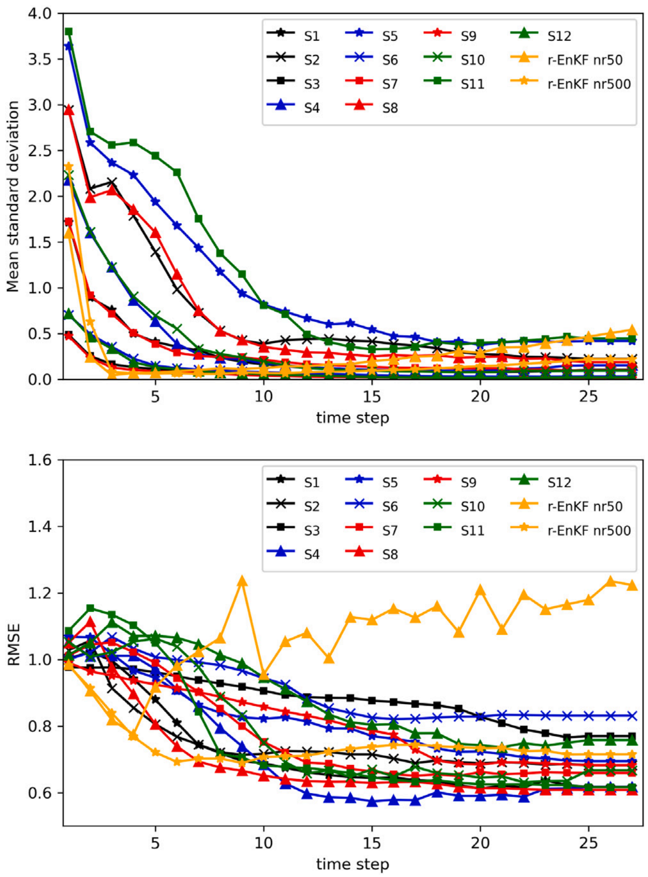


Fig. 10. ASD and RMSE.

Table 1
Scenarios considered.

Scenario	# observations	# realizations	SD
S1	18	50	1.7
S2	18	50	2.5
S3	18	50	1.0
S4	18	100	1.7
S5	18	100	2.5
S6	18	100	1.0
S7	36	50	1.7
S8	36	50	2.5
S9	36	50	1.0
S10	36	100	1.7
S11	36	100	2.5
S12	36	100	1.0

head differences and log-conductivity differences exists beyond this distance.

Finally, for completeness, the r-EnKF was also applied to scenario S1 and used as a benchmark for ERRF.

4. Results and discussion

Fig. 4 shows, for scenario S1, how the mean of the ensemble of realizations evolves as observations are assimilated. It can be observed that, starting from a homogeneous mean, heterogeneity is gradually introduced in the ensemble of realizations after each assimilation step. By step 26, the mean of the ensemble is a good estimate of the reference. The large-scale features of the reference are already visible in step 10, and by step 20, the short-scale features are displayed, too; not many changes are noticeable after step 20. Similar time evolutions are observable in the rest of the scenarios, although not shown here. These results are promising, mainly since no prior information about spatial heterogeneity is used. Fig. 5 shows the evolution of the histograms of all realizations for each scenario. In the first column, the histograms for all values in the initial ensembles of realizations for each scenario are shown as solid gray bars. The histograms of the updated fields are shown in the second and third columns. In all three columns, the hollow red histogram is the histogram in the reference field. There is not much difference between the initial and the updated histograms, although it is clear that there is a shift towards a better fit to the reference histogram in the updated realizations; however, it is important to notice that the spatial heterogeneity of the realizations has gone from homogeneous values in each realization in the first column to heterogeneous ones trying to replicate the reference so as to match the observed piezometric heads in the other two columns. As already said, the only statistical information used for the generation of the initial ensemble is the probability distribution from which to draw the homogeneous values for each realization; these distributions were chosen to match the ones used to generate the references, but it can be said that starting from a uniform distribution with reasonable ranges will yield the same results, meaning that the method is capable of retrieving the spatial patterns of the heterogeneous log-conductivity field with virtually no prior information on this parameter. The difference between the scenarios in the second and third columns of Fig. 5 is the number of observation points, 18 and 36, respectively. With 36 observations, the final updated histograms are slightly closer to the reference ones.

The performance of the method was further analyzed through sensitivity analysis to three variables: number of observation points, ensemble size, and hydraulic conductivity variance.

Fig. 6 shows the ensemble mean of the updated log-conductivity fields after the 26th assimilation time step for all scenarios. The left column shows the final mean log-conductivity field corresponding to a standard deviation of 1.7 ln (m/d), while the center and right columns show the final fields corresponding to standard deviations of 2.5 and 1.0 ln (m/d), respectively. The first row presents the reference fields for comparison purposes, the second and third rows refer to scenarios with 18 observation points, and the fourth and fifth rows show the scenarios with 36 observation points. Fig. 7 presents the ensemble variance of the updated log-conductivity fields after the 26th assimilation time step for all scenarios. And Fig. 8 shows the standardized discrepancy between the reference and the ensemble mean of the updated fields for each scenario computed as the difference between reference value and ensemble mean over the scenario standard deviation. (In the latter two figures, no reference row is displayed, the first and second rows correspond to the scenarios with 18 observation points, and the third and fourth rows scenarios with 36 observation points. Left, center and right columns correspond to scenarios with log-conductivity standard deviations of 1.7, 2.5, and 1.0 ln (m/d), respectively.) From these three figures, one can observe that the method successfully reproduces the heterogeneity of the reference fields regardless of the scenario. It is worth noting that the results are very similar, independently of the number of simulations, the number of observations, or the standard deviation of the reference field. Only a slight improvement is found when the number of observations is doubled. Further analyses carried out and not presented here showed that the number of observations could be reduced to ten and still, the ERFF recovered the heterogeneity of the underlying conductivity fields. The success of the approach must be related to the ability of random forests to extract non-linear relationships between explanatory variables (piezometric head differences) and the parameters (hydraulic conductivity differences).

Fig. 9 compares the ERFF and the r-EnKF using the same number of observation points, ensemble size, and standard deviation of the reference field. The first row shows the reference field and the covariance map computed on it. The second row shows the ensemble mean for the ERFF (left) and the r-EnKF (right). As noticeable, the ERFF mean is much closer to the reference than the r-EnKF. The third row shows the ensemble average covariance maps, where again the covariance

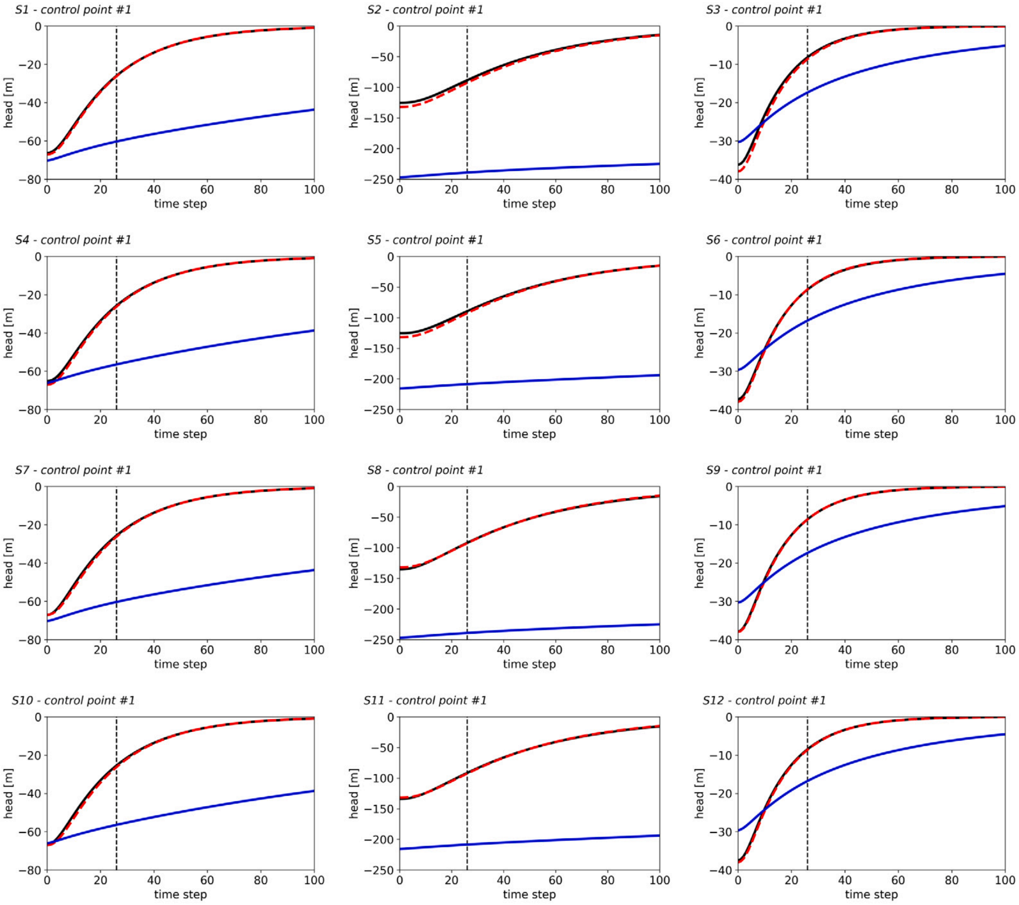


Fig. 11. Head evolution at control point #1. Reference field (dashed line). Mean of head simulations in the initial log-conductivity ensembles (solid blue line). Mean of head simulations in the final ensembles (solid black line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

map of ERFF is closer to the reference than the one from r-ENKF. It should be noticed that the ellipse of anisotropy is slightly smoothed for the ERFF, while for the r-ENKF, the covariance map seems to display a hole effect behavior with maximum continuity close to the north–south direction. The fourth row shows the ensemble variance, which is quite close to zero for both ERFF and r-ENKF. The fifth row shows the error had the ensemble mean been used as an estimate for the reference; again, the ERFF outperforms the r-ENKF. Finally, the sixth row shows the histograms of the final realizations as compared with the reference histogram. The same results can be noticed. The conclusion would be that for 50 realizations, the ERFF is superior to the r-ENKF: even with localization, the small number of realizations in r-ENKF takes an important toll. This does not mean that the r-EnKF is disqualified for inverse modeling, but the ERFF is better under these settings.

Aware of the very good results that the r-EnKF had given in the past, the exercise was repeated with an ensemble of 500 realizations, and then it yielded results as good as the ERFF. The problem was with the number of realizations.

The computational costs of both methods and scenarios were also evaluated by measuring the CPU runtime in an 11th Gen Intel Core i9-11900KF 3.5 GHz with 64 GB of RAM. Table 2 shows the run times in minutes. For the ERFF scenarios with 18 observation locations, the CPU runtime nearly doubles when we go from 50 to 100 realizations; with 36 observation locations, going from 50 to 100 observations triples the CPU runtime. The CPU runtime for the r-EnKF with 50 realizations and 18 observations is extremely low compared to scenarios with the same characteristics (S1, S2, and S3), reflecting the additional time required by the ERFF to generate the realizations of the differences and train the RF for each cell. However, the computational cost needed by the r-EnKF to arrive at satisfactory results is 2,2 times greater when compared to the ERFF of the same characteristics.

Table 2
Computational costs.

Scenario	CPU runtime (minutes)	# observations	# realizations	SD
S1	17	18	50	1.7
S2	17	18	50	2.5
S3	18	18	50	1.0
S4	40	18	100	1.7
S5	41	18	100	2.5
S6	40	18	100	1.0
S7	21	36	50	1.7
S8	21	36	50	2.5
S9	21	36	50	1.0
S10	67	36	100	1.7
S11	65	36	100	2.5
S12	75	36	100	1.0
r-EnKF 500	38	18	500	1.0
r-EnKF 50	5	18	50	1.0

For quantitative analysis, the root-mean-square errors (RMSE) and the average standard deviations (ASD) were computed according to

$$RMSE = \sqrt{\frac{1}{n_e n_p}\sum_{i=1}^{n_p}\sum_{j=1}^{n_e}(x_{ij} - x_i^{ref})^2} \tag{15}$$

$$ASD = \frac{1}{n_p}\sum_{i=1}^{n_p}\sigma_{x_i} \tag{16}$$

where n_e is the number of realizations in the ensemble, n_p is the number of cells, x_{ij} represents the log-conductivity at cell i in realization j, x_i^{ref}

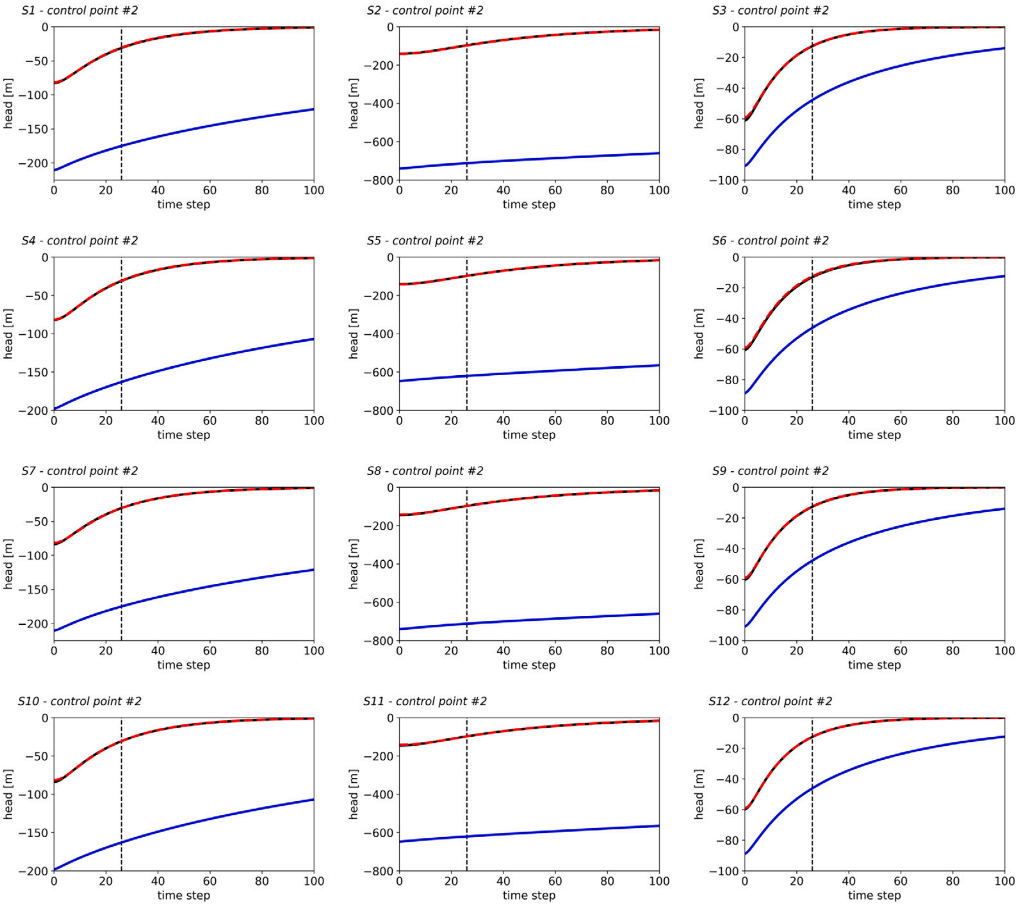


Fig. 12. Head evolution at control point #2. Reference field (dashed line). Mean of head simulations in the initial log-conductivity ensembles (solid blue line). Mean of head simulations in the final ensembles (solid black line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

is the log-conductivity in the reference, and σ_{x_i} is the log-conductivity ensemble standard deviation at cell i. Their evolution in time is shown in Fig. 10, for all 12 scenarios plus the r-EnKF with 50 and 500 ensemble realizations. Both values decrease in magnitude as time passes, with the best performer being S11 (highest values for number of observations, number of realizations, and reference variance), followed by S8 (same as S11 but with only 50 realizations). Note also how the RMSE goes chaotic for the r-EnKF with 50 realizations after iteration 5, probably due to a problem with filter inbreeding, very common in ensemble Kalman filtering with few realizations.

Finally, Figs. 11, 12, and 13 show how the piezometric heads are reproduced at the three control points. Observations were assimilated only until time step 26 (vertical dashed line in all plots), but the piezometric head evolution is shown until the end of the simulation period at time 100. All figures show the head simulation in the reference field from time zero (dashed red line), the average of all head simulations in the initial ensemble of realizations (solid blue line), and the average of the simulations in the updated log-conductivity fields after 26 assimilation steps (solid black line). Note that piezometric head axes vary for each plot to best display the results. The graphs have been grouped by columns, with each column corresponding to one of the three reference cases. It is quite remarkable how the piezometric heads change from being completely off target at time zero to matching, almost perfectly, the reference head curves. The minimal discrepancies between the mean of the simulated values and the reference happen in some of the scenarios with the smaller number of observations, i.e., S4 and S5. For comparison purposes, the head evolution in the log-conductivity realizations obtained using the r-EnKF with 50 realizations is shown in Fig. 14, where it can be seen that the reproduction of the reference values is not as good as for the ERFF, particularly for control points #1 and #2.

5. Conclusion

A new data assimilation method, the ensemble random forest filter (ERFF), has been proposed. It is inspired by the ensemble Kalman filter but replaces the linear updating step with a non-linear update computed using random forests. The ERFF uses an ensemble of log-conductivity realizations and its associated ensemble of predicted piezometric heads to build a large training dataset that is an order of magnitude larger than the initial set of realizations (the dataset size grows with the square of the number of realizations). The random forest analyzes the differences in the predicted piezometric heads at observation locations with the differences in log-conductivities throughout the domain, learns from this training set, and then predicts what should be the difference to be added to the log-conductivity at each location in the domain once the head observations are collected and their differences with respect to the predictions evaluated.

The method has been tested in a number of scenarios with varying degrees of heterogeneity (as measured by the standard deviation), different number of realizations in the ensemble, and different number of observation locations, and it has been found to perform well in all scenarios and better than its benchmarking the restart ensemble Kalman filter when the same number of realizations are used. Only when the number of realizations rises to 500 is the Kalman filter capable of providing similar results but at a cost 2,2 times larger than the ERFF.

The main caveat of the proposal is, as in most machine learning applications, the choice of the hyperparameters that control the building of the random forests. This task could be time-consuming until a suitable set of hyperparameters is found that performs appropriately for the problem at hand.

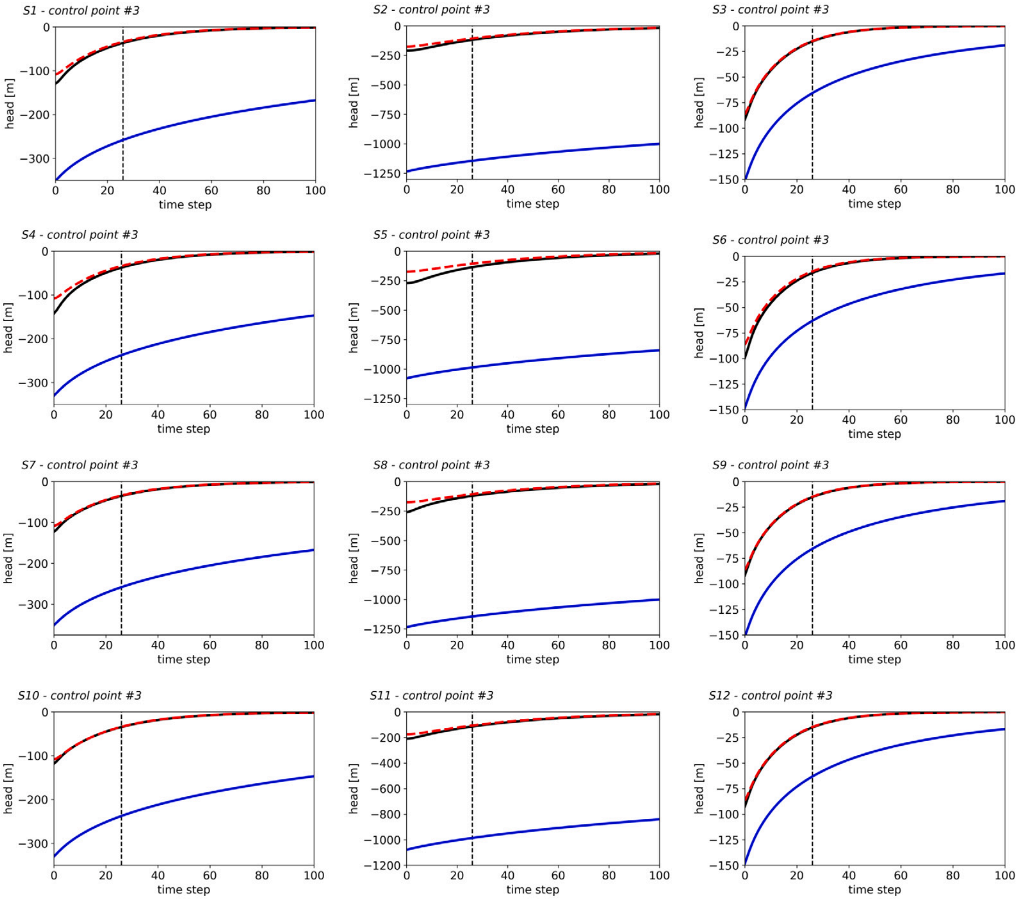


Fig. 13. Head evolution at control point #3. Reference field (dashed line). Mean of head simulations in the initial log-conductivity ensembles (solid blue line). Mean of head simulations in the final ensembles (solid black line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

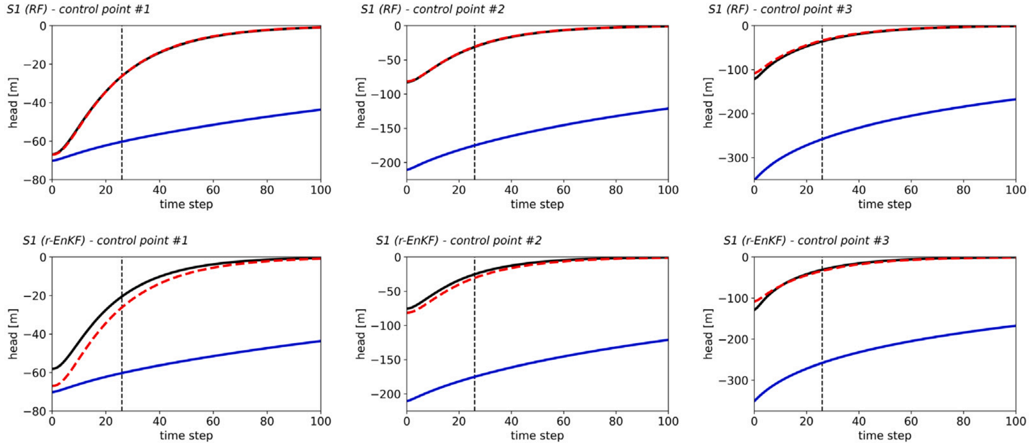


Fig. 14. Head evolution at the three control points for scenario S1 in the final ensembles of realizations obtained by the ERFF and the r-EnKF with 50 realizations. Reference field (dashed line). Mean of head simulations in the initial log-conductivity ensembles (solid blue line). Mean of head simulations in the final ensembles (solid black line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Research continues on the application of the ERFF to more complex problems, such as those involving the identification of external stresses and boundary and initial conditions or the identification of more complex log-conductivity patterns.

Funding

The authors acknowledge grant PID2019-109131RB-I00 funded by MCIN/AEI/10.13039/501100011033 and project InTheMED, which is part of the PRIMA Programme supported by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 1923.

CRediT authorship contribution statement

Vanessa A. Godoy: Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Visualization, Reviewing. **Gian F. Napa-García:** Methodology, Software, Reviewing. **J. Jaime**

Gómez-Hernández: Methodology, Software, Validation, Formal analysis, Investigation, Writing – review & editing, Visualization, Supervision, Funding acquisition, Reviewing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Al-Abadi, A.M., Alsamaani, J.J., 2020. Spatial analysis of groundwater flowing artesian condition using machine learning techniques. Groundw. Sustain. Dev. 11, 100418.

An, Y., Yan, X., Lu, W., Qian, H., Zhang, Z., 2021. An improved Bayesian approach linked to a surrogate model for identifying groundwater pollution sources. Hydrogeol. J. 1–16.

Asher, M.J., Croke, B.F., Jakeman, A.J., Peeters, L.J., 2015. A review of surrogate models and their application to groundwater modeling. Water Resour. Res. 51 (8), 5957–5973.

Bakker, M., Post, V., Langevin, C.D., Hughes, J.D., White, J.T., Starn, J., Fienen, M.N., 2016. Scripting MODFLOW model development using Python and FloPy. Groundwater 54 (5), 733–739.

Bao, J., Li, L., Davis, A., 2022. Variational autoencoder or generative adversarial networks? A comparison of two deep learning methods for flow and transport data assimilation. Math. Geosci. 1–26.

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

Biau, G., 2012. Analysis of a random forests model. J. Mach. Learn. Res. 13, 1063–1095.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Capilla, J.E., Rodrigo, J., Gómez-Hernández, J.J., 1999. Simulation of non-Gaussian transmissivity fields honoring piezometric data and integrating soft and secondary information. Math. Geol. 31 (7), 907–927.

Carrera, J., Alcolea, A., Medina, A., Hidalgo, J., Slooten, L.J., 2005. Inverse problem in hydrogeology. Hydrogeol. J. 13 (1), 206–222.

Chen, Z., Gómez-Hernández, J.J., Xu, T., Zanini, A., 2018. Joint identification of contaminant source and aquifer geometry in a sandbox experiment with the restart ensemble Kalman filter. J. Hydrol. 564, 1074–1084.

Chen, Y., Zhang, D., 2006. Data assimilation for transient flow in geologic formations via ensemble Kalman filter. Adv. Water Resour. 29 (8), 1107–1122.

Cutler, A., Cutler, D.R., Stevens, J.R., 2012. Random forests. In: Ensemble Machine Learning. Springer, pp. 157–175.

Emerick, A.A., Reynolds, A.C., 2013. Ensemble smoother with multiple data assimilation. Comput. Geosci. 55, 3–15.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res.: Oceans 99 (C5), 10143–10162.

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

Fernàndez-Garcia, D., Gómez-Hernández, J., 2007. Impact of upscaling on solute transport: Traveltimes, scale dependence of dispersivity, and propagation of uncertainty. Water Resour. Res. 43 (2).

Feyen, L., Gómez-Hernández, J., Ribeiro Jr., P., Beven, K.J., De Smedt, F., 2003. A Bayesian approach to stochastic capture zone delineation incorporating tracer arrival times, conductivity measurements, and hydraulic head observations. Water Resour. Res. 39 (5).

Gelsinari, S., Doble, R., Daly, E., Pauwels, V.R., 2020. Feasibility of improving groundwater modeling by assimilating evapotranspiration rates. Water Resour. Res. 56 (2), e2019WR025983.

Gómez-Hernández, J.J., Journel, A.G., 1993. Joint sequential simulation of multigaussian fields. In: Geostatistics Troia'92. Springer, pp. 85–94.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. Adv. Neural Inf. Process. Syst. 27.

Harbaugh, A.W., 2005. MODFLOW-2005, the US Geological Survey Modular Ground-Water Model: The Ground-Water Flow Process. US Department of the Interior, US Geological Survey Reston, VA, USA.

He, J., Yue, X., Ren, Z., 2021. The impact of assimilating ionosphere and thermosphere observations on neutral temperature improvement: Observing system simulation experiments using EnKF. Space Weather 19 (10), e2021SW002844.

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6, e5518.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems.

Kim, J., Yoo, J., Do, K., 2020. Wave data assimilation to modify wind forcing using an ensemble Kalman Filter. Ocean Sci. J. 55 (2), 231–247.

Knoll, L., Breuer, L., Bach, M., 2019. Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. Sci. Total Environ. 668, 1317–1327.

Li, L., Zhou, H., Gómez-Hernández, J.J., 2011. A comparative study of three-dimensional hydraulic conductivity upscaling at the macro-dispersion experiment (MADE) site, Columbus Air Force Base, Mississippi (USA). J. Hydrol. 404 (3–4), 278–293.

Liu, K., Vardon, P., Hicks, M., 2018. Sequential reduction of slope stability uncertainty based on temporal hydraulic measurements via the ensemble Kalman filter. Comput. Geotech. 95, 147–161.

Mariethoz, G., Gómez-Hernández, J.J., 2021. Machine learning for water resources. Front. Artif. Intell. 4, 63.

Mo, S., Zabaras, N., Shi, X., Wu, J., 2019. Deep autoregressive neural networks for high-dimensional inverse problems in groundwater contaminant source identification. Water Resour. Res. 55 (5), 3856–3881.

Nguyen, X.H., et al., 2020. Combining statistical machine learning models with ARIMA for water level forecasting: The case of the Red river. Adv. Water Resour. 142, 103656.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Sachdeva, S., Kumar, B., 2021. Comparison of gradient boosted decision trees and random forest for groundwater potential mapping in Dholpur (Rajasthan), India. Stoch. Environ. Res. Risk Assess. 35 (2), 287–306.

Shuai, Y., White, C., Sun, T., Feng, Y., 2016. A gathered EnKF for continuous reservoir model updating. J. Pet. Sci. Eng. 139, 205–218.

Sit, M., Demiray, B.Z., Xiang, Z., Ewing, G.J., Sermet, Y., Demir, I., 2020. A comprehensive review of deep learning applications in hydrology and water resources. Water Sci. Technol. 82 (12), 2635–2670.

Tahmasebi, P., Sahimi, M., 2021. Special issue on machine learning for water resources and subsurface systems. Adv. Water Resour. 103851.

Todaro, V., D'Oria, M., Tanda, M.G., Gómez-Hernández, J.J., 2019. Ensemble smoother with multiple data assimilation for reverse flow routing. Comput. Geosci. 131, 32–40.

Wen, X.-H., Capilla, J.E., Deutsch, C., Gómez-Hernández, J., Cullick, A., 1999. A program to create permeability fields that honor single-phase flow rate and pressure data. Comput. Geosci. 25 (3), 217–230.

Xu, T., Gómez-Hernández, J.J., 2016. Joint identification of contaminant source location, initial release time, and initial solute concentration in an aquifer via ensemble Kalman filtering. Water Resour. Res. 52 (8), 6587–6595.

Xu, T., Gómez-Hernández, J.J., 2018. Simultaneous identification of a contaminant source and hydraulic conductivity via the restart normal-score ensemble Kalman filter. Adv. Water Resour. 112, 106–123.

Xu, T., Jaime Gómez-Hernández, J., Zhou, H., Li, L., 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. Adv. Water Resour. 54, 100–118.

Yin, J., Zhan, X., Zheng, Y., Hain, C.R., Liu, J., Fang, L., 2015. Optimal ensemble size of ensemble Kalman filter in sequential soil moisture data assimilation. Geophys. Res. Lett. 42 (16), 6710–6715.

Zhang, J., Lin, G., Li, W., Wu, L., Zeng, L., 2018. An iterative local updating ensemble smoother for estimation and uncertainty assessment of hydrologic model parameters with multimodal distributions. Water Resour. Res. 54 (3), 1716–1733.

Zhang, J., Zheng, Q., Wu, L., Zeng, L., 2020. Using deep learning to improve ensemble smoother: Applications to subsurface characterization. Water Resour. Res. 56 (12), e2020WR027399.

Zhou, H., Gómez-Hernández, J.J., Li, L., 2014. Inverse methods in hydrogeology: Evolution and recent trends. Adv. Water Resour. 63, 22–37.

Zhu, P., Shi, L., Zhu, Y., Zhang, Q., Huang, K., Williams, M., 2017. Data assimilation of soil water flow via ensemble Kalman filter: Infusing soil moisture data at different scales. J. Hydrol. 555, 912–925.