Data assimilation in density-dependent subsurface flows via localized iterative ensemble Kalman filter

Chuan-An Xia[1, 2], Bill X. Hu[1, 2, 3 *], Juxiu Tong[1, 2] Alberto Guadagnini[4, 5 *]

[1]. School of Water Resources and Environment, China University of Geosciences, Beijing, China 100083

[2]. Key Laboratory of Groundwater Cycle and Environment Evolution (China University of Geosciences), Ministry of Education, Beijing, China 100083

[3]. Department of Earth, Ocean and Atmospheric Science, Florida State University, Tallahassee, FL, USA

[4]. Dipartimento di Ingegneria Civile e Ambientale, Politecnico di Milano, Milano, Italy

[5]. Department of Hydrology and Atmospheric Sciences, The University of Arizona, Tucson, USA

Accepted Manuscript

Water Resources Research

*Corresponding author: Bill X. Hu; Alberto Guadagnini

Tel:

Email: bill.x.hu@gmail.com; alberto.guadagnini@polimi.it

Abstract

Parameter estimation in variable-density groundwater flow systems is confronted with challenges of strong nonlinearity and heavy computational burden. Relying on a variant of the Henry problem, we evaluate the performance of a domain localization scheme of the iterative ensemble Kalman filter in the framework of data assimilation settings for variable-density groundwater flows in a seawater intrusion scenario. The performance of the approach is compared against (*a*) the corresponding domain localization scheme of the ensemble Kalman filter in its standard formulation as well as (*b*) a covariance localization scheme of the latter. The equivalent freshwater head, h_f, and salinity, S_a, are set as the target state variables. The randomly heterogeneous field of equivalent freshwater hydraulic conductivity, K_f, is considered as the system parameter field. Density-independent and density-driven flow settings are considered to evaluate the assimilation results using various methods and data. When only h_f data are assimilated, all tested approaches perform generally well and a localization scheme embedded in the iterative ensemble Kalman filter appears to consistently outperform the domain localized version of the standard ensemble Kalman filter in a density-driven scenario; Dirichlet boundary conditions tend to show a more pronounced negative effect on estimating K_f for density-independent than for density-dependent flow conditions; h_f data are more informative in a density-dependent than in a density-independent setting. The sole use of S_a information does not yield satisfactory updates of h_f for the covariance localization scheme of the standard ensemble Kalman filter while the sole use of h_f does. The domain localization scheme leads to difficulties in the attainment of global filter convergence when only S_a data are used. A covariance localization scheme associated with a standard ensemble Kalman filter can significantly alleviate this issue.

Keywords: variable density flow; value of data; iterative ensemble Kalman filter; ensemble Kalman filter

1. Introduction

The process of seawater intrusion (SWI) is documented to be critically plaguing several coastal areas. This poses serious concerns, in light of competitive use of groundwater resources, the latter being subject to diverse anthropogenic stresses in the context of, e.g., human consumption for domestic use, irrigation and farming activities, industrial operations and processes, increased urbanization and tourism, or population dynamics in coastal regions. A variety of studies targeting optimization of water withdrawals for appropriate management of water resources in coastal zones have been conducted, comprehensive reviews being provided by Sreekanth and Datta (2015) and Datta and Kourakos (2015). A key assumption used in several management-oriented optimization studies is that model parameters can be deterministically characterized, which is rarely the case under field conditions.

1.1 Challenges of parameter estimation in density-dependent flow

One of the main complexities associated with variable density settings is that flow (and transport) dynamics are driven jointly by hydraulic and density gradients. Strong nonlinearities, heavy computational requirements and measurement difficulty are three key challenges for parameter estimation in density-dependent groundwater flows (Carrera et al., 2010; Hu et al., 2016; Colombani et al., 2016). The combination of a traditional (deterministic) inverse modeling approach with automatic model calibration procedures, such as those embedded in the widely used codes PEST (Doherty, 2002) and/or UCODE (Poeter et al., 2005), is prone to yield suboptimal parameter estimation results in density-dependent flow settings and is plagued by strong nonlinearity and remarkable computational costs (Carrera et al., 2010). The latter aspect is a critical concern in realistic scenarios where system zonation through a small number of uniform sub-regions has been shown to provide inaccurate results (e.g., Sanford et al., 2009; Sanford and Pope, 2010). It is also worth pointing out that increasing the dimensionality of the model parameter space to attempt improving the quality of inverse modeling can induce suboptimal parameter estimates for a strongly nonlinear flow system.

The difficulty to obtain a sufficient amount of reliable data to describe the states of a target subsurface system is an additional challenge for model parameter estimation and updating of system states. Typical data of state variables available across a coastal aquifer include hydraulic head and salinity. With reference to the former, head variations can be informative only when measurement depth and salinity variation along the borehole are precisely detected (Post et al., 2007). In addition, head fluctuations detected at a well in the mixing zone between fresh and salt water can be much higher than those taking place in the actual aquifer (Shalev et al., 2009). Direct salinity measurements can be complemented by geophysical campaigns (e.g., Beaujean et al., 2014; Pidlisecky et al., 2015; El-Kaliouby and Abdalla, 2015; Kourgialas et al., 2016). Salinity data associated with an Integrated Depth Sampling approach represent an integral value along a borehole screen and may not be suitable to fully constrain a SWI model. This is markedly evident in cases where water elevation (as opposed to pressure head) data are collected from a borehole, because the influence of density on head measurements somehow shadows the actual head value at the well. Multi-Level Sampling measurements could provide useful information, albeit the need to achieve this degree of detail should be carefully balanced against increased model complexity requirements (Colombani et al., 2016). It should also be noted that characterization of aquifer heterogeneity on the basis of salinity data is complex and not always robust, because of the indirect relationship between salinity and permeability. A discussion of limitations and advantages of inverse modeling in coastal aquifers relying on these types of information can be found in Werner et al. (2013) and Carrera et al. (2010).

In the broad context illustrated above, quantification of the actual effectiveness of data to constrain predictions of a SWI process is recognized as a challenging research issue, which is still not completely resolved. Shoemaker (2004) suggested that hydraulic head data are less informative than salinity data to characterize the hydraulic conductivity field in a simple model of Biscayne Bay (Florida, USA). Sanz and Voss (2006) found that pressure data are more useful than salinity to estimate permeability in the context of the Henry problem.

1.2 Iterative ensemble Kalman filter

Aquifers are generally heterogeneous, their hydraulic parameters, such as permeability, significantly varying in space. Our inability to obtain a deterministic characterization of the system at a scale of interest has led to the development of stochastic approaches to quantify uncertainty propagating from (often unknown) model parameters to system states of interest (e.g., Zhang, 2002).

A stochastic approach such as the ensemble Kalman filter (EnKF) is widely used to estimate model parameters under uncertain conditions as well as to update system state variables as new data are available. As opposed to the extended Kalman filter (EKF), EnKF does not require linearization and is suitable for nonlinear systems. It has been widely used in groundwater flow and solute transport scenarios (e.g., amongst others, Chen and Zhang, 2006; Liu et al., 2008; Huang et al., 2009; Tong et al., 2010, 2013; Gharamti et al., 2015; Shi et al., 2015; Crestani et al., 2015; Zovi et al., 2017).

Nonlinearities in the data assimilation (DA) process are much more pronounced in density-dependent than in density-independent flow (or solute transport) scenarios. As such, several iterative forms of EnKF (e.g., Reynolds et al., 2006; Li and Reynolds, 2007; Gu and Oliver, 2007; Sakov et al., 2012) have been developed.

Gu and Oliver (2007) proposed an iterative ensemble Kalman filter (IEnKF) scheme, termed ensemble randomized maximal likelihood filter (EnRML), to perform data assimilation in the context of a nonlinear problem, with special focus on multiphase flow in porous media. Wang et al. (2010) suggested that the EnRML may be prone to divergence in some highly nonlinear cases. Sakov et al. (2012) introduced another IEnKF scheme to eliminate suboptimal solutions by replacing the standard EnKF with the ensemble square root filter (ESRF) and rescaling ensemble anomalies with the ensemble transform matrix at each iteration. Employing a square root technique during the iterative process enables one to use ensemble anomalies to conduct state analysis without the need of perturbing observations (Whitaker and Hamill, 2002; Sakov et al., 2012; Gharamti et al., 2013; Gharamti and Hoteit, 2014). As such, the proposed scheme can alleviate suboptimality in the standard EnKF. In the IEnKF, the filtering probability density function (*pdf*) at the present state relies on the maximum of a smoothing *pdf* associated with the last ensemble state given present observations. Note that the IEnKF is still based on the filtering theory (Jazwinski, 1970). The current target *pdf* of the system state of interest (termed filtering *pdf*) is conditional on present and past observations. On these bases, Bocquet and Sakov (2014) extended the IEnKF through an iterative ensemble Kalman smoother (IEnKS) that outperforms the standard Kalman filters and smoothers (Evensen, 2003; 2009), using an extended data assimilation window (DAW). In the IEnKS, the DAW with length $L\Delta t$ ($L \geq 1$), Δt representing the fixed time difference between two sequential times associated with available observations, is allowed to shift with fixed length $V\, \Delta t$ ($L \geq V \geq 1$), as time unfolds. Note that when $L = V = 1$ (indicating lag-one IEnKS) there are no overlaps between diverse assimilation cycles, and the IEnKS reduces to an IEnKF form. Gharamti et al. (2015) proposed a new iterative framework of the one-step-ahead smoothing EnKF which can iteratively maximize the smoothing *pdf* without the need of ensemble propagation, with the assumption that parameter evolution during the iteration procedure weakly influences the innovation term (see equation (13) in Gharamti et al., 2015).

Various ad hoc iterative applications of EnKF have been used to solve targeted subsurface and surface hydrology problems (e.g., Moradkhani et al., 2005; Wen and Chen, 2006; Krymskaya et al., 2009; Song et al., 2014; Ng et al., 2014; Gharamti et al., 2013, 2014).

Spurious correlation is a key problem which needs to be properly tackled in the ensemble data assimilation technique. It is typically caused by the limited number of Monte Carlo realizations for the system parameters and state variables (Houtekamer and Mitchell, 1998). The ensuing rank-deficiency of the sample error covariance matrix could dramatically decrease the performance of a data assimilation approach (Houtekamer and Mitchell, 1998; Hamill et al., 2001). Inflation and multi-ensemble configuration (Houtekamer and Mitchell, 2001) could mitigate this issue to some extent. Localization techniques, including covariance and/or domain localization, could also be used to dampen long-range spurious correlations (e.g., Evensen, 2003;

Nan and Wu, 2011; Tong et al., 2012). Covariance localization is typically performed by the Schur product of a smoothing function with the regularized error covariance matrix (Nan and Wu, 2011). Domain localization is obtained by using solely measurements within a region near the location where system state variables and parameters need to be updated. The size of this region is usually selected empirically (Evensen, 2003). Some approaches can be used to alleviate (or even eliminate) spurious correlations. These include, e.g., efficient sampling schemes (Hendricks-Franssen and Kinzelbach, 2008), finite-size ensemble Kalman filter (Bocquet, 2011), and moment-equations based ensemble Kalman filter (Panzeri et al., 2013, 2014, 2015).

Even as localization can be a useful technique to alleviate the rank-deficiency issue, its level of compatibility with the use of an iterative technique for highly nonlinear systems is not entirely clear. In this context, a domain localized scheme of the IEnKS was proposed by Bocquet (2016) to reduce such issues. While all of the above referenced studies have led to interesting results, none of these focuses on density-dependent flow scenarios. Therefore, their performance in such scenarios is still unexplored.

1.3 Motivation and objectives

In this study, we primarily aim at assessing the use of a domain localization scheme embedded in IEnKF (i.e., the localized IEnKS with L = V =1 by Bocquet (2016)) to effectively update density-dependent flow system parameter and state variables while coping with strong nonlinearities and heavy computational requirements. Our work is motived by the observations that (*i*) the domain localized IEnKS (Bocquet, 2016) can tackle highly nonlinear systems with small ensemble size;

(*ii*) a detailed assessment of the ability of stochastic data assimilation methods to estimate hydraulic parameters and update model states are relevant and still scarcely explored for density-dependent groundwater flows in heterogeneous coastal aquifers (e.g., Sreekanth and Datta, 2015), and (*iii*) the relative value of diverse types of data for parameter estimation in density-dependent groundwater flows is still unclear.

We consider a seawater intrusion scenario corresponding to a variant of the Henry problem, which includes the action of a pumping well operating in a heterogeneous domain. For comparison purposes, we also analyze the performance in the same setting of the ensemble Kalman filter in its standard formulation and by embedding in it a covariance localization scheme. Aspects associated with corresponding computational complexities are also analyzed. To establish a baseline for comparison, the data assimilation schemes are also compared in the absence of density effects. The performances of the approaches are assessed in terms of their potential to deal with effects of (*a*) various types of observations (with diverse spatial arrangements of sampling locations), (*b*) magnitude of measurement errors, (*c*) temporal frequency of data assimilation, (*d*) the number of realizations forming the collection (or ensemble) of random fields employed in the calculations, as well as (*e*) uncertainties associated with our prior knowledge of the correlation length of the underlying (randomly heterogeneous) hydraulic conductivity field.

The study is structured as follows. The domain localized IEnKF scheme we employ is introduced in Section 2. The conceptual model setting and the mathematical description of the variable and constant density groundwater flow scenarios are presented in Section 3. Results are illustrated and discussed in Section 4. Conclusions are presented in Section 5.

2. Domain localized iterative ensemble Kalman filter

Forecast and update are the two key elements associated with each data assimilation step. The (nonlinear) model which propagates the system states (here, specified as a collection of system state variables and model parameters) from time t_0 to t_1 is denoted as $\mathcal{M}_{1\leftarrow 0}$. We denote the model state vectors at t_0 and t_1 as $\mathbf{x}_0$ and $\mathbf{x}_1$, respectively. An observation vector $\mathbf{y}_1$ is assimilated at time t_1.

Let us consider the ensemble matrix $\mathbf{E}=\left\{\mathbf{x}_{(1)},\ \ldots,\mathbf{x}_{(N)}\right\}$ whose entries are the N state vectors $\mathbf{x}_{(i)}$ (subscript i = 1, 2, ..., N referring to the i^{th} ensemble member; where N is the number of the system state realizations collected in **E**), including state variables (i.e., equivalent freshwater head and salinity for density-driven flow, and only freshwater head for density-independent flow in this study) and the natural logarithm of equivalent freshwater hydraulic conductivity. Note that the size of vector $\mathbf{x}_{(i)}$ is $3\times m$ or $2\times m$, respectively for density- dependent or independent flow, m being the number of grid cells employed in the numerical model.

The empirical mean vector ($\bar{\mathbf{x}}$) and covariance matrix ($\mathbf{P}$) of these N vectors are calculated as

$$\bar{\mathbf{x}}=\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_{(n)},\qquad\qquad \mathbf{P}=\mathbf{A}\mathbf{A}^{\mathrm{T}}/(N-1) \tag{1}$$

where the anomaly matrix $\mathbf{A}$ is defined as $\mathbf{A}=\left\{\mathbf{x}_{(1)}-\bar{\mathbf{x}},\ ...,\ \mathbf{x}_{(N)}-\bar{\mathbf{x}}\right\}$. We also denote by $\mathbf{x}_0^{(0)}$ the vector of initial (or prior) guesses, and by $\mathbf{P}_0$ the covariance of forecast state errors at t_0, used in the data assimilation process.

Similar to Sakov et al. (2012), for each data assimilation cycle (from t_0 to t_1) we compute the smoothing *pdf* $f\left(\mathbf{x}_0 \mid \mathbf{y}_1\right)$ for state $\mathbf{x}_0$ given $\mathbf{y}_1$ and the analysis *pdf* $f\left(\mathbf{x}_1 \mid \mathbf{y}_1\right)$. According to Bayes' rule, one can note that $f\left(\mathbf{x}_0 \mid \mathbf{y}_1\right) \propto f\left(\mathbf{y}_1 \mid \mathbf{x}_0\right) f\left(\mathbf{x}_0\right)$, where $f\left(\mathbf{y}_1 \mid \mathbf{x}_0\right)$ and $f\left(\mathbf{x}_0\right)$ are the likelihood and prior *pdf* conditioned on the prior guessed mean state $\mathbf{x}_0^{(0)}$, respectively. One can write the prior *pdf* according to a Gaussian model as $f\left(\mathbf{x}_0\right) = \mathcal{N}\left(\mathbf{x}_0 \mid \mathbf{x}_0^{(0)}, \mathbf{P}_0\right)$ and then obtain

$$f\left(\mathbf{x}_0\right) \propto \exp\left(-\frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}_0^{(0)}\right)^{\mathrm{T}} \mathbf{P}_0^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^{(0)}\right)\right) \tag{2}$$

Here, superscript T denotes transpose. The likelihood is assumed to be Gaussian and can be written as

$$f\left(\mathbf{y}_1 | \mathbf{x}_0\right) \propto \exp\left(-\frac{1}{2}\left[\mathbf{y}_1 - \mathcal{H}_1\left(\mathcal{M}_{1\leftarrow 0}\left(\mathbf{x}_0\right)\right)\right]^{\mathrm{T}} \mathbf{R}_1^{-1}\left[\mathbf{y}_1 - \mathcal{H}_1\left(\mathcal{M}_{1\leftarrow 0}\left(\mathbf{x}_0\right)\right)\right]\right) \tag{3}$$

where $\mathcal{H}_1$ is a nonlinear observation operator (i.e., an operator which relates model parameters and states to available data); $\mathbf{R}_1$ is the covariance matrix of observation errors, which are usually modeled according to a zero-mean Gaussian distribution.

A cost function, $\mathcal{J}(\mathbf{x}_0)$, can then be formulated as

$$\begin{aligned}\mathcal{J}\left(\mathbf{x}_0\right) = & \frac{1}{2}\left[\mathbf{y}_1 - \mathcal{H}_1\left(\mathcal{M}_{1\leftarrow 0}\left(\mathbf{x}_0\right)\right)\right]^{\mathrm{T}} \mathbf{R}_1^{-1}\left[\mathbf{y}_1 - \mathcal{H}_1\left(\mathcal{M}_{1\leftarrow 0}\left(\mathbf{x}_0\right)\right)\right] \\ & + \frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}_0^{(0)}\right)^{\mathrm{T}} \mathbf{P}_0^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^{(0)}\right)\end{aligned} \tag{4}$$

and its minimization (being equivalent to maximize the smoothing *pdf* $f\left(\mathbf{x}_0 \,|\, \mathbf{y}_1\right)$) yields an estimate of $\mathbf{x}_0$.

2.1 The cost function in an ensemble space

The vector of true system states can be expressed as $\mathbf{x} = \bar{\mathbf{x}} + \mathbf{A}\mathbf{w}$, where $\mathbf{w}$ is a coordinate vector in the ensemble space (Bocquet, 2011). A cost function $\tilde{\mathcal{J}}\left(\mathbf{w}\right)$ defined in the ensemble space can then be written as (Hunt et al., 2007; Bocquet and Sakov, 2012)

$$\begin{aligned}\tilde{\mathcal{J}}\left(\mathbf{w}\right) = \frac{1}{2}\left[\mathbf{y}_1 - \mathcal{H}_1\left(\mathcal{M}_{1\leftarrow 0}\left(\mathbf{x}_0^{(0)} + \mathbf{A}_0\mathbf{w}\right)\right)\right]^{\mathrm{T}} \mathbf{R}_1^{-1}\left[\mathbf{y}_1 - \mathcal{H}_1\left(\mathcal{M}_{1\leftarrow 0}\left(\mathbf{x}_0^{(0)} + \mathbf{A}_0\mathbf{w}\right)\right)\right] \\ + \frac{1}{2}\left(N-1\right)\mathbf{w}^{\mathrm{T}}\mathbf{w}\end{aligned} \tag{5}$$

where $\mathbf{A}_0$ is the anomaly matrix associated with the collection of realizations forming the initial guess in the approach.

2.2 Minimization of the cost function in the ensemble space

Minimization of function (5) is performed via the Gauss-Newton algorithm as

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \tilde{H}_{(k)}^{-1}\nabla\tilde{\mathcal{J}}_{(k)}\left(\mathbf{w}^{(k)}\right) \tag{6}$$

Here, superscripts and/or subscripts in brackets represent iteration indices, the system state at iteration k being expressed as $\mathbf{x}_0^{(k)} = \mathbf{x}_0^{(0)} + \mathbf{A}_0\mathbf{w}^{(k)}$; the Jacobian $\nabla\tilde{\mathcal{J}}_{(k)}\left(\mathbf{w}^{(k)}\right)$ and the approximated Hessian $\tilde{H}_{(k)}$ at iteration k are given by (Bocquet and Sakov, 2014)

$$\nabla \tilde{\mathcal{J}}_{(k)}\left(\mathbf{w}^{(k)}\right) = -\mathbf{Q}_{(k)}^{\mathrm{T}} \mathbf{R}_1^{-1} \left[\mathbf{y}_1 - \mathcal{H}_1\left(\mathcal{M}_{1\leftarrow 0}\left(\mathbf{x}_0^{(k)}\right)\right)\right]^{\mathrm{T}} + \left(N-1\right)\mathbf{w}^{(k)} \tag{7}$$

and

$$\tilde{H}_{(k)} = \left(N-1\right)\mathbf{I}_N + \mathbf{Q}_{(k)}^{\mathrm{T}} \mathbf{R}_1^{-1} \mathbf{Q}_{(k)} \tag{8}$$

where $\mathbf{I}_N$ is the identity matrix in the ensemble space; and $\mathbf{Q}_{(k)}$ is a tangent linear operator acting from the ensemble to the observation space. One can obtain $\mathbf{Q}_{(k)}$ through two algorithms, i.e., either using an ensemble transform matrix or a scaling factor (Sakov et al., 2012; Bocquet and Sakov, 2012). The use of the ensemble transform matrix can outperform resorting to a scaling factor in the presence of multiple minima for the cost function (Sakov et al., 2012). The scaling factor algorithm, also termed as bundle variant (Bocquet and Sakov, 2012), has provided good performances in the localization scheme of IEnKS (Bocquet, 2016). Otherwise, the ensemble transform matrix algorithm has not yet been applied in conjunction with a localization method. We do so in our study, relying on the form

$$\mathbf{Q}_{(k)} \approx \mathcal{H}_1\left[\mathcal{M}_{1\leftarrow 0}\left(\mathbf{x}_0^{(k)}\mathbf{1}^{\mathrm{T}} + \mathbf{A}_0\right)\right]\left(\mathbf{I}_N - \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{N}\right)\mathbf{T}_{(k)}^{-1} \tag{9}$$

where $\mathbf{T}$ is the ensemble transform matrix, and $\mathbf{1} = [1,\ ...,\ 1]_{1\times N}$ is the identity vector. The former can be obtained at iteration k as

$$\mathbf{T}_{(k)} = \left(\left(N-1\right)\mathbf{I}_N + \mathbf{Q}_{(k-1)}^{\mathrm{T}} \mathbf{R}_1^{-1} \mathbf{Q}_{(k-1)}\right)^{-1/2} \tag{10}$$

Note that $\mathbf{T}_{(1)} = \mathbf{I}_N$ at the first iteration. The maximization procedure to obtain the smoothing *pdf* $f\left(\mathbf{x}_0 \mid \mathbf{y}_1\right)$ through iterative evaluation of equations (6)-(10) is set to stop if a predefined maximum iteration number (i.e., 10 in our computational examples) is reached or $\left\|\mathbf{w}^{(k+1)}-\mathbf{w}^{(k)}\right\| \le e$, e being a threshold which is tuned to ensure high performance of data assimilation in terms of quality of results (see also Section 4). Denoting by the superscript * a given quantity estimated after optimization, we write

$$\mathbf{E}_0^* = \mathbf{x}_0^* \mathbf{1}^{\mathrm{T}} + \tau \mathbf{A}_0 \mathbf{T}^* \tag{11}$$

Here, $\mathbf{E}_0^*$ indicates the estimated $f\left(\mathbf{x}_0 \mid \mathbf{y}_1\right)$, and corresponds to the ensemble of realizations after optimization; $\mathbf{x}_0^*$ is the vector of optimized mean states of the ensemble; τ is an inflation factor which acts on the (ensemble) anomalies; $\mathbf{T}^*$ is the optimal ensemble transform matrix, which is obtained by relying on equation (10). The most computationally intensive part during the iteration is attributed to the iterative time integration of the state ensemble through the forward model $\mathcal{M}_{1\leftarrow 0}$, at least one iteration (i.e., $k \ge 1$) being required for the computation of $f\left(\mathbf{x}_0 \mid \mathbf{y}_1\right)$.

2.3 Localization and spurious correlation

The covariance localization in the IEnKF is not straightforward (Bocquet, 2016). A domain localization scheme might introduce discontinuities across the parameter space, because grid nodes are separately analyzed by confining the approach to the use of the observations within the predefined filter length (Lorenc, 2003). To alleviate these difficulties, we rely on the following function

$$u(\mathbf{d}) = \begin{cases} \exp\left[\alpha\left(\dfrac{d_x}{\lambda_x} + \dfrac{d_z}{\lambda_z}\right)\right] & \text{when } \dfrac{d_x}{\lambda_x} \leq \beta \text{ or } \dfrac{d_z}{\lambda_z} \leq \beta \\ \exp(2\alpha\beta) & \text{when } \dfrac{d_x}{\lambda_x} > \beta \text{ and } \dfrac{d_z}{\lambda_z} > \beta \end{cases} \tag{12}$$

where $\mathbf{d} = [d_x, d_z]$ is a (spatial) lag separation vector between two points on the computational grid; λ_x and λ_z are correlation scales along the x and z directions, respectively; $\alpha > 0$ and $\beta > 0$ are constants which need to be tuned. Taking the Schur product between function (12) and the observation error enables one to magnify somehow artificially the observation error and then impact on the measurement error covariance $\mathbf{R}_1$ in equations (7), (8) and (10) as a function of the distance between a target location on the computational grid and points where measurements are available. This, in turn, can alleviate the emergence of spurious correlations in a way which is more effective than simply grounding the assimilation algorithm on the use of observations comprised within a predefined filter length (see also Nan and Wu (2011)). Therefore, the domain localization approach is used on both EnKF and IEnKF in this study.

2.4 Ensemble analysis

The analysis *pdf* $f(\mathbf{x}_1 | \mathbf{y}_1)$ is estimated by

$$\begin{aligned} f(\mathbf{x}_1 | \mathbf{y}_1) &= \int f(\mathbf{x}_1 | \mathbf{x}_0, \mathbf{y}_1) f(\mathbf{x}_0 | \mathbf{y}_1) d\mathbf{x}_0 \\ &= \int \delta\left(\mathbf{x}_1 - \mathcal{M}_{1\leftarrow 0}(\mathbf{x}_0)\right) f(\mathbf{x}_0 | \mathbf{y}_1) d\mathbf{x}_0 \end{aligned} \tag{13}$$

where δ indicates the Dirac distribution. According to equation (13), the optimized (or analysis) ensemble indicating the filtering distribution at time t_1 is obtained by propagating each member of $\mathbf{E}_0^*$ through $\mathcal{M}_{1\leftarrow 0}$, i.e., $\mathbf{E}_1^* = \mathcal{M}_{1\leftarrow 0}\left(\mathbf{E}_0^*\right)$, where $\mathbf{E}_1^*$ (associated with (ensemble) mean $\mathbf{x}_1^*$) is the best ensemble estimate based on the domain localized IEnKF algorithm. Note that $\mathbf{x}_1^*$ is then used in a new data assimilation cycle as a prior to estimate the smoothing and analysis *pdf*s.

It is remarked that at least two ensemble propagations are needed in the domain localized IEnKF, i.e., one for estimating the smoothing *pdf* $f\left(\mathbf{x}_0 \mid \mathbf{y}_1\right)$ and another one for estimating the analysis *pdf* $f\left(\mathbf{x}_1 \mid \mathbf{y}_1\right)$. In case the predefined maximum iteration number is equal to one, the structure of the IEnKF resembles the one-step-ahead smoothing EnKF (Gharamti et al., 2015).

3. Numerical simulations

3.1. Density-independent groundwater flow

We start by considering a transient groundwater flow scenario without density effects, here termed as density-independent flow (DIF) and described by

$$\nabla \cdot \left[K\left(\nabla h\right)\right] + W = S\frac{\partial h}{\partial t} \tag{14}$$

where K represents hydraulic conductivity $\left[\mathrm{LT}^{-1}\right]$, that is considered to be a random function of space; h is hydraulic head $\left[\mathrm{L}\right]$; W is a sink / source term $\left[\mathrm{T}^{-1}\right]$; S is specific storage $\left[\mathrm{L}^{-1}\right]$; and t represents time $\left[\mathrm{T}\right]$.

3.2. Variable-density groundwater flow (VDF)

For a variable-density flow (VDF), hydraulic head depends on fluid density, which is in turn a function of salt concentration (or salinity). Tackling a typical VDF problem entails jointly solving the flow and transport problems. In this context, one relies on the concept of equivalent freshwater head at a given point B, defined as

$$h_f = \frac{P_\mathrm{B}}{\rho_f g} + Z_\mathrm{B} = \frac{\rho}{\rho_f} h - \frac{\rho - \rho_f}{\rho_f} Z_\mathrm{B} \tag{15}$$

where P_B $\left[\mathrm{ML^{-1}T^{-2}}\right]$ and Z_B $\left[\mathrm{L}\right]$ respectively are pressure and elevation at B; ρ and ρ_f $\left[\mathrm{ML^{-3}}\right]$ respectively are the salt- and fresh-groundwater density. Note that $h = h_f$ for the density independent flow.

Based on the concept of equivalent freshwater head, the governing equation for VDF is expressed as (Guo and Langevin, 2002)

$$\begin{cases} \nabla \cdot \left[\rho K_f \left(\nabla h_f + \frac{\rho - \rho_f}{\rho} \nabla z \right) \right] = \rho S \frac{\partial h_f}{\partial t} + \theta \frac{\partial \rho}{\partial C} \frac{\partial C}{\partial t} - \rho_s q_s' \\ K_f = \frac{\rho_f g}{\mu_f} k \\ \rho = \rho_f + \frac{\partial \rho}{\partial C} C \end{cases} \tag{16}$$

where K_f is the freshwater hydraulic conductivity $\left[\mathrm{LT^{-1}}\right]$, that is considered to be a random function of space; k is permeability $\left[\mathrm{L^2}\right]$; μ_f and θ respectively are the dynamic viscosity for freshwater $\left[\mathrm{ML^{-1}T^{-1}}\right]$ and effective porosity $\left[-\right]$, that are considered as constants in this study; C is salt concentration $\left[\mathrm{ML^{-3}}\right]$; q_s' is a source /

sink $\left[\mathrm{T}^{-1}\right]$ of fluid with density ρ_s $\left[\mathrm{ML}^{-3}\right]$. Note that $K = K_f$ for the density-independent flow. Hereinafter, we also refer to hydraulic head and conductivity in a density-independent flow as equivalent freshwater head and conductivity (i.e., h_f and K_f), respectively.

Transport of salt is described by (e.g., Zheng and Bennett, 1995)

$$\frac{\partial(\theta C)}{\partial t} = \nabla\cdot(\theta D\cdot\nabla C) - \nabla\cdot(qC) - q_s' C_s \tag{17}$$

where q $\left[\mathrm{LT}^{-1}\right]$ is groundwater flux calculated from equation (16); D is dispersion $\left[\mathrm{M}^2\mathrm{T}^{-1}\right]$, which is here taken as a constant; and C_s is salt concentration in the sink /source term $\left[\mathrm{ML}^{-3}\right]$.

3.3. Problem setup

We analyze the performances of the domain localized iterative ensemble Kalman filter (LIEnKF) and of the domain localized standard EnKF (LEnKF) on the VDF and DIF cases using a setting which is adapted from the Henry problem, as described in the following.

3.3.1. Variable-density flow

The modified Henry problem considered in this study is depicted in Figure 1a. The size of the rectangular domain is 100 m (in the vertical direction) × 200 m (in the horizontal direction). It is discretized into 25×50 cells, each of these with a uniform size of 4 m. The upper and bottom boundaries are considered to be impermeable. The left and right boundaries are respectively defined as inland flow and sea boundaries.

A time-varying inland flow rate is prescribed at the left boundary, while a periodic tidal fluctuation is fixed along the right boundary. A partially penetrating well (located as depicted in Figure 1) is operated under a transient regime, as detailed in the following. The (natural) logarithm of equivalent freshwater hydraulic conductivity ($Y = \ln K_f$) is considered as a heterogeneous field (see Figure 1a). A random realization of the field is generated according to the procedure illustrated in the following and is used as the reference field in this study.

The spatial distribution of Y is assumed to be statistically stationary and characterized by the covariance function

$$C(\mathbf{d}) = \sigma^2 \exp\left(-\left[\frac{d_x^2}{\lambda_x^2} + \frac{d_z^2}{\lambda_z^2} \right]^{1/2} \right) \tag{18}$$

where σ^2 is the variance of Y. Generation of a reference conductivity field is performed via the well-known GSLIB software (Deutsch and Journel, 1998) by setting unit variance and $\lambda_x = 40$ m, and $\lambda_z = 24$ m. The (arithmetic) mean value μ for the reference $\ln K_f$ field is set as 6.76, which is equivalent to a hydraulic conductivity of 864 [m/day], a value typically employed for the Henry problem.

The temporal dynamics of the well pumping rate are modeled as a set of uncorrelated and randomly selected values sampled from a Gaussian distribution with mean of 570.20 [m^3/day] and standard deviation equal to 20% of the mean. We consider a temporally varying inland flow rate, uniformly distributed along the inlet. Values for these flow rate values are sampled from a Gaussian distribution with mean of 500 [m^3/day] and standard deviation equal to 20% of the mean. Tide elevations are described by a sine function with an amplitude of 4 m and a period of 30 days. Values for pumping rate and boundary conditions are generated with a daily frequency, the realization selected as input to our computations being depicted in Figure 2. As a consequence, the variant of the Henry problem we consider is characterized by a temporal alternation of confined and unconfined conditions. In the latter case, we do not account for the effects of unsaturated flow above the water table, for simplicity.

A constant salt concentration of 35 [g/L] is assigned along the sea boundary, a freshwater boundary condition being imposed along the inland boundary. The total simulation time is 50 days. Table 1 lists the model parameters that are considered as uniform in our simulations. The initial flow field and concentration distribution are depicted in Figure 1b for the variable-density groundwater flow. These initial conditions have been obtained in the absence of pumping and by setting a uniform tidal level of 100 m at the seaside boundary and a constant inland flow rate coinciding with the mean value of 570.2 [m^3/day], which is then uniformly distributed along the domain inlet (see Table 1).

3.3.2. Density-independent flow

The setting for the density-independent case is similar to the one illustrated in Section 3.3.1, the only differences being the actual values for the initial conditions considered for the flow problem. The initial distribution of pressure heads and fluxes are obtained by the same method used for the VDF case and are depicted in Figure 1c.

3.3.3. Simulation scenarios for data assimilation

The spatial distribution of the 30 points where we sample pressure heads in the reference field is depicted in Figure 1b. Pressure heads sampled at these locations are transformed into equivalent freshwater heads by equation (15) and employed for both

VDF and DIF simulations. The 50 and 10 locations at which pressure head (and salinity, for the VDF scenario) and reference hydraulic conductivity values are respectively measured in our simulations are depicted in Figure 1c. Observations of equivalent freshwater head, h_f , (natural) logarithm of equivalent freshwater hydraulic conductivity, *Y*, and/or salinity, S_a , employed in the data assimilation procedure are obtained by perturbing the corresponding reference values of the (DIF and/or VDF) scenarios considered by a zero-mean Gaussian error with a given standard deviation. The values of the latter (i.e., 0.001 m, 0.01 m, and 0.1 m for h_f measurements and 0.001 g/L for S_a) considered in our simulations enable us to assess the importance of data error on assimilation results (see also our results illustrated in Section 4). These values are partially consistent with measurement accuracies associated with some typical devices deployed in the field (e.g., water level loggers whose measurement accuracies for pressure heads can range between ~ ± 0.005 m and ~ ± 0.05 m).

To ease the interpretation of the results stemming from our analysis of the worth of diverse data types (i.e., h_f or S_a), we intentionally used a constant (in both time and space) standard deviation to characterize the error fluctuation of both h_f and S_a. Model parameters and boundary conditions which are assumed to be deterministically known are also listed in Table 1.

To explore the potential of the approaches analyzed, we consider several showcases, each highlighting key features of interest. We group our exemplary settings according to the following configurations: (*a*) Groups A (see Table 2) allows exploring the effects of a range of measurement errors, number of data, and temporal frequencies to be included in the assimilation procedure; (*b*) Group C (see Table 2) includes diverse observation types; and (*c*) Group D (Table 2) considers uncertainties linked to our incomplete knowledge of the correlation scales of the randomly heterogeneous Y field. Test Cases TCs 1-6 (each of them structured into two sub-components, e.g., TC1 articulated into $TC1_c$ and $TC1_v$, see Table 2) are designed to establish a baseline for both density-independent flow (hereafter termed DIF) and variable-density flow cases with differing numbers of h_f data assimilated in the model during the simulation period (i.e., 30 or 50 h_f data are assimilated with a daily frequency, respectively in TCs 1-3 and TCs 4-6) and considering the effect of diverse values of the standard deviation of measurement errors.

Test Case 7 (structured through the three components collected in Group B in Table 2, i.e., $TC7_i$, $TC7_{ii}$ and $TC7_{iii}$) is designed to study the limitations of domain localization in the VDF case. Test Cases 8-12 (corresponding to Group C in Table 2) are designed to investigate the effects associated with the use of diverse data sets. Due to the limitations of domain localization schemes, TCs 8-12 are assessed through the covariance localization scheme of the standard ensemble Kalman filter ($LEnKF_{cov}$) (see Appendix A). With reference to these cases, note that 50 h_f and/or 50 S_a data (associated with sample standard deviations 0.001 m and/or 0.001 g/L, respectively) are assimilated with a daily frequency. Hydraulic conductivities in TCs 10-11 (associated with a sample standard deviation of $Y = \ln K_f$ equal to 0.001 (for K_f given in (m/day)) are available at the 10 points shown in Figure 1c and are assimilated only after the first day of simulation.

With the exception of $TC7_{iii}$, TCs 1-6, $TC7_i$, $TC7_{ii}$ and TCs 8-12 are designed by generating the initial realizations of the conductivity fields considering $Y = \ln K_f$ to be normally distributed with (ensemble) mean and variance of 6.0 and 1.69, respectively, these values being different from their counterparts (i.e., 6.76 and 1.0) employed in the generation of the reference *Y* field. In TC7$_{iii}$, the guessed mean and variance of *Y* coincide with those employed in the reference field. For simplicity, the random conductivity fields employed in TCs 1-12 are generated according to the covariance function (18) with values of scale lengths equal to those used in the generation of the reference field (i.e., λ_x = 40 and λ_z = 24).

Test Cases 13-16 (Group D in Table 2) are designed to investigate the effects of uncertainties on the employed values for horizontal correlation scales of *Y* and are constructed by generating the initial realizations of the conductivity fields considering *Y* to be normally distributed with (ensemble) mean and variance equal to the values characterizing the reference conductivity field. In these cases, 50 h_f data associated with a sample standard deviation 0.001 m are assimilated with a daily frequency. The random conductivity fields employed in TC13 are generated according to the covariance function (18) with λ_x equal to the value associated with the reference field. Values of the horizonal correlation scales for TCs 14-16 are 80 m, 160 m and 20 m, respectively. Vertical correlation scales are assumed to be perfectly known for TCs 13-16.

The variants (with different data assimilation frequencies) corresponding to TC6$_c$ and TC6$_v$ (Group A in Table 2) are analyzed to assess the effect of the temporal frequency of assimilation of h_f data, respectively for both the DIF and VDF cases. All of the above test cases are performed by relying on a collection of $N = 100$ Monte Carlo (MC) replicates. The effect of the number of MC realizations on the assimilation results is explored (*a*) by performing the variants (with various N

employed) corresponding to $TC6_c$ and $TC6_v$, respectively for the DIF and VDF cases, as well as (*b*) by performing the variants (with various N employed) of TC8 when solely h_f measurements are assimilated, or (*c*) by conducting the variants (with various N employed) of TC9 when only S_a data are assimilated.

The value of the localization parameter α (see function (12)) is selected by minimizing the root-mean-square error of the estimated *N* fields of *Y* based on TCs 4-6, where 50 equivalent freshwater head data are collected and used at each assimilation step. Note that one should avoid values for *e* which are too low, mainly due to its feedback with localization (Bocquet, 2016). Here, we use *e* = 0.2 for both DIF and VDF cases. No inflation is used, i.e., $\tau = 1$ in equation (11). The root-mean-square error

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{l=1}^{m}\left(\mathrm{S}_l^{\mathrm{t}} - \overline{\mathrm{S}}_l^{\mathrm{a}}\right)^2} \tag{19}$$

is used to evaluate performances of the two data assimilation methods (Chen and Zhang 2006). Here, $\mathrm{S}_l^{\mathrm{t}}$ is the l^{th} true system state (i.e., equivalent freshwater head, salinity, or *Y*); $\overline{\mathrm{S}}_l^{\mathrm{a}}$ represents the estimated (ensemble mean) value for the l^{th} state.

The sample (ensemble) variance at the i^{th} grid cell and spread for $\ln K_f$ on the entire grid are respectively defined as

$$\mathrm{Var}_Y(i) = \frac{1}{N-1}\sum_{l=1}^{N}\left(Y_l - \overline{Y}_i^a\right)^2 \tag{20a}$$

$$\text{spread}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}\mathrm{Var}_Y(i)} \tag{20b}$$

where Y_l represents the l^{th} realization of Y on the i^{th} grid cell, $\overline{Y}_i^a$ corresponding to the associated estimated value. Filter inbreeding arises when the ensemble variance tends to artificially decrease as data assimilation proceeds in time. This effect might be related to a variety of reasons, including, e.g., the reliance on a limited number of realizations explored. The observation that the RMSE (19) is less than the spread (20b) is considered as an indicator of filter inbreeding, a large difference between these two quantities suggesting the occurrence of serious filter inbreeding.

3.4. Computational burden

We break down the evaluation of computational complexities associated with data assimilation procedures (respectively through LEnKF, LIEnKF and $\text{LEnKF}_{\text{cov}}$) at a given assimilation cycle into two components, i.e., the forecast- and the update-step. With the assumption that the number of observations N_{obs} is much smaller than the size of the state vector $\mathbf{x}$, these computational complexities are analyzed and shown in Table 3. It can be noted that the use of LIEnKF is computationally equivalent to the LIEnKF in Sakov et al. (2012), and is more intensive than the use of LEnKF and $\text{LEnKF}_{\text{cov}}$. However, it is worth noting that the required CPU times for ensemble propagations can be efficiently decreased by performing parallel computations. With 12 processors, CPU times required to perform TC6_{v} (corresponding to the variable-density flow scenario) are 29344 s (about 8 hours) and 63578 s (about 18 hours), respectively by relying on LEnKF and LIEnKF. It is worth noting that resorting to model reduction techniques (Li et al., 2013) can contribute to decreasing the required the CPU time for propagating ensemble, at the expenses of accuracy loss.

4. Results and discussion

Table 4 lists the results of the process of tuning the value of α based on TCs 4-6 for the diverse values of the magnitude of data error considered. We can note that the values of α obtained for the VDF cases are higher than their counterparts for the DIF cases, thus implying that a higher level of localization is required in the VDF than in the DIF settings. These results suggest that strong nonlinearity and large dimensionality of the state vector $\mathbf{x}$ can aggravate the occurrence of spurious correlations (see also Houtekamer and Mitchell, 1998). The final calculated RMSE and spread values for Y in various test cases are depicted in Figure 3, with exceptions of TC7. We illustrate and discuss our results in details in the following Sections.

4.1. Effect of data quantity and measurement error on simulation (TCs 1-6)

As expected, one can note that RMSE values in Figures 3a, b for TCs 1-6 (Group A in Table 2) are generally lowest in the presence of reduced measurement error for both assimilation methods used, regardless the amount of data assimilated at each time step. Values of spread are general consistent with (and mostly slightly higher than) those of RMSE. When the standard deviation of head observation error is 0.001 m, the optimal value of α is set as 4 to reduce spurious correlation. One can note (see Table 4) that such a large value for α can be directly tied to the high accuracy of the data (as reflected by a low measurement error) which can in turn aggravate the emergence of spurious correlations. At the same time, a large value of α can contribute to dampen the effect of some otherwise informative data, while reducing the effect of spurious correlations. The lower RMSE value observed for LIEnKF in $TC6_v$, as opposed to $TC5_v$, can be a consequence of such contrasting effects.

Houtekamer and Mitchell (1998) pointed out that a strong nonlinearity of the setting and a large dimension of the state vector can aggravate the occurrence of spurious correlations. Nan and Wu (2011) used various filter lengths on different types of data to reduce such spurious correlations. Here, we document that the effect of spurious correlation on assimilation results can change with observation accuracy.

One can note that LIEnKF generally outperforms LEnKF in terms of RMSE values in all of the TCs analyzed here. It is worth noting that RMSE values for the VDF TCs are generally lower than their counterparts related to the DIF TCs, even as the VDF setting is associated with a higher nonlinearity than the DIF case. We observe that the data assimilation performance is also affected by the quality and amount of available observations. Based on equation (15), it is clear that h_f data in the VDF settings are associated with an information value that is higher than in the DIF case, because they also embed salinity information.

Figure 4 depicts the initial ensemble-averaged (i.e., as a result of averaging across the generated 100 Monte Carlo realizations) *Y* field (Figure 4a), the reference *Y* field (Figure 4b), and the ensemble-averaged *Y* fields obtained at the end of the data assimilation process for $TC6_c$ (Figures 4c, d) and $TC6_v$ (Figures 4e, f). The estimated spatial patterns of the average *Y* fields for $TC6_c$ and $TC6_v$ are similar and close to the reference one, consistent with the RMSE results depicted in Figures 3a, b. These results suggest that LIEnKF leads to consistent estimates of the *Y* field for both DIF and VDF settings. Even as the RMSE associated with LEnKF is somewhat higher than its counterpart resulting from LIEnKF, it is noted that LEnKF can lead to a reasonably good estimate of *Y* field.

Figures 5a, b depict the spatial patterns of the variance of Y in $TC6_c$ at the end of the assimilation period, respectively for LEnKF and LIEnKF. Corresponding results for $TC6_v$ are depicted in Figures 5c, d. These results show that the values of Y variance for density-independent groundwater flow ($TC6_c$) are clearly influenced by the given head boundary conditions, a finding which is consistent with the results of Tong et al. (2010). The spatial pattern obtained for $TC6_v$ is significantly different from that for $TC6_c$. The results in Figures 5c, d indicate that the highest values of Y variance lie within regions that are clearly related to the intruding salt-water wedge. This result is associated with the influence of the tidal and inflow conditions acting along the seaside boundary. The reduced values of variance observed within the remaining portion of the domain suggest that the value of information associated with head data is higher in the VDF than that in the corresponding DIF case. These results are consistent with the findings by Shoemaker (2004) and Sanz and Voss (2006). Shoemaker (2004) pointed out that flux observations in the submarine zone are useful for the estimation of model parameters, including hydraulic conductivity. In this context, one can note that the spatial distributions of the Y variance obtained in our settings are consistent with the pattern of scaled sensitivity of pressure to permeability depicted in Figure 2 of Sanz and Voss (2006). The lowest variances are associated with locations around the pumping well for the DIF TCs. Otherwise, the lowest variances are found in the proximity of the pumping well as well as of the transition zone in the VDF TCs. These results further support the conclusions that (*a*) h_f data in a VDF scenario contain not only information about hydraulic head, but also about salinity condition, and (*b*) a reasonably realistic spatial distribution of Y can be estimated even in the presence of the high nonlinearity associated with the VDF setting.

4.2. Impact of data type on assimilation result for variable-density flow (TCs 7-12)

The values of α and β in $TC7_i$ (see Group B in Table 2) coincide with those used in $TC6_v$, while their counterparts in $TC7_{ii}$ are set as 30 and 0.2 to avoid effects of spurious correlation. Figure 6 depicts the spatial distribution of the ensemble mean of Y obtained after the first assimilation step (or the first filter iteration) in $TC7_i$ and $TC7_{ii}$ (note that the values of $\ln K_f$ in Figure 6 are higher than those in Figure 4). These results reveal that some of the updated ensemble mean Y values obtained after the first assimilation step in $TC7_i$ and $TC7_{ii}$ markedly differ from their counterparts in the reference field (and/or initial guessed field, see Figures 4a, b) for both approaches. Note that we report only values after the first step because some of the updated Y values markedly differ from values at their neighbor cells. Such a strong contrast across the Y fields can cause severe numerical issues in the simulation of density-dependent flow systems.

One can note that the largest deviations between reference and estimated Y values in Figure 6 are confined to a region where salinity information is lacking (but flow dynamic are strong due to the action of the pumping well) during the first day of assimilation (see also Figure 1b). The RMSE values for Y in $TC7_i$ after the first assimilation step for LEnKF and LIEnKF are 1.89 and 1.37, respectively, which are very different from the initially guessed one. The reason for this behavior is that the domain localization scheme corresponds to a collection of minimized local cost functions that might have not converged to the global cost function. Therefore, the local scheme may not guarantee convergence of results.

We then consider $TC7_{iii}$ where the values of α and β correspond to those of $TC7_i$ and the initial mean and variance of *Y* are respectively set as 6.76 and 1.0, corresponding to the values associated with the reference field. The calculated *Y* RMSE values after the final assimilation time are 0.39 and 0.28 for LEnKF and LIEnKF, respectively. One can then conclude that both of these methods provide viable solutions in the presence of low nonlinearities (i.e., perfect initial guesses for the mean and variance of *Y*).

We note that some of the results obtained in the previous TCs could be influenced by considering that domain localization schemes might lack a guaranteed convergence (see also Bocquet, 2016). A covariance localized EnKF ($\mathrm{LEnKF_{cov}}$, see Appendix A) is then developed and applied in TCs 8-12 (Group C in Table 2) to investigate this issue for various data types (i.e., h_f, S_a and *Y*) considered in the assimilation process.

Test Cases 8 and 9 are analyzed through $\mathrm{LEnKF_{cov}}$ and are respectively based on assimilating solely h_f or S_a data. The values of RMSE obtained for the (ensemble) average *Y* field at the end of the assimilation period are 0.60 and 0.73, respectively for TC8 and TC9. The spatial distributions of the (ensemble) average and variance of *Y* are depicted in Figure 7. One can note that these results differ from those of $TC6_v$ (see Figures 4c, d), even as these two cases are characterized by very similar values of RMSE. One can also observe the occurrence of high values of variance embedded within a generally low variance field (see Figure 7b) and mainly related to localization. The high variance regions clearly visible in Figure 7d (i.e., the upper left and lower right corners in the figure) are in the areas within which salinity variation is small (i.e., salinity values are about 0 or 35 g/L).

Spatial distributions of (ensemble) averaged h_f and S_a in TC8 and TC9 are respectively depicted in Figures 8 and 9 for early, intermediate and late assimilation time periods. These results suggest that the updated h_f and S_a distributions obtained by assimilating h_f agree well with their reference counterparts. Otherwise, when only S_a data are assimilated, the updated S_a distributions agree well with their reference counterpart (Figures 9d, e and f), but h_f distributions do not (Figures 8d, e and f). The results are similar to what we observed in Section 4.1, because h_f is informative to hydraulic / flow conditions as well as to salinity distributions. This supports the conclusion that h_f data are more informative as compared with S_a data. Based on the results for TCs 7-9, we conclude that both LEnKF and LIEnKF suffer from local convergence problems, so that global convergence in the whole domain is not guaranteed.

Figure 10 depicts the temporal variations of the RMSE and spread associated with Y for TCs 8-12. These results suggest that assimilation of h_f data (TC8) would lead to optimal results in terms of RMSE values. Otherwise, jointly assimilating h_f and S_a (TC12) information would not improve the assimilation results, but can potentially deteriorate the performance of $\mathrm{LEnKF_{cov}}$ (see the values of spread for Y in Figure 10b). There are two main reasons for this latter result. The first one is that there is some data redundancy between h_f and S_a. Thus, jointly assimilating both h_f and S_a data does not necessarily imply that information content is increased. One should note that jointly assimilating h_f and S_a data could lead to more severe underestimation of the covariance (A2), leading to a more severe filter inbreeding possibility in TC12 than in TCs 8-11. This effect can then cause RMSE for TC12 to decrease first and then continuously increase with time. We remark that jointly assimilating h_f and Y (TC10), or S_a and Y (TC11) measurements would also yield more severe filter inbreeding issues than those observed in TC8 and TC9. These findings are consistent with those obtained by Hendricks-Franssen and Kinzelbach (2008), who pointed out that jointly assimilating hydraulic head and conductivity data in a classical groundwater flow model would render the system prone to filter inbreeding.

4.3. Effect of temporal frequency of data assimilation

Here, we consider TC6 and its variants, constructed by relying on differing data assimilation frequencies. Figure 11 depicts the temporal evolution of RMSE associated with Y and h_f for LEnKF and LIEnKF in $TC6_c$ and its variants constructed by considering data assimilation frequencies corresponding to 5 and 10 days (in the DIF setting). Figure 12 depicts corresponding results for $TC6_v$ and its variants constructed by considering the same data assimilation frequencies in the VDF setting, i.e., the temporal evolutions of RMSE related to Y (Figures 12a, b), h_f (Figures 12c, d) and S_a (Figures 12e, f) for LEnKF and LIEnKF. Note that only RMSE results with assimilation frequencies of 1 day and 5 days are depicted for the complete temporal window of assimilation because the $TC6_v$ variant (with assimilation frequency corresponding to 10 days) displayed a filter convergence issue at the first assimilation step (which corresponds to day 10). Both assimilation techniques are linked to large RMSE values for Y, S_a and h_f at the initial assimilation steps, a feature which is mainly due to the initial (head and salinity) conditions that are markedly different from the reference (head and salinity) fields (see the high RMSE values for head and salinity in Figures 12a, b). One can see that daily assimilation yields the best performance, in terms of RMSEs, for both schemes and for both (DIF and VDF) cases.

Comparing Figures 12c, d against Figures 11b, d reveals that RMSE values for h_f are much higher in the VDF than in the DIF case. With identical DA frequency and Gaussian priors, VDF scenarios are not only subject to the increased nonlinearity of the model, in comparison with DIF cases, but are also associated with uncertainty stemming from both h_f and S_a. When the data assimilation frequency is decreased to incorporate observations at a 10 days interval, the quality of the simulation results tends to deteriorate drastically with time. This behavior would in turn increase uncertainty, as an effect of nonlinear system behavior, thus reflecting on system states (i.e., h_f and S_a), so that neither LIEnKF nor LEnKF are effective. This result suggests that a high observation frequency would be critically beneficial to VDF settings, especially in the presence of uncertain initial conditions.

4.4. Effect of the ensemble size

Here, we consider TCs 8 and 9 and their variants, constructed by relying on differing size of the collection of realizations. Figure 13 shows the effect of the ensemble size on assimilating h_f (TC8 with N = 100, and its variants with N = 300, and 1000) and S_a data individually (TC9 with N = 100, its variants with N = 300, and 1000). The final values of RMSE obtained for Y in TC8 and its two variants are 0.60, 0.58 and 0.57, and for TC9 and its two variants are 0.73, 0.64 and 0.59, respectively. It should be pointed out that the ensemble size has a more pronounced effect in the cases where solely S_a measurements are assimilated. This finding suggests that assimilating only S_a data would yield a more severe filter inbreeding problem than assimilating solely h_f data. When 1000 Monte Carlo realizations are employed, final RMSE values for Y are 0.57 and 0.59 for the TC8 and TC9 variants, respectively. This result suggests that h_f or S_a data provide a similar information content when their assimilation is targeted to estimate hydraulic conductivity. Otherwise, when solely assimilating S_a data, increasing the ensemble size yields no visible improvement for the estimation of equivalent freshwater head.

4.5. Effect of uncertain correlation scale of *Y* (TCs 13-16)

Figure 14 depicts the temporal variation of RMSE and spread for TCs 13-16 (Group D in Table 2). It is interesting to note that filter inbreeding becomes increasingly serious for settings corresponding to imposed values of λ_x which are larger than those characterizing the reference Y field) (TCs 14-15), a finding which is consistent with the results of Camporese et al. (2011). The reason for this is related to the observation that a constant correlation scale is used in function (12) during the data assimilation process, while the spatial correlation across the estimated ensemble might change as time elapses. One can also note that when we rely on a value of λ_x which is smaller than the one corresponding to the reference Y field (TC16), filter inbreeding might be small, the final RMSE being larger than the one observed in TC13, where the true value of λ_x is assumed (see Group D in Table 2 and Figure 3d). We remark that even as relying on a small value for λ_x can somehow alleviate filter inbreeding, it can also shadow the importance of useful information that can assist in improving the data assimilation performance. In this context, the use of algorithms that can either temporally adjust estimates of correlation scales (see, e.g., Anderson and Lei, 2013; Ménétrier et al., 2015), or determine the adequate inflation level to be enforced on the state covariance (e.g., Wang and Bishop, 2003; Zovi et al., 2017) can be beneficial when dealing with uncertain correlation scales.

5. Summary and Conclusion

In this study, a variant Henry problem is used to investigate the performances of domain localization schemes of iterative ensemble Kalman filter (LIEnKF) and ensemble Kalman filter (LEnKF), as well as the covariance-localized scheme of the ensemble Kalman filter ($LEnKF_{cov}$) in a variable density groundwater flow (VDF) scenario. As a baseline setting, the performances of both LEnKF and LIEnKF are assessed in the absence of density effects (here termed as DIF scenario). Results are compared and analyzed by considering diverse values for the magnitude of measurement errors, data quantity and type, assimilation frequency, size of the collection of Monte Carlo realizations employed and considering incomplete knowledge of the correlation length of the (randomly) heterogeneous conductivity field of the porous medium. Our numerical study leads to the following major conclusions.

1) Even as VDF is characterized by a higher nonlinearity and dimension of system state vector than DIF, the use of either LEnKF or LIEnKF yields lower RMSE

values (which indicates a good data assimilation performance) in VDF than in DIF settings (see Section 4.1).

2) Our results suggest that equivalent freshwater head, h_f, contains more information in VDF than in DIF. This is related to the observation that h_f also contains information about fluid density. Optimal locations for h_f observations in the DIF settings examined in this study should correspond to regions (*a*) far away from the given head boundary and (*b*) in the proximity of the pumping well. Placing of observation points for h_f which are effective to data assimilation purposes in the VDF scenario are seen to be set around the pumping well as well as in the proximity of the transition zone.

3) We note that when the data assimilation frequency is equal to the lowest one here tested (i.e., corresponding to an assimilation every 10 days) the quality of the simulation results tends to deteriorate drastically with time. This corresponds to increased uncertainty as a result of system nonlinearity, so that neither LIEnKF nor LEnKF are effective. Our results suggest that a high temporal observation frequency would be critically beneficial to VDF settings, especially in the presence of uncertain initial conditions.

4) Overestimating the horizonal correlation length of the heterogeneous log-conductivity field can lead to filter inbreeding issues. Relying on underestimated correlation scale values can alleviate filter inbreeding, while negatively affecting the overall performance of the data assimilation process.

5) We have implemented a covariance-localized ensemble Kalman filter ($\mathrm{LEnKF_{cov}}$) to investigate the value of data collected in the density-driven groundwater flow

scenario. Our results (see TC8 and TC9) suggest that h_f data are associated with an information content that is superior to that of S_a for an accurate estimation of h_f, S_a, and K_f. In this sense, it is seen that these latter quantities could be updated with good accuracy using only h_f observations. Otherwise, equivalent freshwater head distributions are not properly estimated when using only S_a data. Additionally, we note that increasing the ensemble size shows virtually no improvement in the accuracy of the updated h_f when only S_a data are assimilated.

6) The domain localization schemes of the ensemble Kalman filter (LEnKF) and iterative ensemble Kalman filter (LIEnKF) we analyze suffer from convergence issues associated with the global optimization when only S_a data are used (see our results for TCs 7-9). This issue is not seen when (*a*) only h_f data are used, or (*b*) solely S_a data are used in the presence of an initial collection of *Y* fields characterized by the (ensemble) statistics closely corresponding to those of the reference *Y* realization (see, i.e., TC7$_{iii}$). The performance of the domain localization scheme can be evaluated by considering that it can be deconstructed into the following two steps: (1) first, one needs to complete all local optimizations, i.e., by minimizing the localized cost function at each node of the computational grid; (2) then, global optimization can be assessed by collecting all locally optimal results. Note that this is related to the observation that the domain localization scheme is not characterized by a guaranteed convergence of the global optimization. This issue is mainly attributed to the fact that the collected local optimization results do not necessarily converge to the results which would

be obtained by minimizing the global cost function. Employing a covariance localization scheme to minimize the localized global cost function enables us to significantly alleviate the problem.

Acknowledgments

This work was supported by the National Key Research Project (Grant No. 2016YFC0402805). Alberto Guadagnini would like to thank the EU and MIUR for funding, in the frame of the collaborative international Consortium (WE-NEED) financed under the ERA-NET WaterWorks2014 Cofunded Call. This ERA-NET is an integral part of the 2015 Joint Activities developed by the Water Challenges for a Changing World Joint Programme Initiative (Water JPI). The codes and model files can be obtained by contacting the author: Chuan-An Xia (c.a.xia@cugb.edu.cn).

We would like to thank editor Martyn P. Clark and the three anonymous reviewers for their constructive comments.

Appendix A

We perform covariance localization of the ensemble Kalman filter using the following local function (Furrer and Bengtsson, 2007) for the Schur product with the covariance (A2)

$$\begin{cases} L = \exp\left(-3\left[\dfrac{d_x}{\gamma_x} + \dfrac{d_z}{\gamma_z}\right]\right) \\ \gamma_i = \delta_i\left(\sqrt{9+8N} - 5\right)/4; \quad (i = x, z) \end{cases} \tag{A1}$$

where γ_x and γ_z are the set filter length along the *x*- and *z*- direction for covariance or cross-covariance

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{h_f} & \mathbf{C}_{h_f S_a} & \mathbf{C}_{h_f Y} \\ \mathbf{C}_{S_a h_f} & \mathbf{C}_{S_a} & \mathbf{C}_{S_a Y} \\ \mathbf{C}_{Y h_f} & \mathbf{C}_{Y S_a} & \mathbf{C}_{Y} \end{bmatrix} \tag{A2}$$

where $\mathbf{C}_{h_f}$, $\mathbf{C}_{S_a}$ and $\mathbf{C}_Y$ are covariance matrices for equivalent freshwater head h_f, salinity S_a, and (natural) logarithm of hydraulic conductivity *Y*; $\mathbf{C}_{h_f Y}$ is the cross-covariance between h_f and *Y*; $\mathbf{C}_{S_a Y}$ is the cross-covariance between S_a and *Y*; and $\mathbf{C}_{h_f S_a}$ is the cross-covariance between h_f and S_a. Here, similar to Nan and Wu (2011), we consider

$$\delta_i\left(h_f, h_f\right) = \lambda_i / 2; \qquad \delta_i\left(S_a, S_a\right) = \lambda_i / 2; \qquad \delta_i\left(\ln K_f, \ln K_f\right) = \frac{1}{4}\delta_i\left(h_f, h_f\right) \tag{A3a}$$

$$\delta_i\left(h_f, S_a\right) = \sqrt{\delta_i\left(h_f, h_f\right)\delta_i\left(S_a, S_a\right)} \tag{A3b}$$

$$\delta_i\left(h_f, \ln K_f\right) = \sqrt{\delta_i\left(h_f, h_f\right)\delta_i\left(\ln K_f, \ln K_f\right)} \tag{A3c}$$

$$\delta_i\left(S_a, \ln K_f\right) = \sqrt{\delta_i\left(S_a, S_a\right)\delta_i\left(\ln K_f, \ln K_f\right)} \tag{A3d}$$

References

Anderson, J. L., and L. Lei (2013), Empirical localization of observation impact in ensemble Kalman filters, Mon. Weather Rev., 141, 4140-4153.

Beaujean, J., F. Nguyen, A. Kemna, A. Antonsson, and P. Engesgaard (2014), Calibration of seawater intrusion models: Inverse parameter estimation using surface electrical resistivity tomography and borehole data, Water Resour. Res., 50, 6828-6849, doi:10.1002/2013WR014020.

Bocquet, M. (2011), Ensemble Kalman filtering without the intrinsic need for inflation, Nonlinear Proc. Geophys., 18,735-750, doi: 10.5194/ npg-18-735-2011/.

Bocquet, M., and P. Sakov (2012), Combining inflation-free and iterative ensemble Kalman filters for strongly nonlinear systems, Nonlinear Proc. Geophys., 19, 383-399, doi: 10.5194/npg-19-383-2012.

Bocquet, M., and P. Sakov (2014), An iterative ensemble Kalman smoother, Q. J. R. Meteorol. Soc., 140, 1521-1535, doi: 10.1002/qj.2236.

Bocquet, M. (2016), Localization and the iterative ensemble Kalman smoother, Q. J. R. Meteorol. Soc., 142, 1075-1089, doi:10.1002/qj.2711.

Creel, L. (2003), Ripple effects: population and coastal regions, Population Reference Bureau, Washington, DC, pp, 1-7.

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

Carrera, J., J. J. Hidalgo, L. J. Slooten, and E. Vázquez-Suñé (2010), Computational and conceptual issues in the calibration of seawater intrusion models, Hydrogeol. J., 18, 131-145.

Camporese, M., G. Cassiani, R. Deiana, and P. Salandin (2011), Assessment of local hydraulic properties from electrical resistivity tomography monitoring of a three-dimensional synthetic tracer test experiment, Water Resour. Res., 47, W12508, doi:10.1029/2011WR010528.

Colombani, N., G. Volta, A. Osti, and M. Mastrocicco (2016), Misleading reconstruction of seawater intrusion via integral depth sampling, J. HYDROL., 536, 320-326.

Crestani, E., and M. Camporese, and P. Salandin (2015), Assessment of hydraulic conductivity distributions through assimilation of travel time data from ERT-monitored tracer tests, Adv. Water Resour., 84, 23-36.

Deutsch, C. V. and A. G. Journel (1998), GSLIB: geostatistical software library and user’s guide, 2nd edn. Oxford University Press, New York, 369 pp.

Datta, B., and G. Kourakos (2015), Preface: Optimization for groundwater characterization and management, Hydrogeol. J., 23, 1043-1049.

Doherty, J. (2002), PEST: Model-independent parameter estimation user manual, 5th edn., Watermark Numerical Computing, Brisbane, Queensland, Australia.

Evensen, G. (2003), The ensemble kalman filter: theoretical formulation and practical implementation. Ocean Dynam., 53(4), 343-367.

Evensen, G. (2009), Data Assimilation: The Ensemble Kalman Filter (2nd edn). Springer-Verlag: Berlin, Heildelberg

El-Kaliouby, H., and O. Abdalla (2015), Application of time-domain electromagnetic method in mapping saltwater intrusion of a coastal alluvial aquifer, North Oman,J. APPL. GEOPHYS., 115, 59-64.

Furrer, R., and T. Bengtsson (2007), Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants, J. Multivariate Anal, 98(2):227-255.

Gharamti, M. E., I. Hoteit, and J. Valstar (2013), Dual states estimation of a subsurface flow-transport coupled model using ensemble Kalman filtering, Adv. Water Resour., 60, 75-88, doi: 10.1016/j.advwatres.2013.07.011.

Gharamti, M. E., and I. Hoteit (2014), Complex step-based low-rank extended Kalman filtering for state-parameter estimation in subsurface transport models, J. Hydrol., 509, 588-600, doi: 10.1016/j.jhydrol.2013.12.004.

Gharamti, M. E., A. Kadoura, J. Valstar, S. Sun, and I. Hoteit (2014), Constraining a compositional flow model with flow‐chemical data using an ensemble‐based kalman filter, Water Resour. Res., 50(3), 2444-2467.

Gharamti, M.E., B. Ait-El-Fquih, and I. Hoteit (2015), An iterative ensemble Kalman filter with one-step-ahead smoothing for state-parameters estimation of contaminant transport models, J. HYDROL., 527, 442-457.

Gu, Y., and D.S. Oliver (2007), An iterative ensemble Kalman filter for multiphase fluid flow data assimilation, Spe. J., 12 (04), 438-446.

Guo, W., and C. D. Langevin (2002), User's Guide to SEAWAT: A Computer Program for Simulation of Three-Dimensional Variable-Density Ground-Water Flow, U.S. Geological Survey, Tallahassee, Florida.

Hu B. X, Y. Z. Cao, W. D. Zhao, and F. Bao (2016), Identification of hydraulic conductivity distributions in density dependent flow fields of submarine groundwater discharge modeling using adjoint-state sensitivities. Sci. China Ser. D, 59: 770–779, doi: 10.1007/s11430-015-5236-x.

Huang, C., B. X. Hu, X. Li, and M. Ye (2009), Using data assimilation method to calibrate a heterogeneous conductivity field and improve solute transport prediction with an unknown contamination source, Stoch. Environ. Res. Risk Assess., 23:1155-1167.

Houtekamer, P.L. and H. L. Mitchell (1998), Data assimilation using an ensemble Kalman filter technique. Mon. Weather Rev. 126(3), 796-811.

Houtekamer, P. L., and H. L. Mitchell (2001), A sequential ensemble Kalman filter for atmospheric data assimilation, Mon. Weather Rev., 129, 123-137.

Hamill, T. M., J. Whitaker, and C. Snyder (2001), Distance-dependent filtering of background error covariance estimates in an ensemble kalman filter. Mon. Weather Rev., 129(11), 2776-2790.

Hunt, B. R., E. J. Kostelich, and I. Szunyogh (2007), Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter. Physica. D., 230, 112-126.

Hendricks Franssen, H. J., and W. Kinzelbach (2008), Real-time groundwater flow modeling with the Ensemble Kalman Filter: Joint estimation of states and parameters and the filter inbreeding problem, Water Resour. Res., 44(9), 354-358, doi:10.1029/2007WR006505.

Jazwinski, A. H. (1970), Stochastic processes and filtering theory. New York: Academic Press.

Kourgialas, N. N., Z. Dokou, G. P. Karatzas, G. Panagopoulos, P. Soupios, A. Vafidis, E. Manoutsoglou, and M. Schafmeister (2016), Saltwater intrusion in an irrigated agricultural area: combining density-dependent modeling and geophysical methods. Environ. Earth Sci., 75(1), 1-13, doi: 10.1007/s12665-015-4856-y

Krymskaya, M. V, R. G. Hanea, and M. Verlaan (2009), An iterative ensemble Kalman filter for reservoir engineering applications. Comput. Geosci., 13 (2), 235-244.

Li, G., and A. Reynolds (2007), An iterative ensemble Kalman filter for data assimilation. In: SPE Annual Technical Conference and Exhibition.

Liu, G., Y. Chen, and D. Zhang (2008), Investigation of flow and transport processes at the MADE site using ensemble Kalman filter, Adv. Water Resour., doi: 10.1016/j.advwatres.2008.03.006.

Lorenc, A. C. (2003), The potential of the ensemble Kalmanfilter for NWP: a comparison with 4D-Var. Q. J. R. Meteorol Soc. 129, 3183-3203.

Li, X., X. Chen, B. X. Hu, and I. M. Navon (2013), Model reduction of a coupled numerical model using proper orthogonal decomposition, J. HYDROL., 507, 227-240.

Moradkhani, H., S. Sorooshian, H. V. Gupta, and P. R. Houser (2005), Dual state–parameter estimation of hydrological models using ensemble kalman filter. Adv. Water Resour., 28(2), 135-147.

Ménétrier, B., T. Montmerle, Y. Michel, and L. Berre (2015), Linear filtering of sample covariances for ensemble-based data assimilation. Part I: Optimality criteria and application to variance filtering and covariance localization, Mon. Weather Rev., 143: 1622–1643.

Nan, T., and J. Wu (2011), Groundwater parameter estimation using the ensemble Kalman filter with localization, Hydrogeol. J., 19, 547-561, doi: 10.1007/s10040-010-0679-9.

Ng, G.-H. C., D. R. Bedford, and D. M. Miller (2014), A mechanistic modeling and data assimilation framework for Mojave Desert ecohydrology, Water Resour. Res., 50, 4662-4685, doi:10.1002/2014WR015281.

Pidlisecky, A., T. Moran, B. Hansen, and R. Knight (2015), Electrical Resistivity Imaging of Seawater Intrusion into the Monterey Bay Aquifer System, Groundwater, 54, 255-261.

Poeter, E.P., M. C. Hill, E. R. Banta, S. Mehl, and S. Christensen (2005), UCODE_2005 and six other computer codes for universal sensitivity analysis, calibration, and uncertainty evaluation. US Geol. Surv. Tech. Methods 6-A11, 283 pp.

Post, V, H. Kooi, and C. Simmons (2007), Using Hydraulic Head Measurements in Variable-Density Flow Analyses, ground water, 45, 664-671.

Panzeri, M., M. Riva, A. Guadagnini, and S. P. Neuman (2013), Data assimilation and parameter estimation via ensemble Kalman filter coupled with stochastic moment equations of transient groundwater flow, Water Resour. Res., 49, 1334-1344, doi:10.1002/wrcr.20113.

Panzeri, M., M. Riva, A. Guadagnini, and S.P. Neuman (2014), Comparison of Ensemble Kalman Filter Groundwater-Data Assimilation Methods Based on

Stochastic Moment Equations and Monte Carlo Simulation, *Adv. Water Resour.*, 66, 8-18, doi:10.1016/j.advwatres.2014.01.007.

Panzeri, M., M. Riva, A. Guadagnini, and S.P. Neuman (2015), EnKF coupled with groundwater flow moment equations applied to Lauswiesen aquifer, Germany, *J. Hydrology*, 521, 205-216, doi:10.1016/j.jhydrol.2014.11.057.

Reynolds, A. C., M. Zafari, and G. Li (2006), Iterative forms of the ensemble kalman filter, In: 10th European Conference on the Mathematics of Oil Recovery.

Sreekanth, J., and B. Datta (2015), Review: Simulation-optimization models for the management and monitoring of coastal aquifers, Hydrogeol. J., 23, 1155-1166.

Sanford, W. E., J. P. Pope, and D. L. Nelms (2009). Simulation of groundwater-level and salinity changes in the eastern shore, virginia. US Geol. Surv. Sci. Invest. Rep. 2009-5066.

Sanford, W. E., and J. P. Pope (2010), Current challenges using models to forecast seawater intrusion: lessons from the Eastern Shore of Virginia, USA. Hydrogeol. J., 18(1), 73-93.

Shoemaker, W. B., (2004), Important Observations and Parameters for a Salt Water Intrusion Mode, Ground water, 42, 829-840.

Shi, L. S., X. H. Song, J. X. Tong. Y. Zhu, and Q. R. Zhang (2015), Impacts of different types of measurements on estimating unsaturated flow parameters, J. HYDROL., 524, 549-561.

Sanz, E., and C. I. Voss (2006), Inverse modeling for seawater intrusion in coastal aquifers: insights about parameter sensitivities, variances, correlations and estimation procedures derived from the Henry problem, Adv. Water Resour., 29(3), 439-457.

Shalev, E., A. Lazar, S. Wollman, S. Kington, Y. Yechieli, and H. Gvirtzman (2009), Biased Monitoring of Fresh Water-Salt Water Mixing Zone in Coastal Aquifers, Groundwater, 47, 49-56.

Sakov, P., D. S. Oliver, and L. Bertino (2012), An iterative EnKF for strongly nonlinear systems, Mon. Weather Rev., 140, 1988-2004.

Song, X., L, Shi, M. Ye, J. Yang, and I. M. Navon (2014), Numerical comparison of iterative ensemble Kalman filters for unsaturated flow inverse modeling, Vadose Zone J. 13(2), doi:10.2136/vzj2013.05.0083

Tong, J, B. X. Hu, and J. Yang (2013), Data assimilation methods for estimating a heterogeneous conductivity field by assimilating transient solute transport data via ensemble Kalman filter. Hydrol. Process, 27: 3873-3884.

Tong, J., B. X. Hu, and J. Yang (2012), Assimilating transient groundwater flow data via a localized ensemble Kalman filter to calibrate a heterogeneous conductivity field. Stoch. Environ. Res. Risk Assess., 26: 467-478.

Tong, J., B. X. Hu, and J. Yang (2010), Using data assimilation for transient flow to calibrate a heterogeneous conductivity field via ensemble Kalman filter, Stoch. Environ. Res. Risk Assess., 24: 1211-1223.

Whitaker, J. S. and T. M. Hamill (2002), Ensemble Data Assimilation without Perturbed Observations, Mon. Weather Rev., 130 (7) :1913-1924.

Wang, X. and C. H. Bishop (2003), A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. J. Atmos. Sci. 60, 1140-1158.

Wang, Y., G. Li, and A. C. Reynolds (2010), Estimation of depths of fluid contacts by history matching using iterative ensemble-kalman smoothers. SPE J., 15(15), 509-525.

Werner, A. D., M. Bakker, V. E. Post, A. Vandenbohede, C. Lu, B. Ataie-Ashtiani, C. T. Simmons, and D. A. Barry (2013), Seawater intrusion processes, investigation and management: recent advances and future challenges, Adv. Water Resour., 51, 3-26.

Wen, X.H., and W. H. Chen (2006), Real-time reservoir model updating using ensemble Kalman filter with confirming option, Spe. J., 11 (04), 431-442.

Zhang, D. (2002), Stochastic Methods for Flow in Porous Media: Copying with Uncertainties. Academic, San Diego

Zheng, C., and G. D. Bennett (1995), Applied contaminant transport modeling, theory and practice: Van Nostrand Reinhold, 440 p.

Zovi, F., M. Camporese, J. H. Hendricks Franssen, J. A. Huisman, and P. Salandin (2017), Identification of high-permeability subsurface structures with multiple point geostatistics and normal score ensemble Kalman filter, J. HYDROL., 548, 208-224.

Figure 1. Schematic representation of the modified Henry problem considered in this study: (a) type of boundary conditions, location of the partially penetrating pumping well (the well screen being denoted as a black block), reference $Y = \ln K_f$ field; (b) initial flow field and concentration distribution (contour lines corresponding to initial salinity of 10%, 50% and 90% are respectively depicted as cyan, blue and red curves) for the variable-density flow (VDF) scenario. Initial spatial distribution of flux vectors (color scale ranges from blue to red, respectively denoting low to high flux norm) for VDF scenario; (c) initial spatial distribution of flux vectors for density-independent flow (DIF) scenario. Location of the 30 observation points for pressure head are shown in (b), location of the 50 observation points for head (and salinity for VDF) being included in (c), black blocks along the boreholes corresponding to measurement locations. Red squares in (c) represent the 10 locations where Y data are collected.

Figure 2. Values for pumping rate and boundary conditions selected as input to our computations: (a) pumping rates (blue curve) and inland boundary flow rates (here given in (m^3/day)); (b) tidal elevations (given in (m)).

Figure 3. RMSE and spread values collected according to the showcase groups of Table 2 (excluding Group B): ((a) and (b) for density-independent (DIF) and density-dependent flow scenarios (VDF), respectively) Group A; (c) Group C; and (d) Group D.

Figure 4. Initial (a) ensemble-averaged and (b) reference Y field; reference and ensemble-averaged Y fields obtained at the end of the data assimilation process for

$TC6_c$ ((c) and (d) corresponding to LEnKF and LIEnKF, respectively) and $TC6_v$ ((e) and (f) corresponding to LEnKF and LIEnKF, respectively).

Figure 5. Spatial patterns of the variance of Y (Var_Y) in $TC6_c$ at the end of the assimilation period for (a) LEnKF and (b) LIEnKF. Corresponding results for $TC6_v$ are depicted in (c) and (d).

Figure 6. Spatial distribution of the ensemble mean Y values obtained after the first assimilation step in (a, b) $TC7_i$ and (c, d) $TC7_{ii}$. The black circle represents the location of the filter of the pumping well.

Figure 7. Spatial distributions of the (ensemble) (a, c) average and (b, d) variance of Y for (a, b) TC8 and (c, d) TC9 obtained through covariance localization ensemble Kalman filter by solely assimilating (a, b) h_f or (c, d) S_a.

Figure 8. Spatial distributions of reference and (ensemble) averaged equivalent freshwater head values for (a, b, c) TC8 and (d, e, f) TC9. Reference contour lines corresponding to some selected head values are respectively depicted by the fine, medium heavy and heavy red curves, respectively; corresponding simulation results are depicted by green, cyan and blue curves. Results are depicted for early (5 days), intermediate (25 days) and late (50 days) assimilation times.

Figure 9. Spatial distributions of reference and (ensemble) averaged salinity values for (a, b, c) TC8 and (d, e, f) TC9. Results are depicted for early (5 days), intermediate (25 days) and late (50 days) assimilation times. Reference contour

lines corresponding to 10%, 50% and 90% of the employed seawater salinity are respectively depicted by the fine, medium heavy and heavy red curves, respectively; corresponding simulation results are depicted by green, cyan and blue curves.

Figure 10. Temporal variations of RMSE (a) and spread (b) associated with *Y* for TCs 8-12.

Figure 11. Temporal evolution of RMSE associated with (a, c) *Y* and (b, d) h_f for (a, b) LEnKF and (c, d) LIEnKF in $TC6_c$ with assimilation frequency corresponding to 1, 5, and 10 days (associated with the DIF setting).

Figure 12. Temporal evolution of RMSE associated with (a, b) *Y*, (c, d) h_f, and (e, f) S_a for (a, c, e) LEnKF and (b, d, f) LIEnKF in $TC6_v$, with assimilation frequency corresponding to 1, 5, and 10 days (associated with the VDF setting).

Figure 13. Temporal variation of the RMSE for *Y* for TCs 8 and 9 with differing ensemble size, *N*.

Figure 14. Temporal variation of RMSE (circle-dot curves) and spread (solid curves) for *Y* in TCs 13-16 (corresponding to values of horizontal correlation scale λ_x = 40 m, 80 m, 160 m, and 24 m, respectively).

Tables

Table 1. Model parameters considered as uniform in the simulations.

C_{sea}	salinity of seawater (g/L)	35
C_{in}	inflow concentration (g/L)	0
ρ_{sea}	density of seawater (kg/m^3)	1025
ρ_f	density of fresh water (kg/m^3)	1000
D	dispersion (m^2/day)	0.57024
θ	effective porosity	0.35
S	specific yield	0.35

Table 2. Test cases considered for the analysis of the effects of: (*i*) data quantity, data assimilation frequency and measurement error (Group A); (*ii*) limitations of domain localization methods (i.e., Group B, comprising the collection of $TC7_i$, $TC7_{ii}$, and $TC7_{iii}$); (*iii*) diverse observation types (Group C); and (*iv*) uncertainties linked to incomplete knowledge of the correlation scales of the randomly heterogeneous Y field (Group D). Symbols corresponds to: equivalent freshwater head (h_f, in units of m); salinity (S_a, in units of g/L); (natural) logarithm of equivalent freshwater hydraulic conductivity ($Y = \ln K_f$, in units of ln(m/day)); standard deviation of measurement error (σ_e, expressed in m); number of observations (N_{obs}); localization parameters α and β in equation (12); initial guesses of ensemble mean (μ) and variance (σ^2) of Y; guessed correlation scale of Y in the x-direction (λ_x, in units of m); subscripts c and v correspond to density-independent and density-dependent flow conditions, respectively.

Group A			Group B				Group C		Group D	
Test case	σ_e	N_{obs}	Test case	Observation	(α, β)	(μ, σ^2)	Test case	Observation	Test case	λ_x (m)
$TC1_c$ /$TC1_v$	0.1	30	$TC7_i$	S_a	4.0, 1.5	6.0, 1.69	TC8	h_f	TC13	40
$TC2_c$ /$TC2_v$	0.01	30	$TC7_{ii}$	S_a	32, 0.2	6.0, 1.69	TC9	S_a	TC14	80
$TC3_c$ /$TC3_v$	0.001	30	$TC7_{iii}$	S_a	4.0, 1.5	6.76, 1.0	TC10	h_f and Y	TC15	160
$TC4_c$ /$TC4_v$	0.1	50					TC11	S_a and Y	TC16	24
$TC5_c$ /$TC5_v$	0.01	50					TC12	h_f and S_a		
$TC6_c$ /$TC6_v$	0.001	50								

Table 3. Approximated computational complexities associated with LEnKF, LIEnKF and $\mathrm{LEnKF_{cov}}$ at a given assimilation cycle using a daily data assimilation frequency. Notations are as follows: C_H (independent from filter iteration), observation operator cost; CP_{12} (independent from filter iteration and the grid cells analyzed), cost for computing function (12) for each numerical grid cell; CP_{A1}, cost for computing localization function (A1) in Appendix A; CP_f, cost of performing the forward model once with time interval equal to one day under transient state; N_{iter} ($N_{iter} \geq 1$), required iteration number for LIEnKF; N_{obs}, number of observations; m_s^u, size of state vector $\mathbf{x}$ ($m_s^u = 2m$ or $3m$ for DIF or VDF scenarios, respectively).

	Forecast step	Analysis step
LEnKF	NCP_f	$N^2 m_s^u + C_H + m(N_{obs} m_s^u + CP_{12})$
LIEnKF	$N(N_{iter}+1)CP_f$	$N_{iter}[C_H + m(N^3 + CP_{12})]$
$\mathrm{LEnKF_{cov}}$	NCP_f	$N^2 m_s^u + C_H + (m_s^u)^2 + CP_{\mathrm{A1}}$

Table 4. Results of the tuning process for parameters α and β in (12) based on TCs 4-6 (including TCs 4c-6c and TCs 4v-6v) for the diverse values of the standard derivation of measurement error (σ_e) considered.

(α, β)	Standard deviation of measurement error, σ_e		
	0.1 m (TC6)	0.01 m (TC5)	0.001 m (TC4)
DIF	0.7, 1.5	1.2, 1.5	2, 1.5
VDF	1, 1.5	2, 1.5	4, 1.5

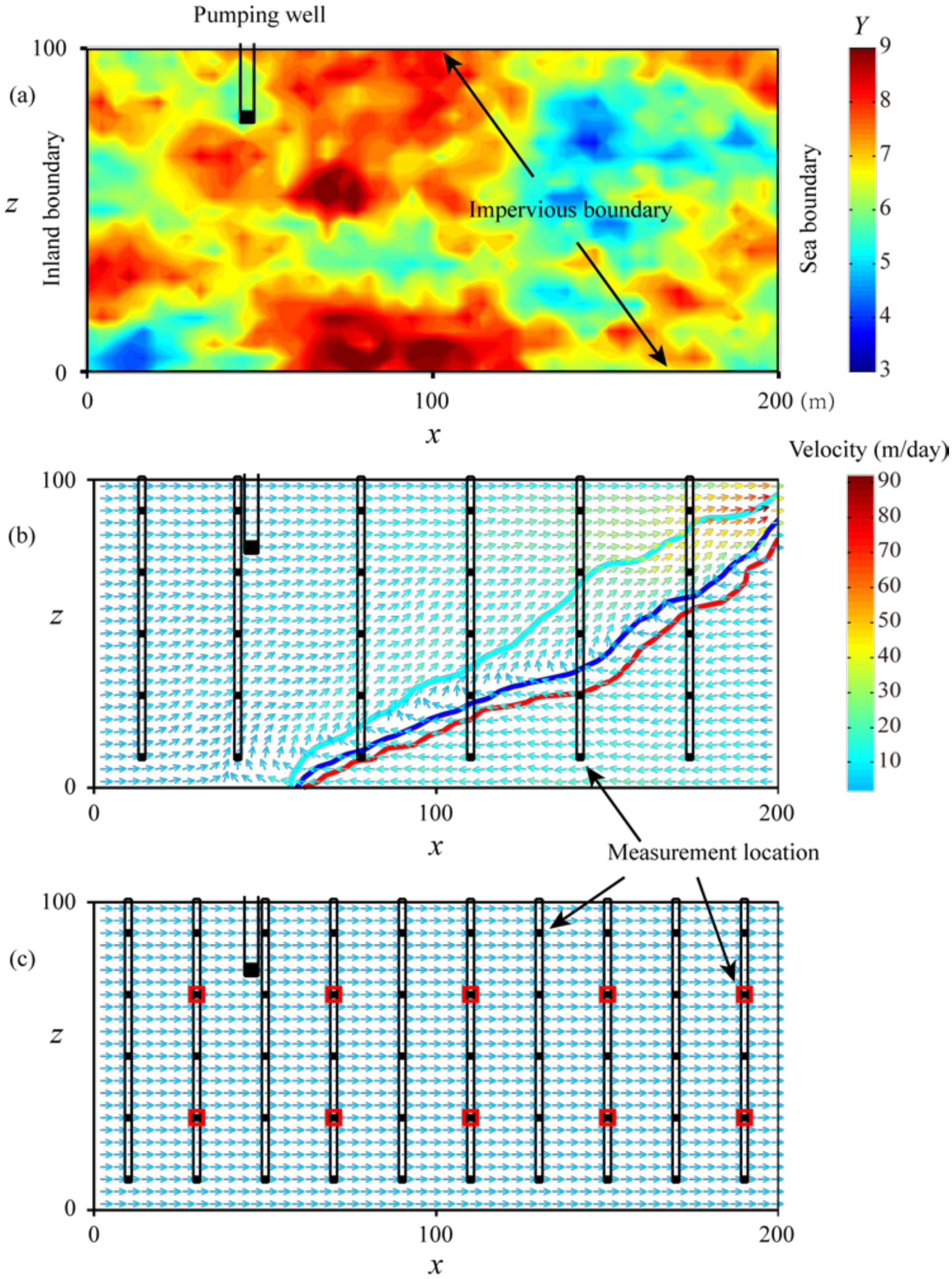

Figure 1. Schematic representation of the modified Henry problem considered in this study: (a) type of boundary conditions, location of the partially penetrating pumping well (the well screen being denoted as a black block), reference $Y = \ln K_f$ field; (b) initial flow field and concentration distribution (contour lines corresponding to initial salinity of 10%, 50% and 90% are respectively depicted as cyan, blue and red curves) for the variable-density flow (VDF) scenario. Initial spatial distribution of flux vectors (color scale ranges from blue to red, respectively denoting low to

high flux norm) for VDF scenario; (c) initial spatial distribution of flux vectors for density-independent flow (DIF) scenario. Location of the 30 observation points for pressure head are shown in (b), location of the 50 observation points for head (and salinity for VDF) being included in (c), black blocks along the boreholes corresponding to measurement locations. Red squares in (c) represent the 10 locations where *Y* data are collected.

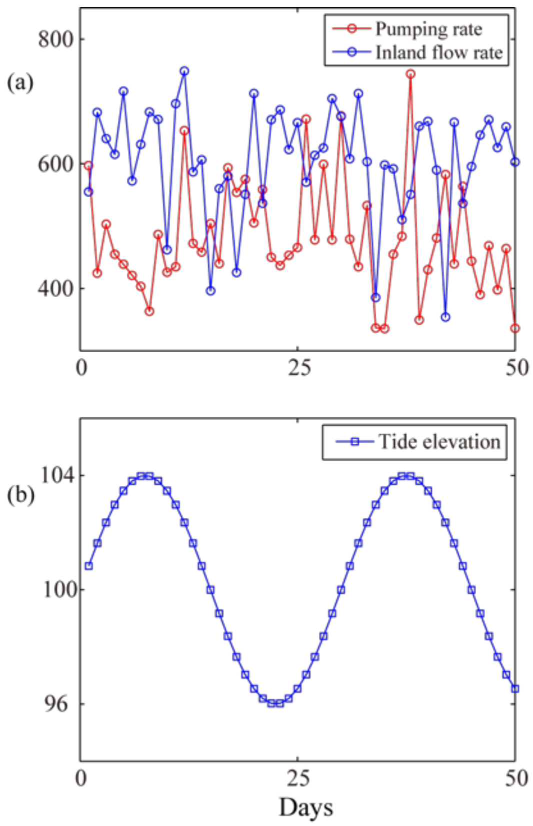

Figure 2. Values for pumping rate and boundary conditions selected as input to our computations: (a) pumping rates (blue curve) and inland boundary flow rates (here given in (m^3/day)); (b) tidal elevations (given in (m)).

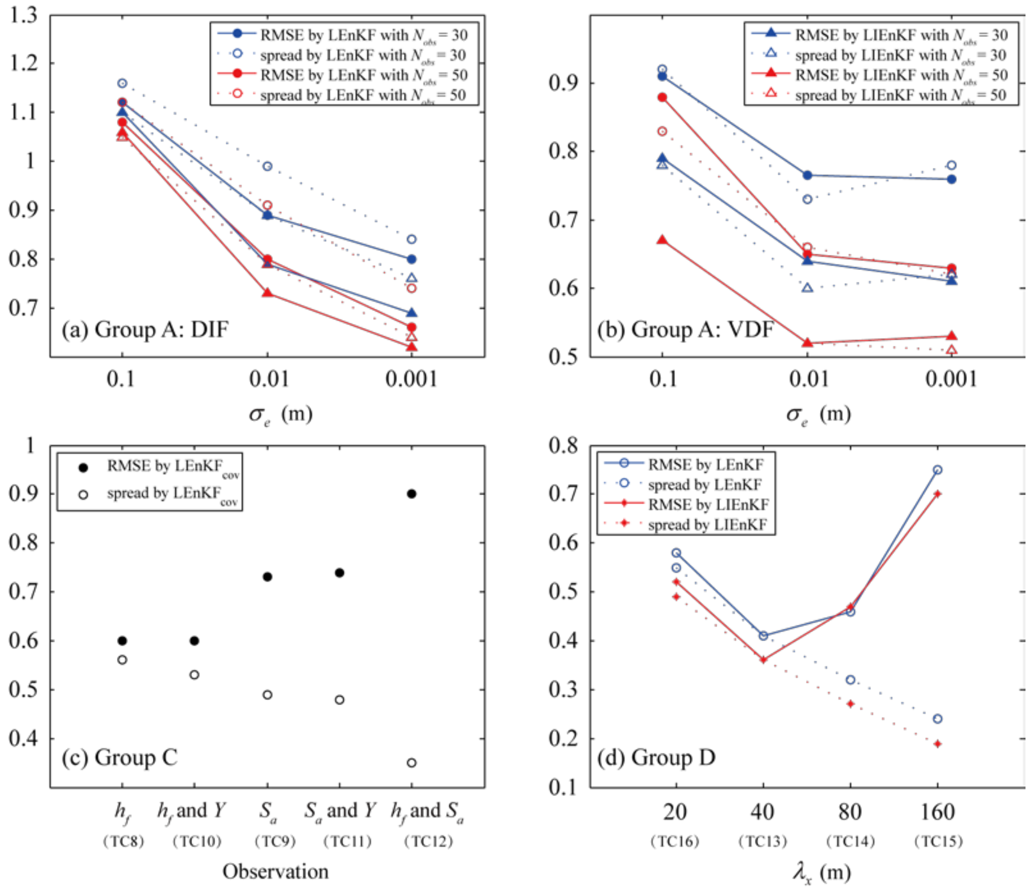

Figure 3. RMSE and spread values collected according to the showcase groups of Table 2 (excluding Group B): ((a) and (b) for density-independent (DIF) and density-dependent flow scenarios (VDF), respectively) Group A; (c) Group C; and (d) Group D.

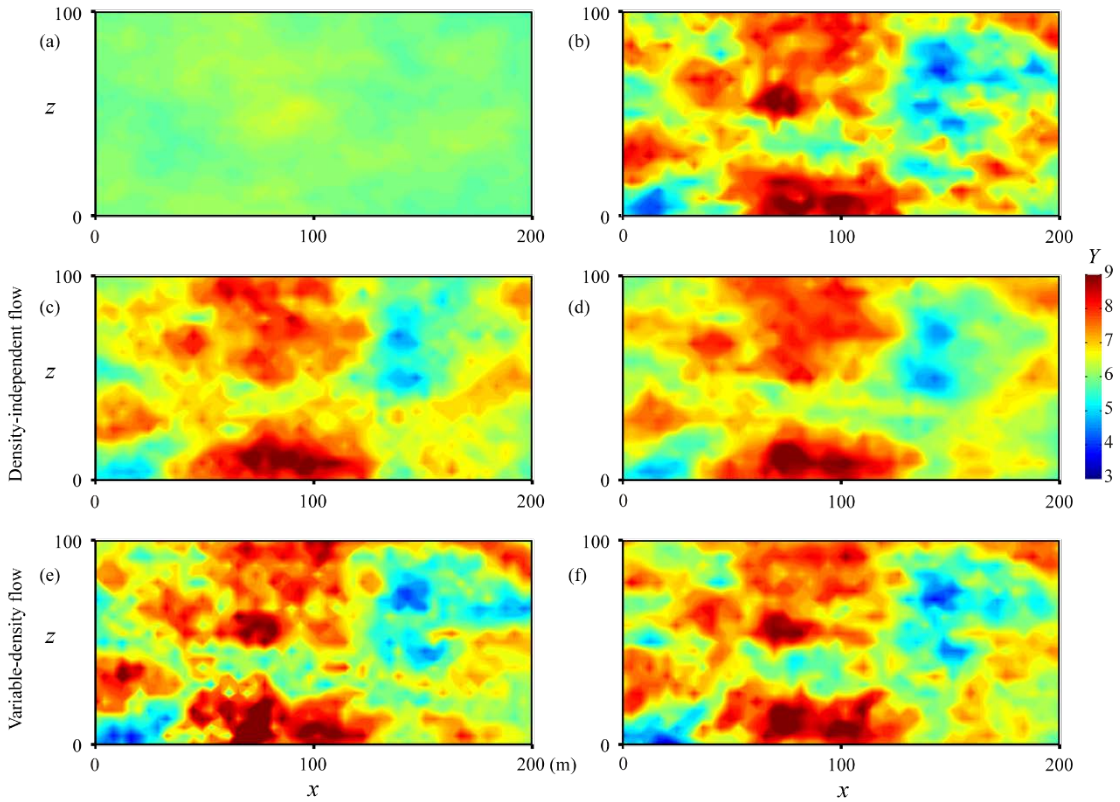

Figure 4. Initial (a) ensemble-averaged and (b) reference Y field; reference and ensemble-averaged Y fields obtained at the end of the data assimilation process for $TC6_c$ ((c) and (d) corresponding to LEnKF and LIEnKF, respectively) and $TC6_v$ ((e) and (f) corresponding to LEnKF and LIEnKF, respectively).

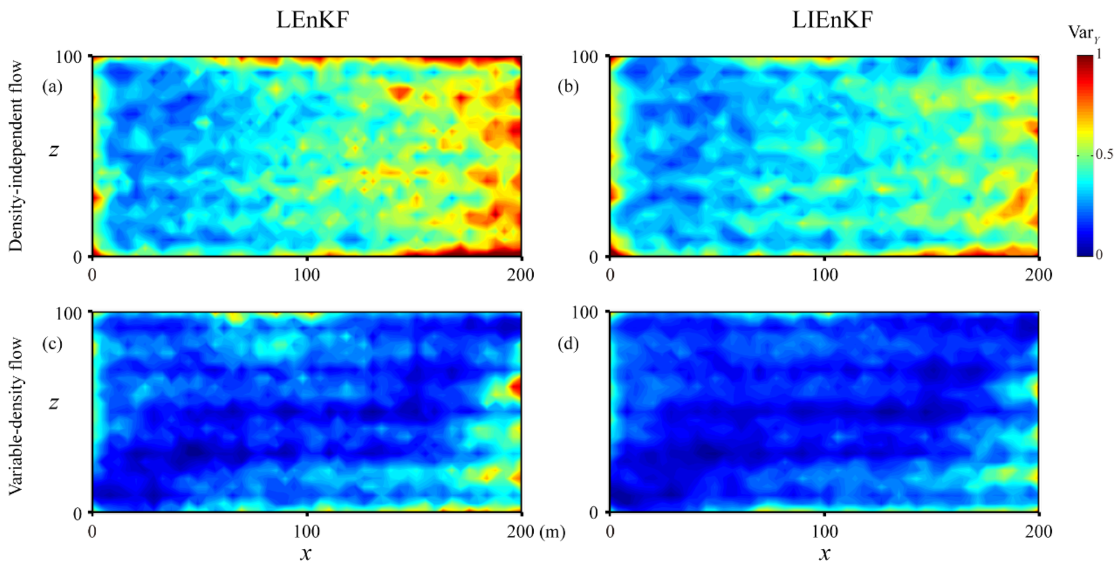

Figure 5. Spatial patterns of the variance of Y (Var_Y) in TC6$_c$ at the end of the assimilation period for (a) LEnKF and (b) LIEnKF. Corresponding results for TC6$_v$ are depicted in (c) and (d).

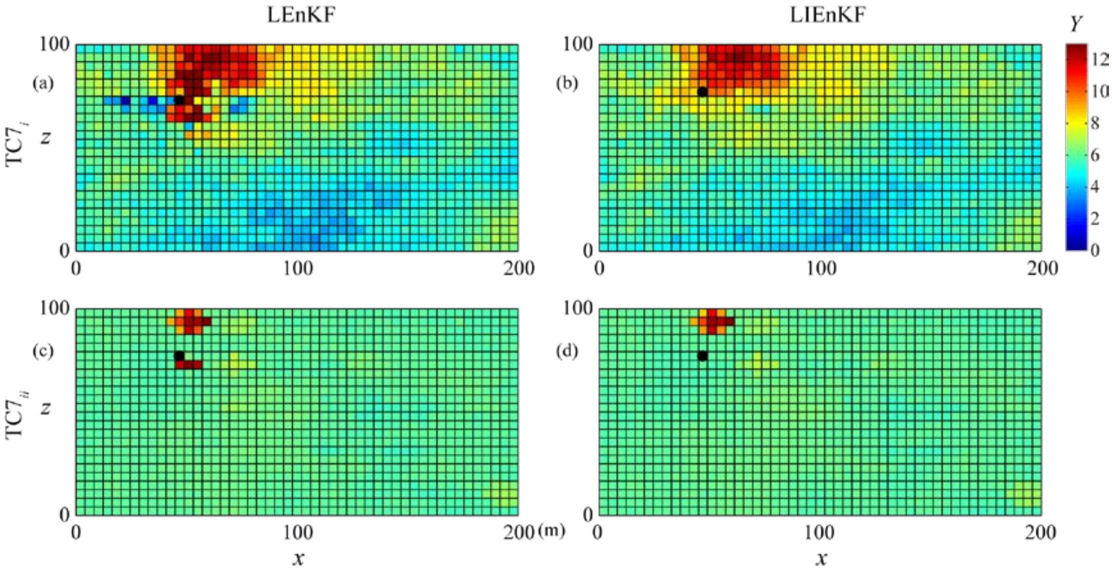

Figure 6. Spatial distribution of the ensemble mean Y values obtained after the first assimilation step in (a, b) $TC7_i$ and (c, d) $TC7_{ii}$. The black circle represents the location of the filter of the pumping well.

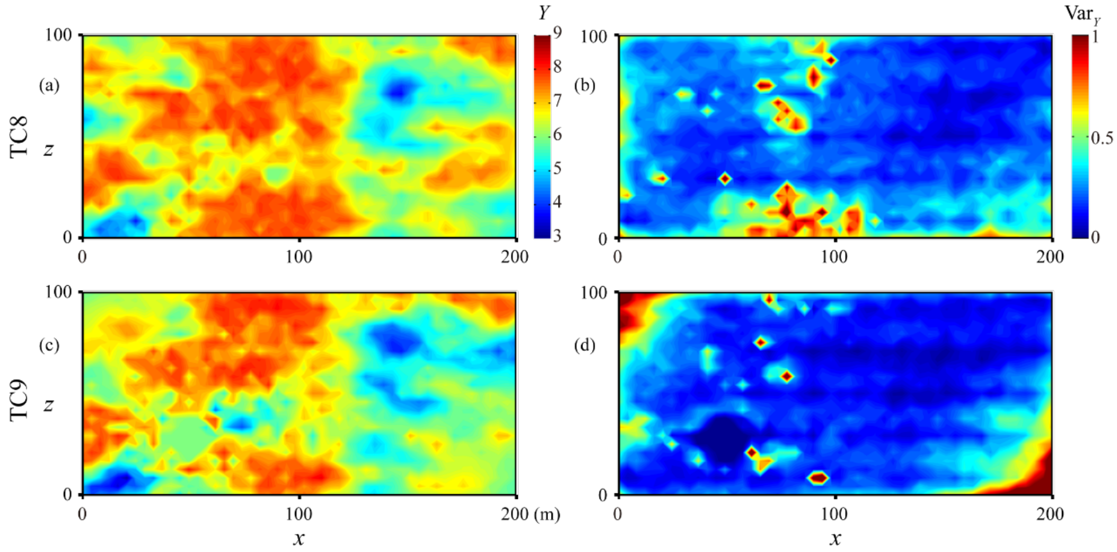

Figure 7. Spatial distributions of the (ensemble) (a, c) average and (b, d) variance of Y for (a, b) TC8 and (c, d) TC9 obtained through covariance localization ensemble Kalman filter by solely assimilating (a, b) h_f or (c, d) S_a.

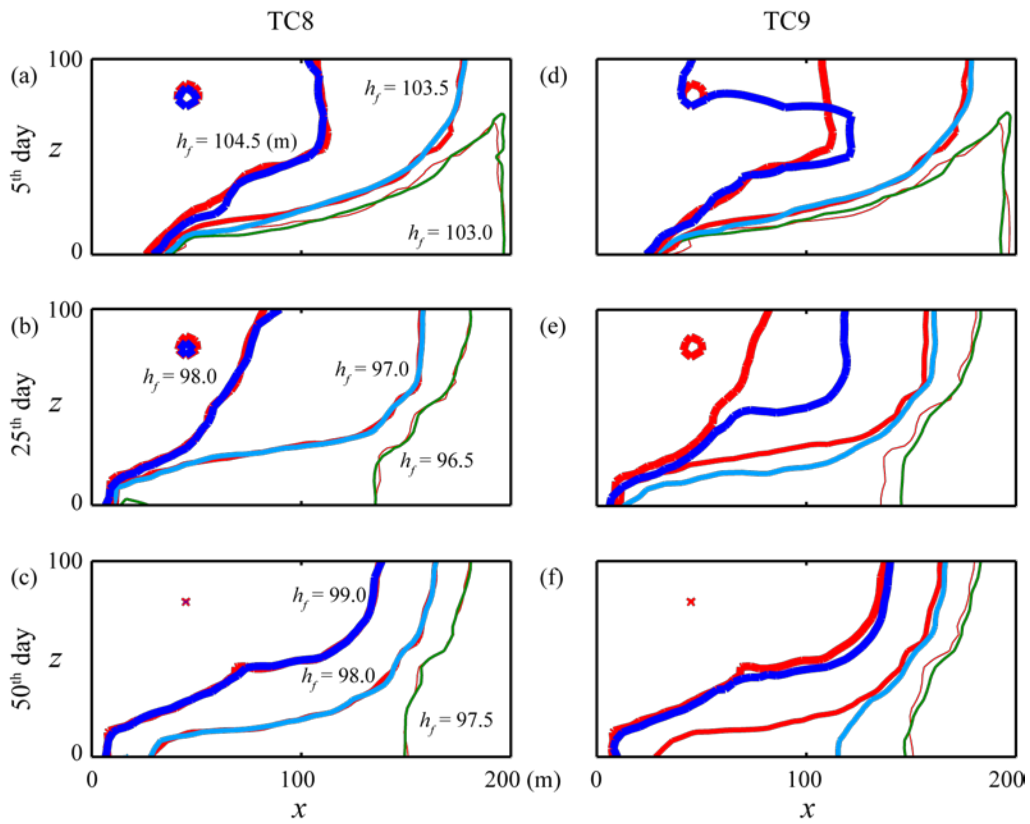

Figure 8. Spatial distributions of reference and (ensemble) averaged equivalent freshwater head values for (a, b, c) TC8 and (d, e, f) TC9. Reference contour lines corresponding to some selected head values are respectively depicted by the fine, medium heavy and heavy red curves, respectively; corresponding simulation results are depicted by green, cyan and blue curves. Results are depicted for early (5 days), intermediate (25 days) and late (50 days) assimilation times.

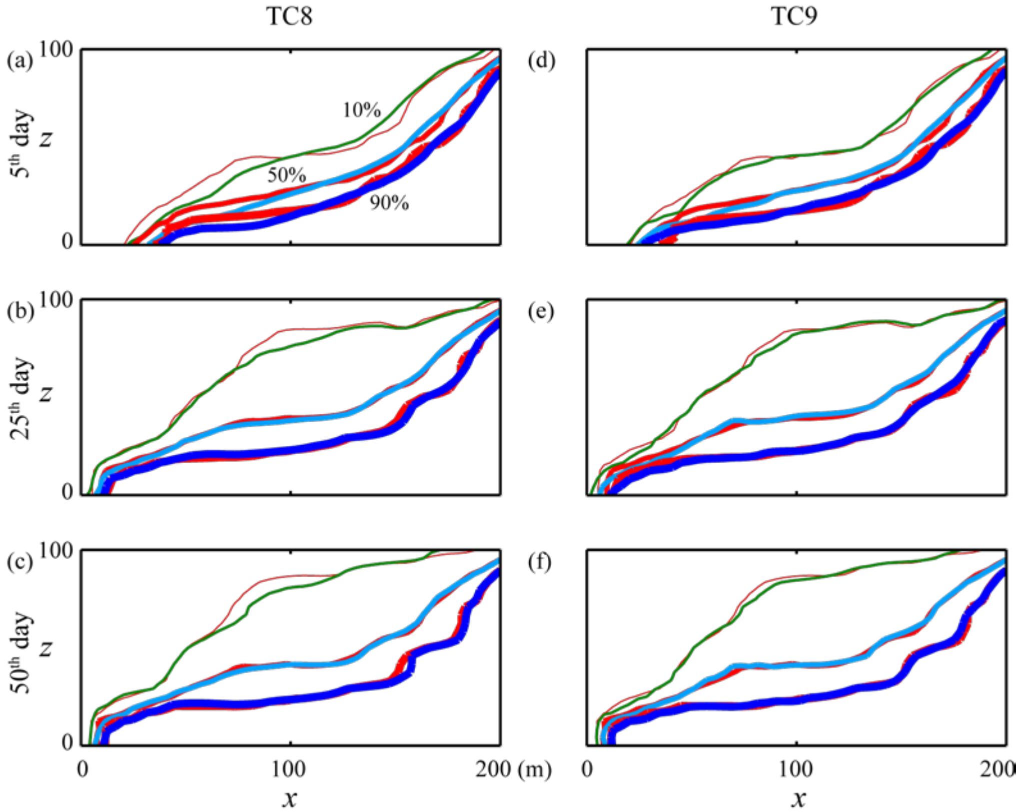

Figure 9. Spatial distributions of reference and (ensemble) averaged salinity values for (a, b, c) TC8 and (d, e, f) TC9. Results are depicted for early (5 days), intermediate (25 days) and late (50 days) assimilation times. Reference contour lines corresponding to 10%, 50% and 90% of the employed seawater salinity are respectively depicted by the fine, medium heavy and heavy red curves, respectively; corresponding simulation results are depicted by green, cyan and blue curves.

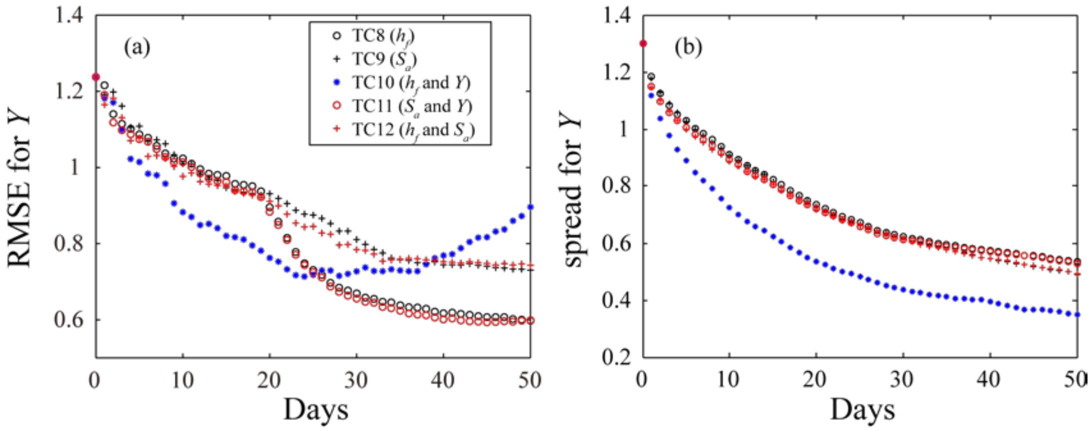

Figure 10. Temporal variations of RMSE (a) and spread (b) associated with Y for TCs 8-12.

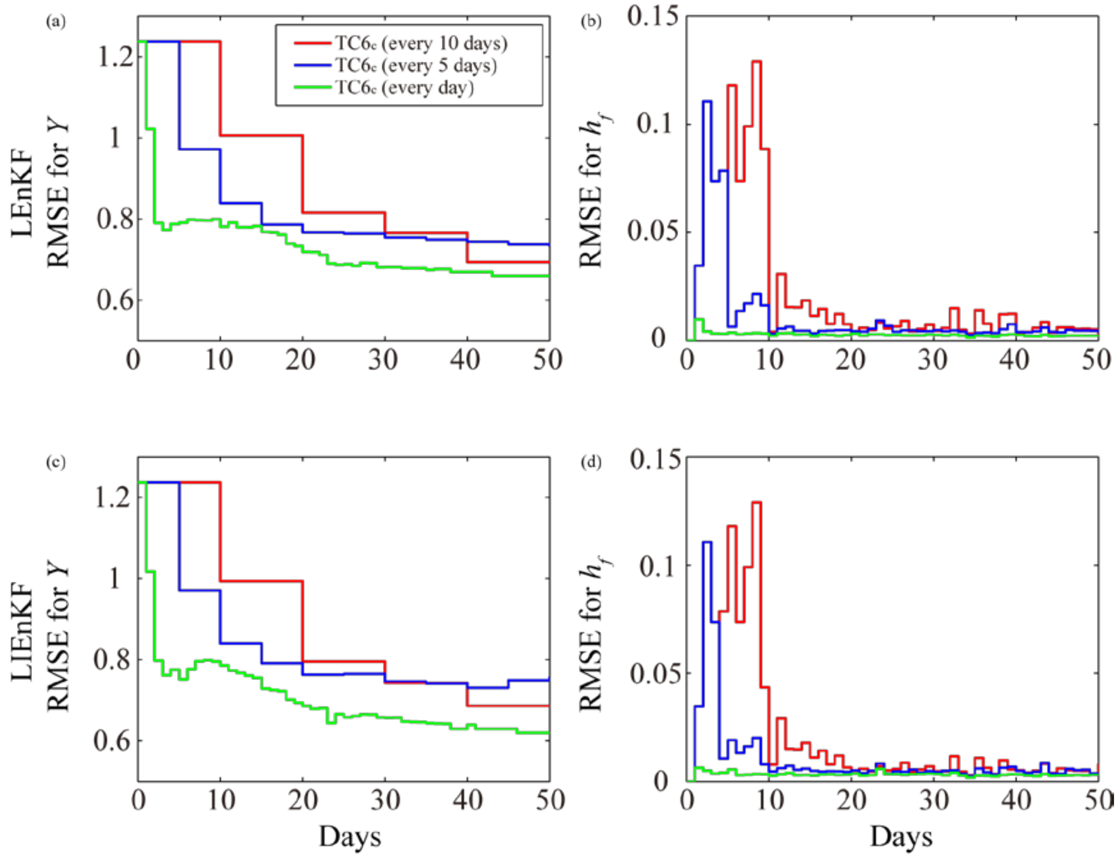

Figure 11. Temporal evolution of RMSE associated with (a, c) Y and (b, d) h_f for (a, b) LEnKF and (c, d) LIEnKF in TC6$_c$ with assimilation frequency corresponding to 1, 5, and 10 days (associated with the DIF setting).

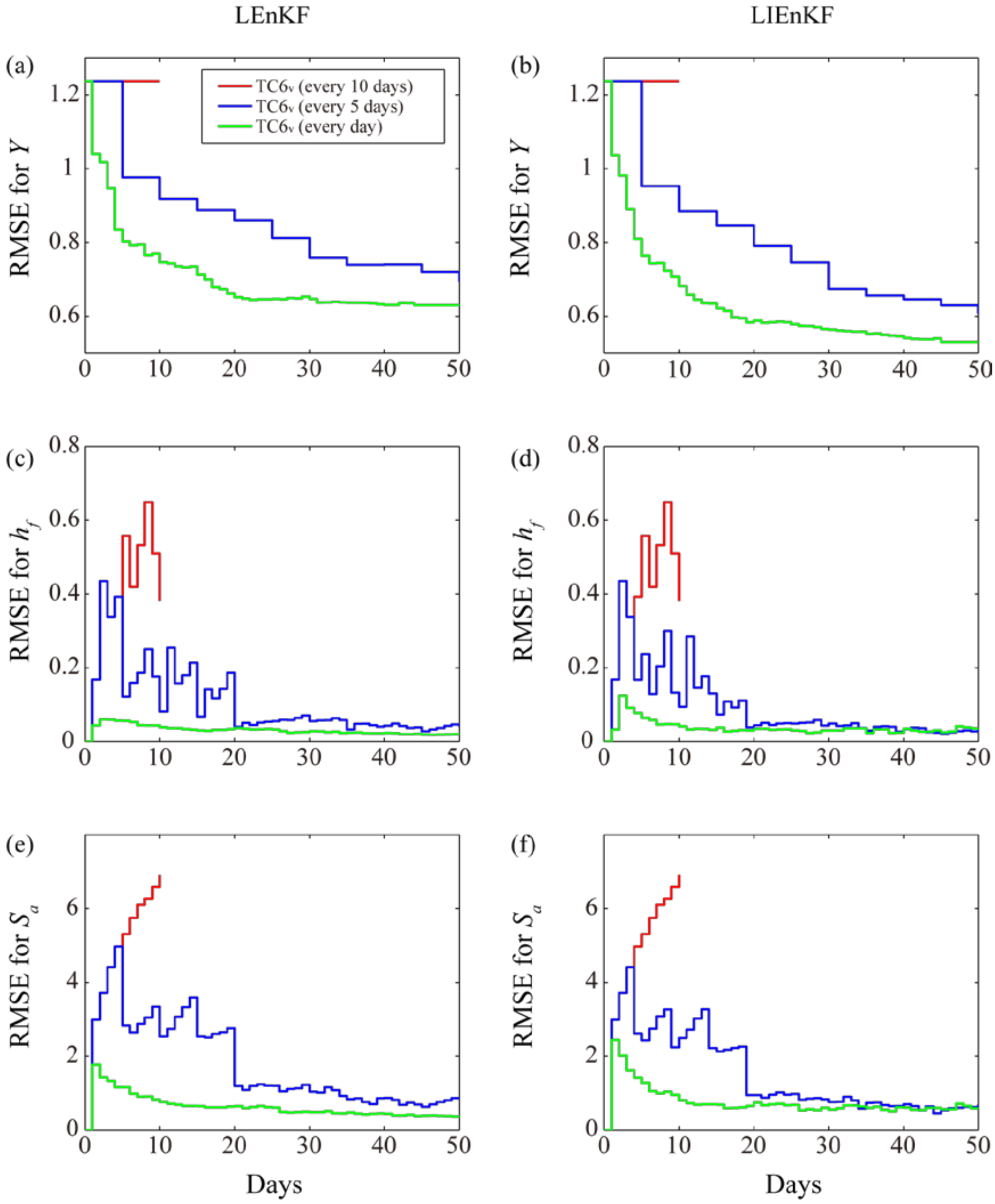

Figure 12. Temporal evolution of RMSE associated with (a, b) *Y*, (c, d) h_f, and (e, f) S_a for (a, c, e) LEnKF and (b, d, f) LIEnKF in TC6$_v$, with assimilation frequency corresponding to 1, 5, and 10 days (associated with the VDF setting).

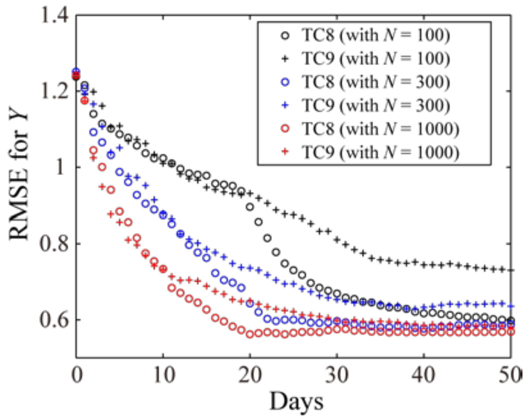

Figure 13. Temporal variation of the RMSE for *Y* for TCs 8 and 9 with differing ensemble size, *N*.

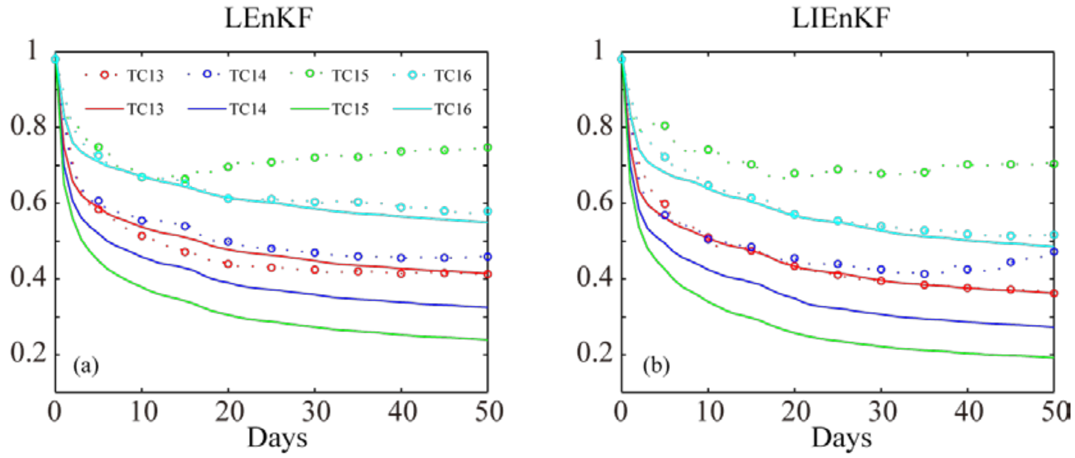

Figure 14. Temporal variation of RMSE (circle-dot curves) and spread (solid curves) for *Y* in TCs 13-16 (corresponding to values of horizontal correlation scale λ_x = 40 m, 80 m, 160 m, and 24 m, respectively).