Data assimilation for unsaturated flow models with restart adaptive probabilistic collocation based Kalman filter

Jun Man¹, Weixuan Li², Lingzao Zeng^{1*}, Laosheng Wu³

¹ Institute of Soil and Water Resources and Environmental Science, College of Environmental and Resource Sciences, Zhejiang Provincial Key Laboratory of Agricultural Resources and Environment, Zhejiang University, Hangzhou 310058, China

² Pacific Northwest National Laboratory, Richland, WA 99352, USA

³ Department of Environmental Sciences, University of California, Riverside, CA 92521, USA

*Corresponding author: L. Zeng, Zhejiang University, Hangzhou, China 310058.
(lingzao@zju.edu.cn)

Abstract. The ensemble Kalman filter (EnKF) has gained popularity in hydrological data assimilation problems. As a Monte Carlo based method, a sufficiently large ensemble size is usually required to guarantee the accuracy. As an alternative approach, the probabilistic collocation based Kalman filter (PCKF) employs the polynomial chaos expansion (PCE) to represent and propagate the uncertainties in parameters and states. However, PCKF suffers from the so-called "curse of dimensionality". Its computational cost increases drastically with the increasing number of parameters and system nonlinearity. Furthermore, PCKF may fail to provide accurate estimations due to the joint updating scheme for strongly nonlinear models. Motivated by recent developments in uncertainty quantification and EnKF, we propose a restart adaptive probabilistic collocation based Kalman filter (RAPCKF) for data assimilation in unsaturated flow problems. During the implementation of RAPCKF, the important parameters are identified and active PCE basis functions are adaptively selected at each assimilation step; the "restart" scheme is utilized to eliminate the inconsistency between updated model parameters and states variables. The performance of RAPCKF is systematically tested with numerical cases of unsaturated flow models. It is shown that the adaptive approach and restart scheme can significantly improve the performance of PCKF. Moreover, RAPCKF has been demonstrated to be more efficient than EnKF with the same computational cost.

1. Introduction

For accurate prediction of flow and solute transport in vadose zone, it is important to identify model parameters in governing equations. Recent researches have focused on estimating these parameters from in-situ measurements by solving an inverse problem [1-3]. For soil hydraulic property (SHP) parameters used in the water retention curve and hydraulic conductivity function, it is possible to estimate them from indirect measurements of the hydraulic states, e.g., pressure head and water content. The ensemble Kalman filter (EnKF) is an increasingly popular data assimilation method, which has been widely used in hydrology [4-6]. For unsaturated flow problems, the SHP parameters of 12 different soil types were estimated by EnKF through one-dimensional homogeneous soil column, in which measurements of pressure head were assimilated [7]. In EnKF, there is no need to calculate the derivatives required in the gradient-based approaches. Instead, this sensitivity information is provided by the cross-correlations between the model parameters and state variables calculated from an ensemble of model realizations. The model response can be simulated using an existing simulator for the model being studied. Once some measurement data become available, the ensemble realizations are updated accordingly using the Kalman gain.

In order to maintain sufficient spread in the ensemble and match the analysis covariance matrix, the observations used in the standard EnKF are considered as random variables and perturbed according to the distributions of observation errors [8]. To eliminate the sampling errors induced by observation perturbations, different forms of deterministic ensemble Kalman filters have been proposed [9-11]. One commonly used non-perturbed method is ensemble square root filter (EnSRF), where a modified analysis scheme is used and there are no perturbations added to the observations [12, 13]. It is of interest to note that, sampling errors introduced by perturbing the observations can still be avoided within the context of the stochastic EnKF using second-order exact sampling algorithms [14]. However, in all the ensemble based Kalman filters, no matter which analysis scheme is used, the

sampling errors derived from generating the initial ensemble realizations of unknown model parameters always exist. As a Monte Carlo based method, the accuracy of EnKF relies on a sufficiently large ensemble size. On the other hand, it could be very time consuming to run these simulations, especially for large scale problems. In practical applications, one has to make a compromise between the accuracy and affordable computational cost.

To mitigate the sampling errors, one possible approach is to construct a surrogate for the flow transport. Surrogate based methods have been more and more popular in uncertainty quantification for subsurface flow and solute transport [15-18]. A probabilistic collocation based Kalman filter (PCKF), which combines Kalman filter with polynomial chaos expansion (PCE), was developed for parameter estimation in [19]. In PCKF, the uncertain model parameters and states variables are represented with PCE. The mean and covariance matrix needed for data assimilation in KF can be calculated from the PCE coefficients, which are determined by the probabilistic collocation method (PCM) [20, 21]. The computational cost of PCKF is mainly spent on solving PCE coefficients, in which any existing simulator can be directly used. Previous studies have shown that, under certain conditions, PCKF can be more computationally efficient than EnKF [22, 23]. However, the number of terms needed in a PCE grows very fast with the increasing number of parameters and system nonlinearity, which will result in an enormous computational burden. This is the so-called "curse of dimensionality". In that case, PCKF could be less efficient than EnKF.

Indeed, some researchers have concluded that, PCE is unsuitable for forward unsaturated flow problems due to the large computational cost [24]. Generally speaking, several hundreds of model evaluations are very challenging for classical PCKF in unsaturated flow problems. It is questionable whether PCKF can outperform EnKF at the same computational cost. To address this issue, an adaptive ANOVA-based PCKF (APCKF) has been proposed in [25], where the important parameters are identified based on analysis of variance (ANOVA). Next, the PCE is built in an adaptive manner in which the terms with respect to important parameters

are kept while those associated with insignificant parameters are largely removed. This approach was shown capable of greatly reducing the number of terms needed in a PCE, and thus improving the efficiency of PCKF for high-dimensional and strongly nonlinear problems.

Another issue of PCKF in a strongly nonlinear problem is the commonly used joint updating approach, in which an augmented state vector, consisting of dynamic states (pressure heads or water contents) and static parameters (SHP parameters), is firstly defined and jointly updated at the analysis step. Although this approach provides decent computational efficiency, it may fail in strongly nonlinear problems and cause the inconsistency between updated model parameters and state variables. For example, the updated SHP parameters and states may not follow the nonlinear Richards equation after the analysis step; the state values may even be physically unreasonable, e.g., water saturation being >1 or <0 . Furthermore, augmenting the model parameters and state variables in a single vector could increase the degrees of freedom in the system, which would make the estimation unstable and intractable. For more detailed issues in this joint updating approach, one may refer to [6, 26].

Gu and Oliver have claimed that the updated model parameters could not be consistent with the updated state variables for a nonlinear problem unless re-solving the nonlinear forward model [27]. To improve the performance of EnKF in strongly nonlinear problems, different iterative forms of EnKF have been proposed. In the “restart” EnKF, only the parameters are updated in the analysis step, the states are obtained by rerunning the nonlinear model from time zero to the current time step with the updated model parameters [28]. The efficiency of restart EnKF has been demonstrated in a real-time groundwater modeling problem [29]. Moreover, a more computationally efficient restart scheme, i.e., modified restart EnKF, was developed to improve the performance of EnKF in unsaturated flow problems [30]. In the “confirming” EnKF, the updated states are obtained by rerunning the simulator with updated parameters from last time step instead [31]. However, there are some debates about the confirming EnKF. When the confirming starts, the inconsistency between updated model parameters and state variables at current step is indeed eliminated, but

the inconsistency at the previous step is inevitable [31]. An iterative ensemble Kalman filter with one-step-ahead smoothing has been recently proposed for state-parameters estimation of contaminant transport models [32]. For the unsaturated flow problem, it has been shown that the scheme of restart from time zero performs better than the confirming scheme, however, at a much larger computational cost [30].

It should be noted that the transformed EnKF and dual EnKF have also been proposed to improve the applicability of EnKF in strongly nonlinear problems. In the transformed EnKF [33, 34], the state variables are represented by alternative variables so that the relations between state variables and observations are more linear, then the state variables can be nearly linearly updated. In the dual EnKF [6, 26], the states and parameters are separately estimated using two parallel filters in a dual manner. The superiority of dual EnKF to joint-updating EnKF has been demonstrated in a subsurface flow-transport coupled model [6].

Motivated by recent advances in EnKF and PCE, we integrate the restart scheme with APCKF for data assimilation of unsaturated flow. Its performance is systematically tested with numerical case studies and compared to the same form of EnKF with the similar computational cost. The rest of the paper is organized as follow: Section 2 reviews the mathematical model for unsaturated flow, whereas Section 3 presents the data assimilation approach restart adaptive PCKF (RAPCKF). Two illustrative numerical examples are shown in Section 4 and then the paper is summarized in Section 5.

2. System Model

In this study, water flowing through unsaturated soil is considered. The governing Richards equation is as follows,

$$\frac{\partial \theta(h)}{\partial t} = \frac{\partial}{\partial x_i} \left[K(h) \frac{\partial h}{\partial x_i} - K(h) \right] \quad (1)$$

where x_i is the distances in all directions [L]; t is the time [T]; h is the pressure head [L]. $\theta(h)$ and $K(h)$ are the volumetric water content [L^3L^{-3}] and hydraulic

conductivity function $[LT^{-1}]$ varied with h , respectively, which are characterized by van Genuchten-Mualem (VGM) model [35, 36],

$$S_e = \begin{cases} \frac{1}{(1 + |\alpha h|^n)^{1-1/n}}, & h < 0 \\ 1, & h \geq 0 \end{cases} \quad (2)$$

$$K(h) = K_s S_e^{0.5} [1 - (1 - S_e^{1/m})^m]^2 \quad (3)$$

where $S_e = \frac{\theta - \theta_r}{\theta_s - \theta_r}$ is the effective saturation; θ_r and θ_s are the residual and saturated water content $[L^3L^{-3}]$, respectively; K_s is the saturated hydraulic conductivity $[LT^{-1}]$, $m = 1 - 1/n$; α $[L^{-1}]$ and n $[-]$ are the empirical parameters which are related to the soil pore-size distribution.

3. Methods

3.1 Ensemble Kalman Filter

Ensemble Kalman filter (EnKF) was proposed as a Monte Carlo variant of the standard Kalman filter (KF) in [37, 38]. Different from the KF, the uncertainty of estimated parameters is represented by the ensemble members. Suppose the response of a system can be simulated by

$$\mathbf{d} = G(\mathbf{x}) \quad (4)$$

where $G(\cdot)$ represents the system model, $\mathbf{d}$ denotes the model output vector, $\mathbf{x}$ denotes the augmented state vector, consisting of model parameters $\mathbf{m}$ and state variables $\mathbf{s}$

$$\mathbf{x} = [\mathbf{m} \ \mathbf{s}]^T \quad (5)$$

The ensemble of N_e members can be written as

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{N_e}] \quad (6)$$

To reduce the uncertainty represented by the ensemble members, the available

measurements $\mathbf{d}_{true}$ are used to update the prior ensemble. In this study, a square root form of EnKF is used [13], which is similar to standard EnKF on all the aspects except for the analysis scheme. The prior ensemble mean $\bar{\mathbf{x}}^f$ and ensemble perturbations $\mathbf{x}_i'^f, i=1,2,...,N_e$ are separately updated with Eqs. (7-8),

$$\bar{\mathbf{x}}^a = \bar{\mathbf{x}}^f + \mathbf{K}(\mathbf{d}_{true} - \mathbf{H}\bar{\mathbf{x}}^f) \quad (7)$$

$$\mathbf{x}_i'^a = \mathbf{x}_i'^f - \mathbf{K}'\mathbf{H}\mathbf{x}_i'^f, \quad i=1,2,...,N_e \quad (8)$$

where

$$\mathbf{K} = \mathbf{P}\mathbf{H}^T(\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R})^{-1} \quad (9)$$

and

$$\mathbf{K}' = \mathbf{P}\mathbf{H}^T[(\sqrt{\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R}})^{-1}]^T[\sqrt{\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R}} + \sqrt{\mathbf{R}}]^{-1} \quad (10)$$

where $\mathbf{P} = \frac{1}{N_e - 1} \sum_{i=1}^{N_e} (\mathbf{x}_i^f - \bar{\mathbf{x}}^f)(\mathbf{x}_i^f - \bar{\mathbf{x}}^f)^T$ is the prior covariance matrix, $\mathbf{H}$ is the

measurement operator relating the true state $\mathbf{x}_{true}$ to the actual measurements $\mathbf{d}_{true}$

$$\mathbf{d}_{true} = \mathbf{H}\mathbf{x}_{true} + \boldsymbol{\varepsilon} \quad (11)$$

where $\boldsymbol{\varepsilon}$ are measurement errors with covariance $\mathbf{R} = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T)$. Thus the posterior

ensemble $\mathbf{x}_i^a, i=1,2,...,N_e$ can be expressed with the updated ensemble mean $\bar{\mathbf{x}}^a$

and ensemble perturbations $\mathbf{x}_i'^a, i=1,2,...,N_e$,

$$\mathbf{x}_i^a = \bar{\mathbf{x}}^a + \mathbf{x}_i'^a, \quad i=1,2,...,N_e \quad (12)$$

In the following discussions, the nomination "EnKF" is used.

3.2 Restart Adaptive Probabilistic Collocation based Kalman Filter

When EnKF is used in the joint estimation of model parameters and states, both the static parameters and dynamic states are updated simultaneously without the need to rerun the model from initial time. This makes EnKF very computationally efficient. However, it may result in many issues as mentioned in the introduction section.

Considering the weakness of EnKF in dealing with unsaturated flow problems, a more efficient approach, i.e., RAPCKF, which integrates restart scheme with adaptive probabilistic collocation based Kalman filter proposed in [25], is presented in this study. In RAPCKF, the model parameters and outputs are approximated by polynomial chaos expansion (PCE) [22].

$$\mathbf{m}(\xi) \approx \sum_{j=0}^Q \mathbf{c}_j \varphi_j(\xi) \quad (13)$$

$$\mathbf{d}(\xi) \approx \sum_{j=0}^Q \mathbf{c}_j^d \varphi_j(\xi) \quad (14)$$

where $\xi = [\xi_1, \xi_2, \dots, \xi_n]^T$ is a vector consisting of n standard Gaussian random variables, $\varphi(\xi)$ is the Hermite polynomial basis function, which satisfies the orthogonal property: $E[\varphi_j(\xi)\varphi_k(\xi)] = \delta_{jk}$, where δ_{jk} is the Kronecker delta. The first term $\varphi_0(\xi)$ denotes the mean part, while the subsequent terms $\varphi_j(\xi), j=1, 2, \dots, Q$ denote the high order terms. The PCE representation of model parameters in Eq. (13) is set based on the prior distribution of $\mathbf{m}$, while $\mathbf{c}_j^d, j=0, 1, 2, \dots, Q$ are the PCE coefficient vectors for model outputs, which are determined by the probabilistic collocation method (PCM) [20, 21].

In standard PCKF, the choice of PCE order is pre-set according to users' experience. After the truncated PCE order is chosen, all the PCE terms for model outputs are kept in the implementation, which will bring enormous computational burden to solve the PCE coefficients, especially when the number of parameters is large.

It is intuitively understandable that, only some of the PCE terms contribute significantly in the solution. To balance the accuracy and computational cost, we employ the adaptive scheme based on analysis of variance (ANOVA) to identify the important PCE terms. The uncertainties of model outputs are represented by low-order components based on ANOVA. The ANOVA decomposition of the model output $\mathbf{d}$ with respect to the n standard Gaussian random variables can be

expressed by

$$f(\xi) = f_0 + \sum_{j=1}^n f_j(\xi) + \sum_{1 \leq j < k \leq n} f_{jk}(\xi_j, \xi_k) + \dots + f_{1,2,\dots,n}(\xi_1, \xi_2, \dots, \xi_n) \quad (15)$$

where the function $f_{1,2,\dots,r}$ with r variables $\xi_1, \xi_2, \dots, \xi_r$ is entitled " r th-order ANOVA component". In particular, the 1st-order ANOVA component f_1 with respect to a standard Gaussian random variable is expressed as follows,

$$f_1(\xi_1) \approx c_0 + c_1 \xi_1 + c_2 (\xi_1^2 - 1) / \sqrt{2} \quad (16)$$

and the 2nd-order ANOVA component $f_{1,2}$ with respect to two standard Gaussian random variables is expressed as follows,

$$f_{1,2}(\xi_1, \xi_2) \approx c_{00} + c_{01} \xi_2 + c_{02} (\xi_2^2 - 1) / \sqrt{2} + c_{10} \xi_1 + c_{11} \xi_1 \xi_2 + c_{12} \xi_1 (\xi_2^2 - 1) / \sqrt{2} + c_{20} (\xi_1^2 - 1) / \sqrt{2} + c_{21} \xi_2 (\xi_1^2 - 1) / \sqrt{2} + c_{22} (\xi_1^2 - 1) (\xi_2^2 - 1) / \sqrt{2} \quad (17)$$

Then PCM is used to determine the PCE coefficients of all the 1st-order ANOVA components. After determining the PCE coefficients in ANOVA components, their variances can be easily calculated with these coefficients. It is assumed that the variance distributed on the 2nd-order ANOVA component $\sigma^2(f_{i,j})$ can be estimated from the variances of 1st-order ANOVA components $\sigma^2(f_i)$ and $\sigma^2(f_j)$. More detailed discussions can be referred to [25]. In this study, all the 1st-order ANOVA components are retained and 2nd-order ANOVA components are adaptively selected according to their variances. After that, PCM is again used to determine the PCE coefficients of these adaptively selected 2nd-order ANOVA components. Thus the statistical moments needed for implementing the Kalman filter can be expressed with PCE coefficients.

$$\mathbf{PH}^T = E \left\{ \left[\sum_{j=1}^Q \mathbf{c}_j \varphi_j(\xi) \right] \left[\sum_{k=1}^Q \mathbf{c}_k^d \varphi_k(\xi) \right]^T \right\} = \sum_{j=1}^Q \mathbf{c}_j \mathbf{c}_j^{dT} \quad (18)$$

$$\mathbf{HPH}^T = E \left\{ \left[\sum_{j=1}^Q \mathbf{c}_j^d \varphi_j(\xi) \right] \left[\sum_{k=1}^Q \mathbf{c}_k^d \varphi_k(\xi) \right]^T \right\} = \sum_{j=1}^Q \mathbf{c}_j^d \mathbf{c}_j^{dT} \quad (19)$$

Owing to the use of restart scheme, only the PCE coefficients of model parameters

are needed to be updated in RAPCKF. The mean part $\mathbf{c}_0^f$ and perturbation parts $\mathbf{c}_j^f, j=1,2,...,Q$ of model parameters are updated according to Eq. (7) and Eq. (8), respectively,

$$\mathbf{c}_0^a = \mathbf{c}_0^f + \mathbf{K}(\mathbf{d}' - \mathbf{c}_0^d) \quad (20)$$

$$\mathbf{c}_j^a = \mathbf{c}_j^f + \mathbf{K}'(-\mathbf{c}_j^d), \quad j=1,2,...,Q \quad (21)$$

where $\mathbf{K}$ and $\mathbf{K}'$ can be calculated by substituting Eqs. (18-19) into Eq. (9) and Eq. (10), respectively.

After the analysis step, the predicted model outputs at next time step are obtained by running the simulator from initial time with the updated model parameters. There are two major merits by doing this. First, the inconsistency between updated model parameters and state variables is completely eliminated. Secondly, only the PCE terms of states at measurement locations instead of those on all nodes are needed to be calculated. Compared with APCKF, the CPU time for a single model evaluation is increased. However, the total number of model evaluations can be reduced, which will somewhat alleviate the computational burden by the use of the restart scheme. This will be illustrated in the following numerical case studies.

4. Case Studies and Discussions

4.1. Case 1: Water flowing through one-dimensional column

In this case, the one-dimensional movement of water in unsaturated soil is considered. To show the accuracy of estimation and predicting ability of soil hydraulic parameters, twin-numerical experiments are conducted, respectively. The experimental setups for parameter estimation are illustrated as follows. As shown in Figure 1, the soil column consists of three uniform layers, which are formed with sand, loamy sand and sandy loam, respectively. The thickness of each soil layer is 30 cm. A 6-mm-thick porous ceramic plate is placed at the bottom of the column. The study domain is discretized with 101 nodes. Initially, the column is saturated and there is a

4.4-cm-depth surface ponding on the top of the column. The lower boundary condition is a constant pressure head equal to -80 cm. There are in total 48 measurements. The pressure heads at the middle of each soil layer are recorded every 2 hours. The total simulation time lasts for 32 hours. There are 12 parameters, i.e., three sets of K_s, θ_r, α, n corresponding to three soil layers, to be estimated. The three saturated water contents θ_s are assumed known as 0.43, 0.43 and 0.41 ($\text{cm}^3 \text{ cm}^{-3}$), respectively. And the hydraulic property of the porous ceramic plate is set as follows: $\theta_r = 0$ ($\text{cm}^3 \text{ cm}^{-3}$), $\theta_s = 1$ ($\text{cm}^3 \text{ cm}^{-3}$), $\alpha = 1e-20$ (cm^{-1}), $n = 1.001$ (dimensionless), $K_s = 0.003$ (cm h^{-1}).

Since KF is based on Gaussian assumption, the prior statistics of hydraulic parameters should be Gaussian. Carsel and Parrish have described the statistics of VGM model parameters and transformed them into Gaussian variables via Johnson system [39]. In the Johnson system [40, 41], there are three main transformation forms, including lognormal (LN), log ratio (SB) and hyperbolic (SU), which are shown as follows,

$$LN : Y = \ln(m) \quad (22)$$

$$SB : Y = \ln[(m-A)(B-m)] \quad (23)$$

$$SU : Y = \sinh^{-1}[w] = \ln[w + (1 + w^2)^{1/2}], w = (m-A)/(B-A) \quad (24)$$

where m denotes the parameter before transformation within the range $[A, B]$, Y denotes the transformed parameter with Gaussian distribution.

The prior statistics of soil hydraulic parameters are listed in Table 1 [39], where different types of Johnson system are applied to transform each soil hydraulic parameter. As shown in Table 2 [39], in most cases, correlations between soil hydraulic parameters are significant and can't be ignored. It has been shown that, the identifiability of soil hydraulic parameters can be improved by incorporating these prior statistics in a formal Bayesian framework [42]. In this work, the factored covariance matrix $\mathbf{L}^T$ shown in Table 2 is obtained by the Cholesky transformation

[43], i.e.,

$$\mathbf{C}_Y = \mathbf{L}\mathbf{L}^T \quad (25)$$

where $\mathbf{C}_Y$ is the prescribed covariance matrix among transformed parameters, $\mathbf{L}$ is a lower triangular matrix while superscript T denotes transpose. Then realizations of soil hydraulic parameters can be generated as follows [43],

$$\mathbf{Y} = \mathbf{u} + \mathbf{L}\boldsymbol{\eta} \quad (26)$$

where $\mathbf{u}$ is a vector consisting of the mean values of transformed parameters, $\boldsymbol{\eta}$ is a vector of independent standard Gaussian random variables.

In this study, one set of SHP parameter realization is randomly chosen as reference (true) values. Measurements of pressure head are generated by running the simulator Hydrus-1D with the reference parameters. The measurement errors are assumed to be normally distributed with standard deviation 1 cm. Then RAPCKF is applied to estimate the unknown parameters. The water retention curve and hydraulic conductivity function are calculated with the finally updated parameter ensemble, respectively. As shown in Figures (2-3), the solid red lines stand for the reference. The dashed blue lines stand for the mean estimations. Considering the uncertainty estimations, shadowed areas stand for 95% confidence interval obtained by sampling the final PCE representation of parameters. On the whole, a very good match between the estimated and reference value is observed for both the water retention curve and hydraulic conductivity function. It is noted that, larger uncertainty occurs when the pressure head or hydraulic conductivity drastically varies with the water content. More specifically, when the water content is relatively low, the uncertainty of estimated water retention curve is large. While for the estimated hydraulic conductivity function, larger uncertainty occurs at the higher water content. A possible reason for this phenomenon is the stronger nonlinearity of VGM model for the pressure head at lower water content, for hydraulic conductivity at higher water content. As shown in Figure 4 (a-c), in addition to the mean values, 95% confidence interval for predicted data (pressure heads on observed nodes at measurement time

steps) are plotted. To test the predicting ability of estimated parameters under new conditions, the porous ceramic plate on the bottom of the column is removed, and the lower boundary condition is changed to seepage condition in this case, then the simulator is run with the reference parameters and those sampled from the final PCE representation, respectively. The scatter plots for the predicted and true pressure heads are shown in Figure 4(d-f). It is shown that the predicted values match with reference values very well.

In each KF loop, RAPCKF can adaptively select PCE terms according to their variances. Once some criterion is satisfied, the selection is stopped. Therefore, the number of PCE terms is not fixed during the whole data assimilation process. It is well known that keeping more PCE terms helps to improve the estimation accuracy, however, at the cost of higher computational cost. In Figure 5, we plot the number of model evaluations (for solving the PCE coefficients) in each data assimilation loops during the implementation of RAPCKF. At the first few assimilation steps, the number of model evaluations is relatively small. It is because the experiment is conducted by simulating water flowing through a column from wetting to drying. At the beginning, the initial states are assumed known without uncertainties, soil column is saturated, the changes of pressure head are relatively small and fewer PCE basis functions are required to approximate the model. When the soil turns evidently unsaturated, there are some drastic changes for the pressure head. At this time, more PCE basis functions are required to capture the uncertainty of model output, which increases the number of model evaluations. Overall, the average number of model evaluations is 36 at each assimilation step.

To examine the superiority of RAPCKF, we also solve the same problem using the restart EnKF with different ensemble sizes, i.e., 25, 50, 100, 200, 300, 500 and 1000. The result provided by the restart EnKF with 1000 realizations is used as the benchmark for accuracy comparison. Specifically, we compare the mean and variance estimations reported by RAPCKF and restart EnKFs against the benchmark using the squared Hellinger distance, which is defined as follows [25],

$$H^2(\mu, \sigma^2) = 1 - \sqrt{\frac{2\sigma\sigma_b}{\sigma^2 + \sigma_b^2}} e^{-\frac{1}{4} \frac{(\mu - \mu_b)^2}{\sigma^2 + \sigma_b^2}} \quad (27)$$

where μ and σ^2 are the estimated mean and variance to be checked, and μ_b and σ_b^2 are the benchmark. H^2 varies in $[0,1]$, and a smaller H^2 value indicates a more accurate estimation.

Since EnKF is based on Monte Carlo simulation, its performance varies with different sets of initial ensemble realizations, especially when the ensemble size is not large enough. To give a fair comparison, we run restart EnKF 20 times with different sets of initial realizations for each tested ensemble size. Then H^2 values between the final ensemble and reference ensemble (with ensemble size 1000) are calculated. The averages and standard deviations of 20 H^2 values for each tested ensemble size are plotted in Figure 6. With the ensemble size increasing from 25 to 500, the average values of H^2 are on the decline, which indicates an improving estimation accuracy. In this work, for fair comparisons, restart EnKF and RAPCKF both employ the restart scheme. Note that some extra computational cost in RAPCKF is spent in the pure matrix/vector calculation in recovering PCE coefficients. However, this is negligible compared to that spent in running the simulator. With the same number of realizations or ensemble members, the CPU time is almost the same for both methods. Thus the computational cost can be compared by the number of model evaluations. It is of interest to find that the H^2 value of RAPCKF is comparable to those of restart EnKFs with ensemble size 200 to 300. This means that they give similar estimation accuracy. It is worth noting that, the average number of model evaluations used in RAPCKF is only 36. Therefore, the superiority of RAPCKF is clearly demonstrated.

In practical applications, since the true parameter values are unavailable, another way to evaluate the accuracy is to see the match between the predicted and actual measurements. To further test the performance of RAPCKF, two performance indicators are defined as follows,

$$DH = \sqrt{\sum_{n=1}^N \sum_{m=1}^M [\mathbf{h}(n, m) - \mathbf{h}_{ref}(n, m)]^2 / (N \times M)} \quad (28)$$

$$DCH = \sqrt{\sum_{n=1}^N \sum_{m=1}^M [\mathbf{h}^c(n, m) - \mathbf{h}_{ref}^c(n, m)]^2 / (N \times M)} \quad (29)$$

where $\mathbf{h}$ and $\mathbf{h}_{ref}$ denote the predicted and reference values of pressure head on all the nodes at measurement time steps, respectively; $\mathbf{h}^c$ and $\mathbf{h}_{ref}^c$ have similar meanings as $\mathbf{h}$ and $\mathbf{h}_{ref}$ but defined under the second scenario; N is the number of repetitive runs, here $N=20$, n is the index; M is the total number of pressure heads, which equals to the product of the numbers of nodes and time steps, m is corresponding index. As shown in Eqs. (28-29), DH and DCH are performance indicators which are defined on the root of mean square error between predicted and true values of pressure head. Theoretically, smaller performance indicator values indicate the better performance.

Figure 7 shows the performance indicator values given by restart EnKF, RAPCKF and their variants without restart option (i.e., with the joint-updating scheme). It is seen that, the DH and DCH values apparently decrease with the utilization of the restart scheme, which means that the performance is much improved. It should be pointed out that the use of restart scheme will increase the CPU time for solving the model. In this simple 1D case, with the joint-updating scheme, the CPU time of running the simulator for one time step is 2 seconds; with the restart scheme, running the simulation for a whole period requires 5 seconds. Note that the temporal resolution is automatically determined by the solver Hydrus, and some extra time is required for file IO implementations for states in the joint-updating scheme. Therefore, the total time of simulation for a restart run is less than the product of time needed for each step and number of steps. It is also noted that, compared with APCKF, the average number of model evaluations at each assimilation step for RAPCKF is decreased from 49 to 36, which has been explained in previous section. This can somewhat offset the extra computational cost caused by restart scheme. In addition, the DH and DCH values of RAPCKF are both smaller than those of restart EnKF,

even when the ensemble size of restart EnKF is as large as 500. In conclusion, the superiority of RAPCKF is well demonstrated on both parameter estimation and state prediction.

4.2. Case 2: The estimation of hydraulic conductivity field

In this case, RAPCKF is applied to estimate the saturated hydraulic conductivity of a two-dimensional heterogeneous saturated hydraulic conductivity field. As shown in Figure 8, the size of study domain is 200 by 200 cm², which is discretized into 31 by 31 nodes. The red dots represent the observation locations for pressure head. Initially, the pressure head is -200 cm at the top and gradually changed to 0 at the bottom. The pressure heads of vertical cross-section have a linear range from -200 to 0 cm. The upper boundary of study domain is always exposed to the atmosphere without precipitation (and soil evaporation is negligible), the lower boundary condition is set as free drainage, and two lateral sides are impervious boundaries. The saturated hydraulic conductivity field is assumed to be uncertain and needs to be estimated from the pressure head values extracted from the 9 observation locations, which are measured every 1 day for the first 20 days of simulation period. The other VGM model parameters are assumed known as $\theta_r = 0.0001$ (cm³ cm⁻³), $\theta_s = 0.399$ (cm³ cm⁻³), $\alpha = 0.0174$ (cm⁻¹), and $n = 1.3757$ (dimensionless).

The spatially varying saturated conductivity field K_s are discretized by 961 nodes, which lead to a high-dimensional problem. Usually, the number of unknown parameters can be greatly reduced by dimension reduction techniques. In this scenario, the log conductivities $Y = \ln(K_s)$ are parameterized using the Karhunen-Loeve (KL) decomposition [23], which has mean $\mu_Y = 2$, variance $\sigma_Y^2 = 1$, and a separable exponential covariance function,

$$C_Y(\mathbf{x}_1, \mathbf{x}_2) = C_Y(x_1, y_1; x_2, y_2) = \sigma_Y^2 \exp \left[-\frac{|x_1 - x_2|}{\lambda_x} - \frac{|y_1 - y_2|}{\lambda_y} \right] \quad (30)$$

where the horizontal and vertical correlation lengths of the study domain

$\lambda_x = \lambda_y = 100$ cm. By keeping the 20 leading terms in KL decomposition, which requires 20 independent Gaussian random variables, 84% of the total variance of the random field can be kept. In this way, dimensionality is reduced from 961 to 20. Then the reference Y field and initial Y realizations are randomly generated by KL decomposition, respectively.

The mean log hydraulic conductivity fields estimated by RAPCKF and restart EnKF with different ensemble sizes (20, 50, 100, 200, 300 and 500) are plotted in Figure 9. It is shown that, the main patterns of the reference field can be captured by RAPCKF and restart EnKF with more than 200 ensemble members. In Figure 10, the variance of estimated log hydraulic conductivity fields provided by RAPCKF and restart EnKF are plotted. The uncertainty of estimated log hydraulic conductivity field is much underestimated when the ensemble size for restart EnKF is lower than 200. Figure 11 shows the number of model evaluations (for solving PCE coefficients) at each assimilation step during the implementation of RAPCKF. With the increase of assimilation step, the uncertainties of model outputs are on the decline and fewer PCE basis functions are required to represent the model outputs. The average number of model evaluations per loop for RAPCKF is 123.

To quantitatively compare the performance, two commonly used indicators are defined,

$$RMSE = \sqrt{\frac{1}{N_g} \sum_{i=1}^{N_g} (\bar{Y}_i - Y'_i)^2} \quad (31)$$

$$Spread = \sqrt{\frac{1}{N_g} \sum_{i=1}^{N_g} Var(Y_i)} \quad (32)$$

where $\bar{Y}_i$ and Y'_i denote the estimated and reference logarithm hydraulic conductivity, respectively; $Var(Y_i)$ denotes the ensemble variance of the logarithm hydraulic conductivity for the i th node; $N_g = 961$ is the total number of nodes in the study domain. The $RMSE$ measures the accuracy of mean estimation, while the $Spread$ measures the uncertainty in the estimation. The $RMSE$ and $Spread$ values of

RAPCKF are plotted in Figure 12. In general, the *RMSE* and *Spread* values are on the decline with the increase of the assimilation step. It can be shown that the difference between the squares of *RMSE* and *Spread* (*RS*) provides another measure of performance [22],

$$RS = RMSE^2 - Spread^2 = \frac{2}{N_g} \sum_{i=1}^{N_g} [\bar{Y}_i(\bar{Y}_i - Y'_i)] + \frac{1}{N_g} \sum_{i=1}^{N_g} [(Y'_i)^2 - (\bar{Y}_i^2)] \quad (33)$$

RS involves with both the mean and uncertainty estimations. If $\bar{Y} \rightarrow Y'$ and $\bar{Y}^2 \rightarrow (Y')^2$, then *RS* will tend to zero. Thus, the smaller *RS* value indicates the better parameter estimation.

We run restart EnKF 20 times with different sets of initial ensemble realizations for each tested ensemble size. The averages of three indicators (*RMSE*, *Spread* and *RS*) are compared in Figure 13. As for restart EnKF, with the increasing ensemble size, the *RMSE* and *RS* values are decreasing, while the *Spread* values are increasing. The *RMSE* value of RAPCKF is slightly smaller than that of restart EnKF with the ensemble size 200. Note that the average number of RAPCKF in each loop is only 123, the RAPCKF is thus still superior to restart EnKF with a similar computational cost if the accuracy of mean estimation is considered. As mentioned above, the match between *RMSE* and *Spread*, i.e., *RS*, is another performance indicator. As shown in Figure 13 (b), RAPCKF gives the smallest *RS* value. Therefore, if both the mean and uncertainty estimations are considered, RAPCKF outperforms restart EnKF, even the ensemble size of restart EnKF is as large as 500.

To further assess the efficiency of RAPCKF, more scenarios are investigated. In scenario 2, the variance of log hydraulic conductivity field is increased to 2, and all the other settings are the same. The performance indicator values of RAPCKF and restart EnKF are shown in Figure 14. It is seen that, due to the increased uncertainty, larger number of model evaluations is needed for RAPCKF to guarantee its accuracy. In scenarios 3 and 4, the vertical correlation length λ_y in Eq. (30) is decreased to 50

cm, and 50 leading KL terms are kept to parameterize the random field, in which 87% of the total variance of the random field is retained. The variances of log conductivity field are $\sigma_Y^2 = 1$ in scenario 3, and $\sigma_Y^2 = 2$ in scenario 4, respectively. As shown in Figures (15-16), the efficiency of RAPCKF decreases with the increasing heterogeneity of parameter field (larger variance and smaller correlation length). However, RAPCKF still has an advantage over restart EnKF with the same computational cost. The *RS* values of RAPCKF are comparable to those of restart EnKF with ensemble size 500, whereas the average numbers of model evaluations are 297 and 435, respectively.

5. Conclusions

In this study, a restart adaptive probabilistic collocation based Kalman filter (RAPCKF) algorithm has been proposed to estimate the soil hydraulic parameters in unsaturated flow problems. Unlike the classical PCKF, it can adaptively select active PCE basis functions and adjust the number of PCE terms during the implementation based on analysis of variance (ANOVA). Therefore, the issue of "curse-of-dimensionality" rooted in this polynomial chaos based approach can be greatly alleviated. In addition, a "restart" scheme is applied to eliminate the inconsistency between updated model parameters and states, which can improve the applicability of PCKF in strongly nonlinear problems.

The superiority of RAPCKF has been systematically demonstrated in one- and two- dimensional unsaturated flow problems. In our cases, the adaptive approach and restart scheme can significantly improve the performance of PCKF. Compared with restart EnKF, RAPCKF provides more accurate parameter estimation and model prediction with less computational cost. Although the use of restart scheme will increase the CPU time for a single model evaluation, the total number of model evaluations for RAPCKF is reduced, which can somewhat offset the extra computational cost.

It should be pointed out that, although the performance of PCKF can be much

improved by the utilization of adaptive scheme, it may still lose its efficiency advantage for extremely high-dimensional problems (>100), because the cost of calculating only low-order PCE terms may be already comparable to that of EnKF. Other dimension reduction techniques are needed in this situation. Furthermore, the requirement of rerunning from time zero indeed poses a significant computational burden for large-scale problems. There may be an optimal restart time window, which balances the computational cost and accuracy. These issues will be investigated in our future work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grants 41371237 and 41271470). Weixuan Li's work was supported by the Laboratory-Directed Research and Development program at Pacific Northwest National Laboratory through the Control of Complex Systems Initiative, and U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program as part of the Multifaceted Mathematics for Complex Energy Systems (M²ACS) project.

References

- [1] Vrugt JA, PH Stauffer, T Wöhling, BA Robinson, VV Vesselinov. Inverse modeling of subsurface flow and transport properties: A review with new developments. *Vadose Zone Journal*. 7 (2008) 843-64.
- [2] Yao Y, F Yang, EM Suuberg, J Provoost, W Liu. Estimation of contaminant subslab concentration in petroleum vapor intrusion. *Journal of hazardous materials*. 279 (2014) 336-47.
- [3] Yao Y, R Shen, KG Pennell, EM Suuberg. Estimation of Contaminant Subslab Concentration in Vapor Intrusion Including Lateral Source–Building Separation. *Vadose Zone Journal*. 12 (2013).
- [4] Chen Y, D Zhang. Data assimilation for transient flow in geologic formations via ensemble Kalman filter. *Advances in Water Resources*. 29 (2006) 1107-22.
- [5] Clark MP, DE Rupp, RA Woods, X Zheng, RP Ibbitt, AG Slater, et al. Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. *Advances in water resources*. 31 (2008) 1309-24.
- [6] Gharamti ME, I Hoteit, J Valstar. Dual states estimation of a subsurface flow-transport coupled model using ensemble Kalman filtering. *Advances in Water Resources*. 60 (2013) 75-88.
- [7] Li C, L Ren. Estimation of unsaturated soil hydraulic parameters using the ensemble Kalman filter. *Vadose Zone Journal*. 10 (2011) 1205-27.

543 [8] Houtekamer PL, HL Mitchell. Data assimilation using an ensemble Kalman filter technique. Monthly
544 Weather Review. 126 (1998) 796-811.

545 [9] Sun AY, A Morris, S Mohanty. Comparison of deterministic ensemble Kalman filters for assimilating
546 hydrogeological data. Advances in Water Resources. 32 (2009) 280-92.

547 [10] Sakov P, PR Oke. A deterministic formulation of the ensemble Kalman filter: an alternative to
548 ensemble square root filters. Tellus A. 60 (2008) 361-71.

549 [11] Bishop CH, BJ Etherton, SJ Majumdar. Adaptive sampling with the ensemble transform Kalman
550 filter. Part I: Theoretical aspects. Monthly weather review. 129 (2001) 420-36.

551 [12] Tippett MK, JL Anderson, CH Bishop, TM Hamill, JS Whitaker. Ensemble Square Root Filters*.
552 Monthly Weather Review. 131 (2003) 1485-90.

553 [13] Whitaker JS, TM Hamill. Ensemble Data Assimilation without Perturbed Observations. Monthly
554 Weather Review. 130 (2002) 1913-24.

555 [14] Hoteit I, D-T Pham, M Gharamti, X Luo. Mitigating observation perturbation sampling errors in the
556 stochastic EnKF. Monthly Weather Review. 143 (2015) 2918-36.

557 [15] Zeng L, L Shi, D Zhang, L Wu. A sparse grid based Bayesian method for contaminant source
558 identification. Advances in Water Resources. 37 (2012) 1-9.

559 [16] Zhang J, L Zeng, C Chen, D Chen, L Wu. Efficient Bayesian experimental design for contaminant
560 source identification. Water Resources Research. 51 (2015) 576-98.

561 [17] Shi X, M Ye, GP Curtis, GL Miller, PD Meyer, M Kohler, et al. Assessment of parametric uncertainty
562 for groundwater reactive transport modeling. Water Resources Research. 50 (2014) 4416-39.

563 [18] Zeng X, M Ye, J Burkardt, J Wu, D Wang, X Zhu. Evaluating two sparse grid surrogates and two
564 adaptation criteria for groundwater Bayesian uncertainty quantification. Journal of Hydrology. 535
565 (2016) 120-34.

566 [19] Saad G, R Ghanem. Characterization of reservoir simulation models using a polynomial chaos -
567 based ensemble Kalman filter. Water Resources Research. 45 (2009) W04417.

568 [20] Li H, D Zhang. Probabilistic collocation method for flow in porous media: Comparisons with other
569 stochastic methods. Water Resources Research. 43 (2007) W09409.

570 [21] Li W, Z Lu, D Zhang. Stochastic analysis of unsaturated flow with probabilistic collocation method.
571 Water resources research. 45 (2009) W08425.

572 [22] Zeng L, H Chang, D Zhang. A probabilistic collocation-based Kalman filter for history matching. SPE
573 J. 16 (2011) 294-306.

574 [23] Zeng L, D Zhang. A stochastic collocation based Kalman filter for data assimilation. Computational
575 Geosciences. 14 (2010) 721-44.

576 [24] Tartakovsky DM, DA Barajas-Solano. Uncertainty quantification in unsaturated flow models:
577 Progress and challenges. (2012).

578 [25] Li WX, G Lin, DX Zhang. An adaptive ANOVA-based PCKF for high-dimensional nonlinear inverse
579 modeling. Journal of Computational Physics. 258 (2014) 752-72.

580 [26] Moradkhani H, S Sorooshian, HV Gupta, PR Houser. Dual state-parameter estimation of
581 hydrological models using ensemble Kalman filter. Advances in Water Resources. 28 (2005) 135-47.

582 [27] Gu Y, DS Oliver. An Iterative Ensemble Kalman Filter for Multiphase Fluid Flow Data Assimilation.
583 SPE Journal. (2007) 438-446.

584 [28] Zafari M, AC Reynolds. Assessing the uncertainty in reservoir description and performance
585 predictions with the ensemble Kalman filter. SPE Annual Technical Conference and Exhibition. Society
586 of Petroleum Engineers. (2005).

- [29] Hendricks Franssen H, W Kinzelbach. Real - time groundwater flow modeling with the ensemble Kalman filter: Joint estimation of states and parameters and the filter inbreeding problem. *Water Resources Research*. 44 (2008) W09408.
- [30] Song X, L Shi, M Ye, J Yang, IM Navon. Numerical Comparison of Iterative Ensemble Kalman Filters for Unsaturated Flow Inverse Modeling. *Vadose Zone Journal*. 13 (2014).
- [31] Wen XH, WH Chen. Real-Time Reservoir Model Updating Using Ensemble Kalman Filter With Confirming Option. *Society of Petroleum Engineers*. (2005) 431-442.
- [32] Gharamti ME, B Ait-El-Fquih, I Hoteit. An iterative ensemble Kalman filter with one-step-ahead smoothing for state-parameters estimation of contaminant transport models. *Journal of Hydrology*. 527 (2015) 442-57.
- [33] Chen Y, DS Oliver, D Zhang. Data assimilation for nonlinear problems by ensemble Kalman filter with reparameterization. *Journal of Petroleum Science and Engineering*. 66 (2009) 1-14.
- [34] Liao Q, D Zhang. Data Assimilation for Strongly Nonlinear Problems by Transformed Ensemble Kalman Filter. *SPE Journal*. (2014).
- [35] Van Genuchten MT. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil science society of America journal*. 44 (1980) 892-8.
- [36] Mualem Y. A new model for predicting the hydraulic conductivity of unsaturated porous media. *Water resources research*. 12 (1976) 513-22.
- [37] Evensen G. The Ensemble Kalman Filter: theoretical formulation and practical implementation. *Ocean Dynamics*. 53 (2003) 343-67.
- [38] Evensen G. Data assimilation: the ensemble Kalman filter. *Springer Science & Business Media*. (2009).
- [39] Carsel RF, RS Parrish. Developing joint probability distributions of soil water retention characteristics. *Water Resources Research*. 24 (1988) 755-69.
- [40] Johnson NL, S Kotz. *Distributions in Statistics: Continuous Univariate Distributions*. Houghton Mifflin. 1 (1970).
- [41] Bertino L, G Evensen, H Wackernagel. Sequential Data Assimilation Techniques in Oceanography. *International Statistical Review*. 71 (2003) 223-41.
- [42] Scharnagl B, J Vrugt, H Vereecken, M Herbst. Bayesian inverse modelling of in situ soil water dynamics: using prior information about the soil hydraulic properties. *Hydrol Earth Syst Sci Discuss*. 8 (2011) 2019-63.
- [43] Høyland K, M Kaut, S Wallace. A Heuristic for Moment-Matching Scenario Generation. *Computational Optimization and Applications*. 24 (2003) 169-85.

631

Table 1. Statistical parameters used for distribution approximation

Soil Texture	Hydraulic Parameter	Transformation Type	Limits of Variation		Mean	Standard Deviation
			A	B		
Sand	K_s	SB	0.00	70.00	-0.394	1.150
	θ_r	LN	0.00	0.10	-3.120	0.224
	α	SB	0.00	0.25	0.378	0.439
	n	LN	1.50	4.00	0.978	0.100
Loamy Sand	K_s	SB	0.00	51.00	-1.270	1.400
	θ_r	SB	0.00	0.11	0.075	0.567
	α	LN	0.00	0.25	0.124	0.043
	n	SB	1.35	5.00	-1.110	0.307
Sandy Loam	K_s	SB	0.00	30.00	2.490	1.530
	θ_r	SB	0.00	0.11	0.384	0.700
	α	SB	0.00	0.25	-0.937	0.764
	n	LN	1.35	3.00	0.634	0.082

632

633

Table 2. Correlations among transformed variables presented with the factored covariance matrix

635

		K_s	θ_r	α	n
Sand	K_s	1.040	-0.109	0.328	0.081
	θ_r		0.182	0.258	-0.047
	α			0.143	-0.011
	n				0.017
Loamy Sand	K_s	1.480	-0.201	0.037	0.211
	θ_r		0.522	0.017	-0.194
	α			0.014	0.019
	n				0.108
Sandy Loam	K_s	1.600	-0.201	0.037	0.211

θ_r	0.538	0.017	-0.194
α		0.014	0.019
n			0.108

Figure 1. The schematic of experimental setup.

Figure 2. True and estimated water retention curve for (a) sand, (b) loamy sand and (c) sandy loam.

Figure 3. True and estimated hydraulic conductivity function for (a) sand, (b) loamy sand and (c) sandy loam.

Figure 4. The predicted pressure heads on observed nodes at measurement time steps generated by running the simulator with SHP parameters sampled from the final PCE representation under two different scenarios (I and II).

Figure 5. Number of model evaluations at each assimilation step for RAPCKF.

Figure 6. The H^2 values of RAPCKF and restart EnKF with different ensemble sizes. The average number of model evaluations for RAPCKF is 36.

Figure 7. Performance indicators values, defined on the root of mean square error between predicted and true values of pressure head, provided by RAPCKF (without restart option) and restart EnKF (without restart option) with different ensemble sizes under two different scenarios.

Figure 8. Study domain, boundary conditions, and 9 locations (red dots) for observations of pressure head.

Figure 9. The mean log hydraulic conductivity fields estimated by RAPCKF (g-123) and restart EnKF with different ensemble sizes (a-20, b-50, c-100, d-200, e-300 and f-500).

Figure 10. The variance of log hydraulic conductivity fields estimated by RAPCKF (g-123) and restart EnKF with different ensemble sizes (a-20, b-50, c-100, d-200, e-300 and f-500).

Figure 11. Number of model evaluations at each assimilation step for RAPCKF.

Figure 12. The *RMSE* and *Spread* values of RAPCKF at each assimilation step.

Figure 13. The *RMSE* and *Spread* values, provided by RAPCKF and restart EnKF with different ensemble sizes in scenario 1 ($\lambda_x = 100$ cm, $\lambda_y = 100$ cm and $\sigma_Y^2 = 1$). The average number of model evaluations for RAPCKF is 123.

Figure 14. The *RMSE* and *Spread* values, provided by RAPCKF and restart EnKF with different ensemble sizes in scenario 2 ($\lambda_x = 100$ cm, $\lambda_y = 100$ cm and $\sigma_Y^2 = 2$). The average number of model evaluations for RAPCKF is 210.

Figure 15. The *RMSE* and *Spread* values, provided by RAPCKF and restart EnKF with different ensemble sizes in scenario 3 ($\lambda_x = 100$ cm, $\lambda_y = 50$ cm and $\sigma_Y^2 = 1$).

The average number of model evaluations for RAPCKF is 297.

Figure 16. The *RMSE* and *Spread* values, provided by RAPCKF and restart EnKF with different ensemble sizes in scenario 4 ($\lambda_x = 100$ cm, $\lambda_y = 50$ cm and $\sigma_Y^2 = 2$).

The average number of model evaluations for RAPCKF is 435.