Development of Integrated Approaches for Hydrological Data Assimilation through Combination of Ensemble Kalman Filter and Particle Filter Methods

Y.R. Fan[a], G.H., Huang[a,b*], B.W. Baetz[c], Y.P., Li[b], K. Huang[d], X., Chen[e], M. Gao[d]

[a] Institute for Energy, Environment and Sustainable Communities, University of Regina, Regina, Saskatchewan, Canada S4S 0A2

[b] School of Environment, Beijing Normal University, Beijing, China, 100875

[c] Department of Civil Engineering, McMaster University, Hamilton, ON L8S 4L8, Canada

[d] Faculty of Engineering and Applied Science, University of Regina, Regina, Saskatchewan, Canada S4S 0A2

[e] State Key Laboratory of Hydrology-Water Resource and Hydraulic Engineering, Hohai University, Nanjing, China, 210098

*Correspondence: Dr. G. H. Huang

Institute for Energy, Environment and Sustainable Communities, University of Regina

Regina, Saskatchewan, S4S 0A2, Canada

Tel: (306) 585-4095

Fax: (306) 585-4855

E-mail: huangg@uregina.ca

Abstract:

This study improved hydrologic data assimilation through integrating the capabilities of particle filter (PF) and ensemble Kalman filter (EnKF) methods, leading to two integrated data assimilation schemes: the coupled EnKF and PF (CEnPF) and parallelized EnKF and PF (PEnPF) approaches. The applicability and usefulness of CEnPF and PEnPF were demonstrated using a conceptual rainfall-runoff model. The performance of two new developed data assimilation methods and traditional EnKF and PF approaches was tested through a synthetic experiment and two real-world cases with one located in he Jing River basin and one located in the Yangtze river basin. The results show that both PEnPF and CEnPF approaches have more opportunities to provide better results for both deterministic and probabilistic predictions than traditional EnKF and PF approaches. Moreover, the computational time of the two integrated methods is manageable. But the proposed PEnPF may need much more time for some large-scale or time-consuming hydrologic models since it generally needs three times of model runs of EnKF, PF and CEnPF.

Keywords: Hydrologic Prediction, Data assimilation, Ensemble Kalman filte, Particle filter, Uncertainty

1. Introduction

The great increase in computing power and hydrologic data availability has resulted in increasingly use of hydrologic models in real world applications (Montanari and Brath, 2004). However, significant uncertainties are associated with rainfall-runoff simulation and it is of great importance to account for these uncertainties in hydrologic predictions (e.g., Pappenberger and Beven, 2006; Schaake et al., 2006; Brown, 2010). Uncertainty in hydrologic predictions may result from several major sources, including errors in the model structure and model parameters, as well as model initial conditions and forcing data (e.g., Ajami et al., 2007; Kavetski et al., 2006a, b; Salamon and Feyen, 2010; Liu et al., 2012). Effective quantification and reduction of these uncertainties is necessary to provide reliable hydrologic forecasts for estimating designated variables in engineering practice, mitigating hydrological risks and improving water resource management policies (DeChant and Moradkhani, 2014; Fan et al., 2015a,c; Kong et al., 2015; Li et al., 2015; Yan et al., 2015).

Previously, a great number of approaches have been proposed for quantifying the uncertainty in hydrologic predictions (De Lannoy et al., 2007; Parrish et al., 2012; DeChant and Moradkhani, 2014; Madadgar and Moradkhani, 2014; Su et al., 2014). Sequential data assimilation techniques are widely used for explicitly dealing with various uncertainties and for optimally merging observations into uncertain model predictions (Reichle et al., 2002; Moradkhani et al., 2005a; Vrugt et al., 2005; Clark et al., 2008; Xie and Zhang, 2013; Fan et al., 2015b). The state variables and parameters in a hydrologic model can be continuously updated when new measurements are available through sequential data assimilation techniques, and such a process can highly improve the model predictions. The ensemble Kalman filter (EnKF) and the particle filter (PF) are two of the most widely used sequential data assimilation schemes.

The EnKF technique approximates the distribution of the system state using random samples, called ensemble, and replaces the covariance matrix by the sample covariance computed from the ensemble, which is used for state updating in the Kalman filter formula (Evensen, 1994). The EnKF approach is much attractive in hydrologic predictions due to its features of real-time adjustment and easy implementation (Reichle et al., 2002). It can provide a general framework for dynamic state, parameter, and joint state-parameter estimation in hydrologic models. For instance, Moradkhani et al. (2005a) initially proposed a dual-state estimation approach based on EnKF for sequential estimation for both the parameters and state variables of a hydrologic model. Weerts and EI Serafy (2006) compared the capability of EnKF and particle filter (PF) methods in reducing uncertainty in the rainfall-runoff update and internal model state estimation for flooding forecasting purposes. Parrish et al. (2012) integrated Bayesian model averaging and data assimilation to reduce model uncertainty. DeChant and Moradkhani (2014) combined ensemble data assimilation and sequential Bayesian methods to provide a reliable prediction of seasonal forecast uncertainty. Shi et al. (2014) conducted multiple parameter estimation using multivariate observations via the ensemble Kalman filter (EnKF) for a physically-based land surface hydrologic model. Pathiraja et al. (2016a, b) proposed EnKF-based approaches to detect non-stationary hydrologic model parameters in a paired catchment systems.

In comparison with EnKF, the particle filter (PF) method also uses random samples (i.e. particles) to approximate the distributions of the model state. However, these particles are updated forward by using sequential Monte Carlo (SMC) simulation. The most significant advantage of PF is that it relaxes the assumption of Gaussian distribution in state-space model errors, which is required for EnKF. Furthermore, Liu et al. (2012) stated that the PF approaches can reduce numerical instability especially in physically-based or process-based models, since they performs updating on the particle weights instead of the state variables

(Liu et al., 2012). The initial implementation of PF is based on sequential importance sampling, which usually leads to severe deterioration for particles (i.e. only several or even one particle would be available). Consequently, sampling importance resampling (SIR) techniques have been proposed to mitigate this problem (e.g. Moradkhani et al., 2005b; Li et al., 2015; Fan et al., 2016). However, previous studies in other fields have concluded that the PF method usually requires more samples than other filtering methods and the sample size would increase exponentially with the number of state variables (Liu and Chen, 1998; Fearnhead and Clifford, 2003; Snyder et al., 2008). Specifically, a great number of samples may be required for reliable characterization of the posterior probability density functions (PDFs) even for small problems with only a few unknown states and parameters (Liu et al., 2012). Thus, the applications of PF suffer from the number requirement of particles, especially for physically-based distributed hydrologic models (Liu et al., 2012). Recent improvements for PF are to combine the strengths of sequential Monte Carlo sampling and Markov chain Monte Carlo simulation to achieve a more complete representation of the posterior distribution (Moradkhani et al., 2012; Vrugt et al., 2013). Such improvements can mitigate sample impoverishment (i.e. a decrease in the diversity of the particles or even a single particle available after resampling steps), and may lead to a more accurate streamflow forecast with small, manageable ensemble sizes (Moradkhani et al., 2012). Recently, Yan and Moradkhani (2016) demonstrated the application of integration of particle filter and Markov chain Monte Carlo (PF-MCMC) methods by a distributed Sacramento Soil Moisture Accounting (SAC-SMA) model.

Both EnKF and PF have been widely used for characterizing uncertainties in hydrologic models. Each of them has its own advantages and drawbacks. The EnKF provides good estimates for very small ensembles but it suffers from its inherent Gaussian assumption (Shen and Tang, 2015). The PF relaxes the Gaussian assumption and is able to outperform the EnKF if the ensemble size is sufficiently large to prevent filter degeneracy (Moradkhani, 2008; Leisenring and Moradkhani 2012; Shen and Tang, 2015), but it may not recuperate quickly if the particle ensemble consistently over or underestimates the respective observation (Vrugt et al., 2013). Integration of EnKF and PF may be an alternative for overcoming the shortcomings in EnKF and PF, (Frei and Künsch, 2013; Rezaie and Eidsvik, 2012; Plaza-Guingla et al., 2013; Shen and Tang, 2015). For instance, Shen and Tang (2015) proposed a modified ensemble Kalman particle filter for non-Gaussian systems with nonlinear measurement functions by providing a continuous interpolation between the EnKF and PF analysis schemes. The results showed that the proposed method, given an affordable ensemble size, can perform better than the EnKF for nonlinear systems with nonlinear observations (Shen and Tang, 2015).

As an extension of previous research, this study aims to develop integrated approaches for hydrologic data assimilation. In detail, two integrated data assimilation approaches are firstly proposed through integrating EnKF and PF: the coupled EnKF and PF (abbreviated as CEnPF) and the parallelized EnKF and PF (abbreviated as PEnPF). The CEnPF sequentially will employ the EnKF and PF to update model parameters and states, in which the EnKF is initially applied to correct model states and parameters, and PF is then adopted to eliminate insignificant particles. In comparison, the PEnPF approach simultaneously updates model states and parameters in parallel through EnKF and PF, and chooses the better estimates as the posterior distributions.

2. Methodology

In a sequential data assimilation process, the state variables in a hydrologic model can be evolved forward as follows:

$$x_t = f(x_{t-1}, u_{t-1}, \theta) + \omega_{t-1} \tag{1}$$

where the subscript t denotes the time step; f is a nonlinear function expressing the system transition from time t - 1 to t; x_t denote the state variables, and θ are the model parameters; ω_{t-1} is considered as process noise (i.e. model error). The model output y_t related to real measurements (e.g. streamflow) can be obtained through the measurement operator $h(.)$, subject to model states and parameters as follows:

$$y_t = h(x_t, \theta) + v_t \tag{2}$$

where h is the nonlinear function producing forecasted observations; v_t is the observation noise.

The essence of the parameter and state estimation problem in the Bayesian filtering framework is to construct the posterior probability density functions (PDFs) of parameters and states conditioned on all previous observations ($y_{1:t-1}$) and current available observation (y_t) (Gordon et al., 1993; Fan et al., 2016). The posterior PDF can be calculated in two steps theoretically: prediction and update, in which the state PDF from the previous state would be integrated through the system model, and the update operation modifies the prediction PDF making use of the latest observations (Han and Li, 2008). The prediction step aims to obtain the prior $p(x_t \mid y_{1:t-1})$ through the following model:

$$p(x_t \mid y_{1:t-1}) = \int p(x_t \mid x_{t-1}) p(x_{t-1} \mid y_{1:t-1}) dx_{t-1} \tag{3}$$

where $p(x_t \mid x_{t-1})$ is the transition probability to describe evolution of states and can be obtained by Equation (1). $p(x_{t-1} \mid y_{1:t-1})$ is the posterior distribution at time step t-1.When new observations at time t are available, the prior can be corrected according to Bayes' rule, formulated as follows:

$$p(x_t \mid y_{1:t}) = \frac{p(y_t \mid x_t)\, p(x_t \mid y_{1:t-1})}{\int p(y_t \mid x_t)\, p(x_t \mid y_{1:t-1})\, dx_t} \tag{4}$$

where $p(x_t \mid y_{1:t-1})$ represents the prior information; $p(y_t \mid x_t)$ is the likelihood.

The optimal Bayesian solution (i.e. Equations (3) and (4)) is difficult to determine since the evaluation of the integrals may be intractable (Plaza-Guingla et al., 2013). Consequently, approximation methods are applied to address the above issues. Ensemble Kalman filter (EnKF) and PF approaches are the two most widely used methods. The central idea of EnKF and PF is to represent the state probability density function (pdf) as a set of random samples and the difference between these two methods lies in the way of recursively generating an approximation to the state PDF (Weerts and EI Serafy, 2005).

2.1. Ensemble Kalman Filter

The EnKF and its variants use ensembles of states to approximate the covariance matrices to achieve suboptimal state estimations in which the error statistics are analyzed by numerically solving the Fokker-Planck equation using the Monte Carlo method (Evensen, 2003; Shen and Tang, 2015). EnKF-based filters normally distributed errors and the Monte Carlo approach is applied to approximate the error statistics, as well as compute an approximate Kalman gain matrix for updating model and state variables. A general framework of EnKF for states and parameters updating is described below, followed the description in Moradkhani et al. (2005b).

In the implementation of EnKF, the prior and posterior distributions for model parameters and state variables are characterized by random samples name "ensembles". At any given time t, the prior and posterior distributions of states and parameter are assumed to be denoted through a set of ensembles below

$$X_t^f = (x_{t,1}^f, \ldots, x_{t,i}^f, \ldots, x_{t,ne}^f) \quad (1)$$

$$\Psi_t^f = (\theta_{t,1}^f, \ldots, \theta_{t,i}^f, \ldots, \theta_{t,ne}^f) \quad (2)$$

$$X_t^a = (x_{t,1}^a, \ldots, x_{t,i}^a, \ldots, x_{t,ne}^a) \quad (3)$$

$$\Psi_t^a = (\theta_{t,1}^a, \ldots, \theta_{t,i}^a, \ldots, \theta_{t,ne}^a) \quad (4)$$

where the superscript f indicates the "forecast" values indicating the prior distributional information and the superscript a indicates the "analyzed" values after assimilation which denotes the posterior distributional information; the subscript i refers to the i^{th} ensemble member, and ne denotes the total number of ensembles. Consider a stochastic dynamic-state model $f(x, u, \theta)$ described by state vector x, parameter vector θ and forcing data u, the state propagation can be expressed as:

$$x_{t+1,i}^f = f(x_{t,i}^a, u_{t,i}, \theta_{t+1,i}^f) + \omega_{t,i}, \; i = 1, 2, \ldots, ne \quad (5)$$

where ω_t is the model error term, which follows a Gaussian distribution with zero mean and covariance matrix P_t. To implement model (5), parameter evolution should be conducted. A number of parameter evolution approaches have been developed (e.g. Fan et al., 2015b; Pathiraja et al., 2016a,b). Among these methods, the random walk method is widely used, in which stochastic perturbations with mean values of zero and heteroscedastic variances are added to the analyzed ensembles in the previous stage as follows:

$$\theta_{t+1,i}^f = \theta_{t,i}^a + \tau_{t,i}, \; \tau_{t,i} \sim N(0, \Sigma_t^\theta) \quad (6)$$

where Σ_t^θ is the covariance matrix of the analyzed parameter ensembles at time t.

Based on the forecasts in model states and parameters, the corresponding observation values can be obtained through an observation equation characterized as:

$$y_{t+1,i}^{f} = h(x_{t+1,i}^{f}, \theta_{t+1,i}^{f}) + v_{t+1,i}, v_{t+1,i} \sim N(0, \Sigma_{t+1}^{y}) \quad (7)$$

where h represents the operator to transfer the states into the observation space, $v_{t+1,i}$ indicates the random perturbation in model prediction, which is drawn from a normal distribution with a mean value of zero and a covariance of Σ_{t+1}^{y}. When new observations at time step $t+1$ are available, model states and parameters are corrected by assimilating the observation into modelling process, leading to analyzed ensembles indicating the posterior distributions for model states and parameters. Before assimilating observations, stochastic perturbations are usually added to the observations to account for the uncertainty in measurements. In this process, Gaussian noise is generally employed expressed as:

$$y_{t+1,i}^{o} = y_{t+1} + \varepsilon_{t+1,i}, \varepsilon_{t+1,i} \sim N(0, \Sigma_{t+1}^{y^o}) \quad (8)$$

where y_{t+1} represents the raw observation and $\Sigma_{t+1}^{y^o}$ denotes the error covariance. Through assimilating the observations, the posterior states and parameters can be updated by the Kalman update equations:

$$x_{t+1,i}^{a} = x_{t+1,i}^{f} + K_{xy}[y_{t+1,i}^{o} - y_{t+1,i}^{f}] \quad (9)$$

$$\theta_{t+1,i}^{a} = \theta_{t+1,i}^{f} + K_{\theta y}[y_{t+1,i}^{o} - y_{t+1,i}^{f}] \quad (10)$$

where K_{xy}, $K_{\theta y}$ are Kalman matrix for states and parameters, which can be expressed as follows (DeChant and Moradkhani, 2012; Pathiraja et al., 2016a):

$$K_{xy} = \Sigma_{t+1}^{xy}(\Sigma_{t+1}^{y} + \Sigma_{t+1}^{y^o})^{-1} \quad (11)$$

$$K_{\theta y} = \Sigma_{t+1}^{\theta y}(\Sigma_{t+1}^{y} + \Sigma_{t+1}^{y^o})^{-1} \quad (12)$$

where Σ_{t+1}^{xy} is the cross covariance of the forecasted states $x_{t+1,i}^{f}$ and the simulated observation $y_{t+1,i}^{f}$; $\Sigma_{t+1}^{\theta y}$ is the cross covariance between model parameters $\theta_{t+1,i}^{f}$ and the simulated observation $y_{t+1,i}^{f}$

2.2. Particle Filter

The PF, similar to the EnKF, is a kind of sequential Monte Carlo method that calculates the posterior distribution of states and parameters by a set of random samples. But PF and its variants are different from EnKF since the ensemble members (or the particles) are not modified, but are combined with different weights (Shen and Tang, 2015). It was found that PF outperforms EnKF by relaxing the assumption of a Gaussian error structure, which allows PF to accurately predict the posterior distribution in the presence of skewed distributions (Moradkhani et al., 2005a; DeChant and Moradkhani, 2012).

In detail, consider *ne* independent and identically distributed random variables $x_{t,i} \sim p(x_t \mid y_{1:t})$ for i = 1, 2, …, *ne*, the posterior density, based on the sequential importance sampling (SIS) method, can then be approximated as a discrete function:

$$p(x_t \mid y_{1:t}) = \sum_{i=1}^{ne} w_{t,i}\delta(x_t - x_{t,i}) \tag{13}$$

where $w_{t,i}$ is the posterior (updated) normalized weight of the ith particle drawn from the proposed distribution; δ is the Dirac delta function. Assume the system state to be a Markov process, and apply the Bayesian recursive expression to the filtering problem. The updating expression for the importance weights (not normalized) is expressed as:

$$w_{t,i}^{a*} = w_{t,i}^{f} \cdot \frac{L_\theta(y_t \mid x_{t,i}^f) p_\theta(x_{t,i}^f \mid x_{t-1,i}^f)}{q_\theta(x_{t,i}^f \mid x_{t-1,i}^f, y_t^f)} \quad (14)$$

where $w_{t,i}^f$ is the prior weight, which is equal to the posterior weight at the previous time step. $w_{t,i}^{a*}$ is the unnormalized posterior weight. Through Equation (14), the importance weights are sequentially updated when an appropriate proposal distribution $q_\theta(x_{t,i}^f \mid x_{t-1,i}^f, y_t^f)$ is given. Consequently, the expression of the proposal distribution will significantly affect the efficiency and complexity of the PF method. Gordon et al. (1993) have suggested to set $q_\theta(x_{t,i}^f \mid x_{t-1,i}^f, y_t^f) = p_\theta(x_{t,i}^f \mid x_{t-1,i}^f)$, resulting in a simplified expression for importance weights:

$$w_{t,i}^{a} = w_{t,i}^{f} L_\theta(y_t \mid x_{t,i}^f) \quad (15)$$

Therefore, the normalized updating weight can then be obtained via the following equation:

$$w_{t,i}^{a} = \frac{w_{t,i}^{f} L_\theta(y_t \mid x_{t,i}^f)}{\sum_{i=1}^{ne} w_{t,i}^{f} L_\theta(y_t \mid x_{t,i}^f)} \quad (16)$$

$w_{t,i}^a$ is the normalized posterior weight. $L_\theta(y_t \mid x_{t,i}^f)$ is the posterior likelihood function. The choice of an adequate likelihood function has been the subject of considerable debate in hydrologic and statistics literature (Vrugt et al., 2013). In the data assimilation process through PF, the Gaussian likelihood is widely used in a number of fields (Moradkhani et al., 2005b; Weerts and EI Serafy, 2006; Salamon and Feyen, 2010; Fan et al., 2016). Consequently, this study will also adopt the Gaussian likelihood expressed as:

$$L_\theta(y_t \mid x_{t,i}^f) = \frac{1}{\sqrt{2\pi R_t}} \exp(-\frac{1}{2R_t}[y_t - y_{t,i}^f]^2) \quad (17)$$

For the particle filter through SIS, a serious limitation is the depletion of the particle set, which means that, after a few iterations (time steps), all the particles except one are discarded because their importance weights are insignificant (Doucet, et al. 2001). To address the above issue, sampling importance resampling (SIR) algorithms are usually applied to eliminate the particles with small importance weights and replace them by particles with large importance weights. A number of resampling approaches have been developed, such as multinomial resampling, systematic resampling, residual resampling, and grouping-based resampling approaches (Li et al., 2015)

2.3. Integration of EnKF and PF for Hydrologic Data Assimilation

The application of EnKF is constrained by its assumption of Gaussian errors while the PF requires a large sample size for providing reliable predictions. In this study, we extend the previous research to provide two integrated data assimilation schemes: the coupled EnKF and PF (abbreviated as CEnPF) and the parallelized EnKF and PF (abbreviated as PEnPF) approaches to characterize uncertainty in hydrologic models.

2.3.1. the coupled EnKF and PF (CEnPF) approach

The CEnPF sequentially uses the EnKF and PF to update model parameters and states, in which EnKF is first applied to correct model states and parameters, and PF is then adopted to eliminate insignificant particles (see Figure 1). The detailed procedures for the implementation of CEnPF are presented as follows:

Step 1. Similar to the implementation of EnKF and PF, the model initial conditions should be assumed before implementing CEnPF. In this study, the initial state variables and parameters are sampled from the corresponding uniform distributions:

$$x_{1,i} \sim \mathrm{U}(x^L, x^U),\ i = 1, 2, \ldots, ne,\quad x \in R^{N_x} \tag{18}$$

$$\theta_{1,i} \sim \sim \mathrm{U}(\theta^L, \theta^U),\ i = 1, 2, \ldots, ne,\quad \theta \in R^{N_\theta} \tag{19}$$

Step 1. Assign prior weights for the ensembles. In general, the prior weights are assigned uniformly as follows:

$$w_{t,i} = 1/ne,\ i = 1, 2, \ldots, ne \tag{20}$$

Step 3. At any time step t, model states at current step can be forecasted based the posterior states in step t-1 and the prior parameters in the current step by using model operator f:

$$x_{t,i}^f = f(x_{t-1,i}^a, u_{t,i}, \theta_{t,i}^f) + \omega_{t,i}, \omega_t \sim N(0, \Sigma_t^m),\ i = 1, 2, \ldots, ne \tag{21}$$

where parameters $\theta_{t,i}^f$ are obtained by Equation (6).

Step 5. Observation simulation: Use the observation operator h to propagate the model state forecast:

$$y_{t,i}^f = h(x_{t,i}^f, \theta_{t,i}^f) + v_{t,i},\quad v_{t+1,i} \sim N(0, \Sigma_t^y),\ i = 1, 2, \ldots, ne \tag{22}$$

Step 6. Parameters and states updating: Update the parameters and states via the EnKF updating equations

$$x_{t,i}^a = x_{t,i}^f + K_{xy}[y_{t,i}^o - y_{t,i}^f] \tag{23}$$

$$\theta_{t,i}^a = \theta_{t,i}^f + K_{\theta y}[y_{t,i}^o - y_{t,i}^f] \tag{24}$$

where $x_{t,i}^a$ and $\theta_{t,i}^a$ are the updated state and parameter values and K_{xy} and $K_{\theta y}$ are the Kalman matrix for states and parameters obtained by Equations (11) and (12).

Step 7. Estimate the likelihood:

$$L(y_t \mid x_{t,i}^a, \theta_{t,i}^a) = \frac{1}{\sqrt{2\pi R_t}} \exp(-\frac{1}{2R_t}[y_{t,i}^o - h(x_{t,i}^a, \theta_{t,i}^a)]^2) \tag{25}$$

$$p(y_t \mid x_{t,i}^a, \theta_{t,i}^a) = \frac{L(y_t \mid x_{t,i}^a, \theta_{t,i}^a)}{\sum_{i=1}^{ne} L(y_t \mid x_{t,i}^a, \theta_{t,i}^a)} = p(y_{t,i}^o - h(x_{t,i}^a, \theta_{t,i}^a) \mid R_t) \tag{26}$$

Step 8. update weight for the analyzed ensemble values:

$$w_{t,i}^{a} = \frac{w_{t,i}^{f} \cdot p(y_{t,i}^{o} - h(x_{t,i}^{a}, \theta_{t,i}^{a}) \mid R_t)}{\sum_{i=1}^{ne} w_{t,i}^{f} \cdot p(y_{t,i}^{o} - h(x_{t,i}^{a}, \theta_{t,i}^{a}) \mid R_t)} \quad (27)$$

where $w_{t,i}^{f}$ are the prior sample weights and are usually set to be 1/*ne*.

Step 9. Resampling: Apply resampling procedure proposed by Moradkhani et al. (2 005a) to eliminate the abnormal samples in $x_{t,i}^{a}$, and $\theta_{t,i}^{a}$, and generate resampled ensembles denoted as $x_{t-resamp,i}^{a}, \theta_{t-resamp,i}^{a}$.

Step 10. Parameter perturbation: take parameter evolution to the next stage through adding small stochastic error around the sample:

$$\theta_{t+1,i}^{f} = \theta_{t-resamp,i}^{a} + \varepsilon_{t,i}, \quad \varepsilon_{t,i} \sim N(0, \eta S(\theta_{t-resamp,i}^{a})) \quad (28)$$

where η is a hyper-parameter which determines the radius around each sample being explored; $S(\theta_{t-resamp,i}^{a})$ is the standard deviation of the analyzed ensemble values.

Step 11. Check the stopping criterion: if measurement data is still available in the next stage, *t* = t + 1 return to step 3; otherwise, stop.

In CEnPF, model parameters and states are initially updated through Kalman update equations, then the updated states and parameters are corrected again through PF procedure to eliminate abnormal or insignificant state and parameters and replace them by significant ones by sampling importance resampling procedure. Compared with EnKF, the CEnPF can be applicable for nonlinear and non-Gaussian systems. At any time step *t*, even though the EnKF procedure may not produce optimal states and parameters under nonlinear and non-Gaussian systems, the following PF procedure can remove non-optimal ensembles (i.e. insignificant samples) and replace them with significant ones. In comparison with PF, the proposed CEnPF firstly reduces the sample requirement for large-scale models since the inherent EnKF procedure can achieve satisfactory performance with a moderate sample size; it can also adjust the ensemble values to fit the observations well especially when the particle ensembles consistently over or underestimates the respective observations.

2.3.2. the parallelized EnKF and PF (PEnPF) approach

In comparison with CEnPF, the PEnPF approach simultaneously updates model states and parameters in parallel through EnKF and PF, and chooses the better estimates as the posterior distributions (see Figure 2). The full description of the PEnPF procedures is illustrated as follows:

Step 1. Model state initialization: Initialize N_x-dimensional model state variables and N_θ-dimensional model parameters from uniform distributions expressed as Equations (18) and (19)

Step 2. Sample weight assignment: Assign the prior weights uniformly to the particles expressed as Equation (20):

Step 4. Model state forecast step: Propagate the *ne* state variables and model parameters forward in time using model operator *f* by Equation (21).

Step 5. Observation simulation: Use the observation operator *h* to propagate the model state forecasts by Equation (22):

Step 6. Parameters and states updating based on EnKF: This step is further divided into two procedures: model parameters and states are updated by Kalman updating scheme and the updated ensembles are evaluated by a mismatch index proposed by Gu and Oliver (2007).

6a. Obtain the analyzed estimations through Kalman updating scheme expressed as Equations (23) and (24)

6b. Evaluate the data match term for the analyzed estimation by the mismatch index expressed by:

$$S(x_{t,i}^{a},\theta_{t,i}^{a})=\sum_{i=1}^{ne}(h(x_{t,i}^{a},\theta_{t,i}^{a})-y_{t,i}^{o})^{T}R_{t}^{-1}(h(x_{t,i}^{a},\theta_{t,i}^{a})-y_{t,i}^{o}) \quad (29)$$

Such an index has been adopted in several data assimilation literatures (e.g. Gu and Oliver

2007; Chen and Oliver, 2013; Zhang et al., 2014) to evaluate history-matching results. In this study, this index is used to evaluate the performance of the updated states and parameters obtained from Kalman updating scheme.

Step 7. Different from the CEnPF in which PF updates model parameters and states based on the analyzed state and parameter values from EnKF, the PF procedure in PEnPF also update model states and parameters from the priori states and parameters at time t. Therefore, the likelihood function can be expressed as:

$$L(y_t \mid x_{t,i}^f, \theta_{t,i}^f) = \frac{1}{\sqrt{2\pi R_t}} \exp(-\frac{1}{2R_t}[y_{t,i}^o - h(x_{t,i}^f, \theta_{t,i}^f)]^2) \tag{30}$$

$$p(y_t \mid x_{t,i}^f, \theta_{t,i}^f) = \frac{L(y_t \mid x_{t,i}^f, \theta_{t,i}^f)}{\sum_{i=1}^{ne} L(y_t \mid x_{t,i}^f, \theta_{t,i}^f)} = p(y_{t,i}^o - h(x_{t,i}^f, \theta_{t,i}^f) \mid R_t) \tag{31}$$

Then, the updated weights denoted as $w_{t,i}^a$ for each particle can be obtained as:

$$w_{t,i}^a = \frac{w_{t,i}^f \cdot p(y_{t,i}^o - h(x_{t,i}^f, \theta_{t,i}^f) \mid R_t)}{\sum_{i=1}^{ne} w_{t,i}^f \cdot p(y_{t,i}^o - h(x_{t,i}^f, \theta_{t,i}^f) \mid R_t)} \tag{32}$$

Based on the updated weights, those particles can be resampled to remove those samples with insignificant weights. A number of resample methods have been developed and the multinomial resampling method proposed by Moradkhani et al. (2005a) is used. Therefore, the resampled particles denoted as $\theta_{t-resamp,i}$ and $x_{t-resamp,i}$ can be obtained. The performance of the resampled particles is also evaluated by the mismatch index expressed as:

$$S(x_{t-resamp,i}, \theta_{t-resamp,i}) = \sum_{i=1}^{ne} (h(x_{t-resamp,i}, \theta_{t-resamp,i}) - y_{t,i}^o)^T R_t^{-1} (h(x_{t-resamp,i}, \theta_{t-resamp,i}) - y_{t,i}^o) \tag{33}$$

Step 8. Choose the posterior estimations for states and parameters by the following criteria:

If $S(x_{t+1-resamp,i}, \theta_{t+1-resamp,i}) \le S(x_{t+1,i}^a, \theta_{t+1,i}^a)$, $\theta_{t-resamp,i}, x_{t-resamp,i}$ would be the posterior estimations at current stage; otherwise, $x_{t+1,i}^a$, and $\theta_{t+1,i}^a$ would be the posterior estimations.

Step 9 Parameter perturbation: take parameter evolution to the next stage through add small stochastic error around the sample (take the EnKF estimation as an example):

$$\theta_{t+1,i}^f = \theta_{t,i}^a + \varepsilon_{t,i}, \quad \varepsilon_{t,i} \sim N(0, \eta S(\theta_{t,i}^a)) \tag{34}$$

where η is a hyper-parameter which determines the radius around each sample being explored; $S(\theta_{t,i}^{a})$ is the standard deviation of the analyzed ensemble values.

Step 10. Check the stopping criterion: if measurement data is still available in the next stage, $t = t + 1$ return to step 3; otherwise, stop.

Through PEnPF, the better estimations from EnKF and PF will be chosen as the posterior states and parameters, which may lead to improved predications for model states and simulated observations. Similar to CEnPF, the PEnPF can be applicable for nonlinear and non-Gaussian systems where once the estimates from EnKF are non-optimal, the estimates from PF will be adopted. Also, the ensembles will be adjusted through EnKF when the resulting predictions are consistently over or underestimates the respective observations.

3. Synthetic Experiments

3.1. Rainfall-Runoff Model

In this study, the Hymod, is adopted to test the efficiency of the CEnPF and PEnPF approaches. Hymod is a non-linear rainfall-runoff conceptual model which can be run in a minute/hour/daily time step (Moore, 1985). In Hymod, the soil moisture storage is characterized by a spatial probability distribution function and the runoff is routed to the catchment outlet by a fast linear-routing process (nominally event runoff) and a slow nonlinear routing process (nominally baseflow), as shown in Figure 3 (Moore, 2007). A cumulative distribution function (CDF) is proposed to describe such variability of soil moisture capacities, expressed as (Moore, 1985, 2007):

$$F(c) = 1 - \left[1 - \frac{c}{C_{\max}}\right]^{b_{\exp}}, 0 \le c \le C_{max} \qquad (35)$$

where C_{max} [L] is the maximum soil moisture capacity within the catchment and b_{exp} [-] is the degree of spatial variability of soil moisture capacities and affects the shape of the CDF. Five parameters are involved in Hymod for calibration based on observations: (i) the maximum storage capacity (C_{max}), (ii) spatial variability of soil moisture capacity (b_{exp}), (iii) the partitioning factor between the two series of reservoir tanks (α), (iv) the residence for the time quick-flow tank (R_q), and (v) the residence time for the slow-tank (R_s). Two inputs are required to force this model: precipitation, P (mm/day), and potential evapotranspiration, ET (mm/day).

Place Figure 3 Here

3.2. Synthetic Experiments

In this study, synthetic experiments are initially applied to test the applicability of the CEnPF and PEnPF approaches. The "true" observations are first defined when the model is run for a set of meteorological and initial conditions in the synthetic experiment (Moradkhani, 2008). The "true" model parameters are predefined before the synthetic experiment. The model inputs, including the potential evapotranspiration, ET (mm/day), and mean areal precipitation, P (mm/day), are generated based on onsite meteorological data, in which the mean areal precipitation data are generated based on the rain station measurements in the watershed, and the potential evapotranspiration values are interpolated based on data from national weather stations near the watershed.

Stochastic perturbations are required in a data assimilation framework to account for the uncertainties in model inputs, parameters and structures. In the synthetic experiments, random perturbations are added to precipitation and potential evapotranspiration (ET) observations to account for their uncertainties. For potential evapotranspiration, a Gaussian noise distribution is recommended by a number of researchers (e.g. DeChant and Moradkhani, 2012; Moradkhani et al., 2012; Chen et al., 2013; Rasmussen et al., 2015). For precipitation, some studies have applied Gaussian noise (e.g. Rasmussen et al., 2015), while other studies have concluded that log-normal noise may perform better (e.g. DeChant and Moradkhani, 2012; Moradkhani et al., 2012). In this study, the log-normal noise is adopted for the synthetic experiments, while Gaussian noises are employed for potential evapotranspiration, synthetic observations and model predictions. The proportionality factors are set to be 0.2 for all data in the synthetic experiments.

3.3. Evaluation Criteria

The root-mean-square error (RMSE), and the Nash-Sutcliffe efficiency (NSE) coefficient will be adopted to evaluate the performance of different data assimilation methods. These two indices also served as the responses in the multi-level factorial design to visualizing the effects of stochastic perturbations. The formations of RMSE and NSE are expressed as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Q_i - P_i)^2} \quad (36)$$

$$NSE = 1 - \frac{\sum_{i=1}^{N}(Q_i - P_i)^2}{\sum_{i=1}^{N}(Q_i - \overline{Q})^2} \qquad (37)$$

where N is the total number of observations (or predictions), Q_i are the observed values, P_i are the estimated values, and $\bar{Q}$ is the mean of all observed and estimated values.

Both RMSE and NSE merely measure the accuracy of the expected value and show the ability of each data assimilation technique to track the observations (Dechant et al., 2012). However, they are unable to evaluate the performance of predictive distribution from ensemble forecasts (Renard et al., 2010). Consequently, probabilistic measures are required to further provide a description of ensemble forecasts for different data assimilation schemes. In this study, the continuous ranked probability score (CRPS) and resolution (π) are used, which are formulated as follows (Murphy and Winkler, 1987; Hersbach, 2000; Madadgar et al., 2014):

$$CRPS = \int_{-\infty}^{+\infty}[F^f(x) - F^o(x)]^2 dx \qquad (38)$$

where where F^f and F^o are CDFs for forecasts and observations, respectively

$$\pi = \frac{1}{T}\sum_{t=1}^{T}\frac{E[y_{t,i}]}{\sigma[y_{t,i}]} \qquad (39)$$

where $E[y_{t,i}]$ is the expected value of ensemble predictions at time t and $\sigma[y_{t,i}]$ is the standard deviation of ensemble predictions at time t.

The CRPS is a measurement of error for probabilistic prediction. A small CRPS value indicates a better model performance, with the value of zero suggesting a perfect accuracy for model prediction. The index of resolution provides a description of precision of ensemble predictions with greater values suggesting larger uncertainty of forecasts (Madadgar et al., 2014)

3.4. Results Analysis

To demonstrate the capability of the proposed CEnPF and PEnPF approaches in parameters and state quantification for hydrologic models, synthetic experiments were performed with Hymod. Table 1 shows the "true" parameter set for the synthetic experiments. The initial ensembles for the five parameters (i.e. i.e. C_{max}, b_{exp}, α, R_s R_q) are sampled uniformly from predefined intervals as shown in Table 1. The initial ensembles for the state variable of storage are sampled from a normal distribution with a mean value of 0.05 and a standard deviation to be proportional to the mean value (the proportional factor is set to be 0.1). The initial samples for the slow flow tank are also sampled from a similar normal distribution with a mean value of 2.14. The initial samples for the three quick flow tanks are set to be 0, and the sample size used in the synthetic experiment was 200.

Figure 4 shows the comparison between the ensemble predictions and the synthetically generated true discharge values obtained from the EnKF, PF, CEnPF and PEnPF approaches. The results indicate that the ensemble means of streamflow predictions from the four methods can track well the observed discharge data. The ranges formulated by 5% and 95% percentiles (i.e. 90% confidence intervals) of streamflow predictions can adequately bracket the observations. In addition, ensemble predictions for two state variables, namely the storage and the flow in the slow tank of Hymod are plotted and compared with their true values in the experiment, as shown in Figure 4. The results show that, for all the four data assimilation schemes, the deterministic predictions (i.e. predictive means in this study) of state variables can well trace the fluctuation of their true values. Moreover, almost all the true values for the two state variables are located in the predictive intervals of the ensemble predictions of the four approaches.

Figure 5 describes the comparison of the convergence of each parameter from the EnKF, PF, CEnPF and PEnPF approaches. It is observed that identifiability of one parameter depends on the filtering approaches. For instance, all five parameters in Hymod are identifiable if the PF is employed, while in comparison the parameters of C_{max} and b_{exp} are unidentifiable for EnKF. For the two developed methods, CEnPF and PEnPF, the five parameters of Hymod can be well identified by CEnPF. Moreover, compared with PF, the proposed CEnPF can still rejuvenate ensembles in larger spaces than PF, which may lead to more reliable estimations for parameter posterior distributions. In comparison, parameter evolution patterns generated by PEnPF are similar with those from EnKF, which means that C_{max} and b_{exp} are unidentifiable in this data assimilation scheme. This is due to the mechanism of ensembles rejuvenation in PEnPF. In PEnPF, parameters and states are updated simultaneously by EnKF and PF, and the better estimations are shoes as the posterior distributions. If at each time step, EnKF performs better than PF, evolution characteristics of parameters and states would be identical to those generated by EnKF. The results in Figure 5 suggest that, parameter and state estimations from EnKF are chosen as the corresponding posteriors in the data assimilation experiment through PEnPF.

Moreover, to further explore the reliability of the four data assimilation approaches, five sample size scenarios (i.e. {20, 50, 100, 200, 500}) are tested. For each scenario, the synthetic experiment is performed for 30 replicates to identify the robustness of the proposed approaches. The performances of EnKF, PF, CEnPF and PEnPF are evaluated through two deterministic indices (i.e. RMSE and NSE) and two probabilistic indices (i.e. CRPS and

Resolution). Figure 6 compares the performance of EnKF, PF, CEnPF and PEnPF through a boxplot. The results indicate that all four methods will perform better with an increase in sample size, and the sample size influence PF more significantly than the other three data assimilation approaches. In detail, the PEnPF produce best deterministic predictions with lowest values for NSE and RMSE, followed by EnKF, CEnPF and PF. The performance of CEnPF is not as well as EnKF in this synthetic experiment. However, it performs better than PF. Especially when the same size is larger than 50, CEnPF would generate more reliable predictions than PF. For probabilistic predictions, the PEnPF would lead to lowest values for CRPS, indicating closest distance between the predictive and observed cumulative distribution functions (CDFs). Moreover, similar with deterministic predictions, the proposed CEnPF does not perform as well as EnKF in this synthetic experiment, but it provide more accurate predictions than PF, especially when the sample size is larger than 50.

4. Real Case Study

4.1. Site Description

Two real watersheds will be used test the applicability of the proposed data assimilation schemes, as presented in Figure 7. The first catchment is the Huanjiang river, located in the northern part of Jing river basin with a drainage area of 4,640 km^2. This catchment has two main tributaries, which converge together at Hongde (107.19 E, 36.76 N). In general, the Jing river basin is characterized by a semi-arid and sub-humid continental monsoon climate, resulting in significant temporal-spatial variations in precipitation. From the northern to southern part, the corresponding annual precipitation ranges from 240 to 710 mm, with approximately 50~60% precipitation occurring in the Summer and Fall seasons. In particular, the Huanjiang in this case is located in the northern part of the Jing River watershed, and the annual precipitation there fluctuates from 240 to 350 mm with mean annual precipitation of approximate 309 mm. For Huanjiang river, the daily precipitation data from Ganjipan, Fanxue, Shancheng, Wuqi, Gengwan, Honglaochi, Siheyuan and Hongde are employed to generate areal precipitation over the entire sub-catchment. The potential precipitation values were obtained through the Penman–Monteith equation, based on meteorological measurements from national meteorological stations (i.e. Changwu, Xifengzhen, Guyuan, Huanxian, Tongchuan) in the Jing river basin. Tables 2 and 3 provide the location information for the rain gauge stations and the national meteorological stations.

The second case is the Xiangxi river basin, located in the Three Gorges Reservoir area, China. The Xiangxi river is located between 30.96 ~ 31.67 ^{0}N and 110.47 ~ 111.13^{0}E in the Hubei part of the China Three Gorges Reservoir (TGR) region, with a draining area of approximately 3,200 km^2. The Xiangxi river originates in the Shennongjia Nature Reserve with a main stream length of 94 km and a catchment area of 3,099 km^2 and is one of the main tributaries of the Yangtze river (Han et al., 2014; Yang and Yang, 2014; Miao et al., 2014). The watershed experiences a northern subtropical climate. The annual precipitation is about 1,100 mm and ranges from 670 to 1,700 mm with considerable spatial and temporal variability (Xu et al., 2010; Zhang et al., 2014). The main rainfall season is from May through September, with a flooding season from July to August. The annual average temperature in this region is 15.6 ^{0}C and ranges from 12 ^{0}C to 20 ^{0}C. For this case, meteorological and streamflow data at Xingshan (31^{0}13'N, 110^{0}45'E) station will be used.

Place Figure 7 here and Tables 2 and 3 here

4.2. Results Analysis for Huanjiang river

In hydrologic sequential data assimilation, two issues are generally predefined before implementation of the sequential data assimilation. The first one is how many ensembles or particles are going to use to represent the distributional information in parameters, state variables and predictions. The other one is that how to account for uncertainty existing in forcing data, model prediction, and streamflow measurements. In the real case study, the sample size is set to be 200 for all the four data assimilation schemes based on the results of the synthetic experiment. Moreover, random perturbations are added to model inputs, outputs, and parameters to reflect their inherent uncertainties. In this study, the precipitation is assumed to follow a lognormal distribution with the proportional factors being 20% of the true, while the potential evapotranspiration, streamflow measurements, and model prediction are normally distributed with the standard errors being 20% of the true values.

Figure 8 shows the comparison between ensemble predictions of the four data assimilation methods and observations. Figure 8(a) indicates the comparison between the mean predictions and predictive intervals from EnKF and model and observations. The result shows that the predictive intervals from EnKF can generally bracket the observations during the low flow period, while underestimations occur during the high flow period. Similar characteristics can be found for both PF. However, as shown in Figure (8b), PF provide better predictions than EnKF. Especially for the high flow periods, the predictive intervals from PF can catch the peak flow better than those from EnKF. In comparison with EnKF and PF, the proposed CEnPF can generate more reliable predictions. As shown in Figure (8c), the predictive intervals from CEnPF can generally bracket the observations while the ensemble means can well track the fluctuation of real discharges for both low and high flow periods. For the PEnPF, it seems to perform slightly worse than CEnPF. In particular, the PEnPF would generate worse (i.e. underestimation) predictions than PF during the high flow periods. However, the PF would produce overestimations in a quite long period after the highest peak flow while PEnPF can provide accurate predictions in this period. In this case, the predictions from CEnPF lead to a NSE value of 0.911, a RMSE value of 5.897, a CRPS value of 2.209 and a Resolution value of 41.685. The four indices (i.e. NSE, RMSE, CRPS and Resolution) correspond to the predictions of PEnPF are 0.861, 7.372 , 1.675 and 15.058, respectively. The four indices for the predictions of EnKF are 0.767, 9.540, 2.234, and 21.697, and those indices for PF predictions are 0.776, 9.354, 4.026, and 38.596. Consequently, the CEnPF leads to best deterministic predictions while the PEnPF generates best probabilistic predictions

----

Place Figure 8 here

To further demonstrate the applicability of the proposed data assimilation methods, four sample scenarios (i.e. {50, 100, 200, 500}) are further tested for this real case with 10 replicates conducted for each sample scenario. Figure 9 compares the performance of EnKF, PF, CEnPF and PEnPF through a boxplot. It shows that as the increase of sample size, the proposed CEnPF, PEnPF as well as traditional EnKF would generate reliable predictions with the four evaluation indices varied within limited intervals. In comparison, the PF can also generate unsatisfactory results even the sample size of 500. Tables 4 to 7 provide the mean, minimum and maximum values for NSE, RMSE, CRPS and Resolution for the 10 replicates by different data assimilation schemes under different sample size scenarios. The results indicate that the proposed CEnPF can generally provide best results for deterministic predictions with lowest NSE and RMSE values. For instance, the CEnPF can lead to a mean NSE value of 0.78 under a sample size of 100, which is higher than the other three approaches (i.e. the mean NSE values would be 0.72, 0.69 and 0.65 for PEnPF, EnKF and PF). In comparison, the PEnPF would produce better probabilistic predictions than CEnPF, EnKF and PF, which generally has lowest CRPS and Resolution values, as presented in Tables 6 and 7. In general, even though the prediction from CEnPF has large degree of uncertainty (i.e. large Resolution values), the proposed CEnPF and PEnPF can provide better results for both deterministic and probabilistic forecasts for the Huanjiang river basin

----

Place Figure 9, Tables 4 to 7 here

4.3. Results Analysis for Xiangxi river

The developed data assimilation approaches are further applied for hydrological data assimilation in Xiangxi river, which is an main tributary of Yangtze river in Hubei Province. The Xiangxi river basin experiences a northern subtropical climate with higher temperature and precipitation than the Huanjiang river basin which has a semi-arid climate. To clearly account uncertainties in meteorological data and streamflow measurements in Xiangxi river, the proportional factor is set to be 30% of the true measurements. In current case, the sample size is 500.

Figure 10 shows the performance of the developed CEnPF and PEnPF as well as traditional EnKF and PF approaches for hydrological data assimilation in Xiangxi river. As presented in Figure (10a), the EnKF approach provide accurate deterministic and probabilistic predictions during the low flow periods, but these predictions cannot well track observations during high flow periods and show underestimated results in these periods. Compared with EnKF, the PF approach seems to provide better predictions, as shown in Figure (10b). Especially in high flow periods, PF performs better than EnKF, but it still provides underestimations in these time steps. In comparison, the developed CEnPF and PEnPF are able to generate reliable results for both deterministic predictions and the associated predictive intervals. As shown in Figures (10c) and (10), the predictive intervals of CEnPF and PEnPF can bracket the real observations at most time periods for this case. Meanwhile, the corresponding deterministic predictions (i.e. predictive means) can trace the variation in streamflow in both high and low flow periods.

Place Figure 10

Table 8 shows the performance of the four approaches for hydrological data assimilation in Xiangxi river basin under different sample size scenarios. The results shows that for deterministic predictions, the proposed CEnPF and PEnPF approach performs better than EnKF in all selected sample scenarios, and these two methods provide better deterministic predictions than PF in three of the four sample scenarios. However, in terms of the probabilistic forecasts, the performances of the fours approaches show different features. EnKF seems to lead to lowest CRPS values for all sample scenarios. However, at least one proposed approach (i.e. CEnPF or PEnPF) can provide better probabilistic predictions than PF for all selected sample scenarios.

Place Tables 8 here

----

5. Discussion

In this study, both CEnPF and PEnPF integrate traditional PF and EnKF into combined framework. This means that the computational demand would increase for CEnPF and

PEnPF since they have additional procedures. Figure 11 presents the computation demand for EnKF, PF, CEnPF and PEnPF. The results show that, among these four approaches, PF requires least computational time, and both CEnPF and PEnPF require more computational time than EnKF and PF since they have more steps. However, the computational time for the two developed methods is manageable. In detail, the PEnPF needs more computational requirement than the other three approaches. For instance, the computational time for PEnPF would be about 590 seconds when the sample size is 500, while the time for EnKF, PF and PEnPF would be 347, 102 and 443 seconds, respectively. This is because that, in spite of update procedures of EnKF and PF, the PEnPF needs two additional steps for putting the updated parameters from EnKF and PF into the original hydrological model to evaluate the mismatch between the resulting outputs and the real observations at each time step. This suggests that for some large hydrological models requiring much computation time, the PEnPF may need much more time than EnKF, PF and PEnPF since the hydrological model would be run for 3**ns* (*ns* is the sample size) times at each time while the other three approaches only need to run the hydrological model *ns* times.

Place Figure 11 here

----

6. Conclusions

This study proposed two integrated data assimilation schemes, i.e. the coupled EnKF and PF (CEnPF) and the parallelized EnKF and PF (PEnPF) approaches through the integration of the capabilities of EnKF and PF. The CEnPF sequentially adopts EnKF and PF to update model parameters and states, in which EnKF is first applied to correct model states and parameters, and PF is then employed to eliminate insignificant particles. In comparison, the PEnPF approach simultaneously updates model states and parameters in parallel through EnKF and PF, and chooses the better estimates as the posterior distributions. The proposed CEnPF and PEnPF approaches were applied for hydrologic data assimilation in two real-world cases to demonstrate their applicability in quantifying uncertainty in hydrologic prediction

A synthetic application firstly illustrated procedures of the proposed CEnPF and PEnPF approaches and compared them with traditional PF and EnKF methods. Five sample size scenarios were tested to evaluate the performance of the proposed methods. The results suggested that PEnPF performed best for both probabilistic and deterministic predictions, while CEnPF could provide better predictions than PF. The improvement of the proposed CEnPF and PEnPF upon EnKF and PF was further illustrated by two real-world catchments with different climate conditions. The results for the Huanjiang river, located in the northern part of Jing river, demonstrated that PEnPF would produce better probabilistic predictions than CEnPF, EnKF and PF, which generally has lowest CRPS and Resolution and the CEnPF could provide better results in deterministic predictions but lead to large uncertainty in its ensemble outputs. For the Xiangxi river located in the Yangtze river basin, the results indicated that the proposed approach improved EnKF and PF in terms of deterministic predictions. For all selected sample size scenarios, at least one method could give better probabilistic predictions than PF.

The ensemble Kalman filter (EnKF) and particle filter (PF) methods have been extensively applied for hydrologic data assimilation. However, both of them have their inherent disadvantages which restrict their application for many cases. In this study, two integrated sequential data assimilation approaches are proposed by integrating the capabilities of EnKF and PF into a general framework. The case studies for synthetic experiment and two real-world hydrologic data assimilation problems demonstrate the significant potential of the proposed CEnPF and PEnPF approaches. Moreover, the computational time for CEnPF and PEnPF is manageable when compared with EnKF and PF. However, the PEnPF may require much more computational time for large-scale or time-consuming hydrological models than EnKF, PF and CEnPF.

Acknowledgement

This work was jointly funded by the Natural Science Foundation of China (51520105013), the National Key Research and Development Plan (2016YFC0502800), and the Natural Sciences and Engineering Research Council of Canada.

References

Ajami N.K., Duan Q.Y., Sorooshian S., (2007). An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research*, 43, W01403.

Brown, J. D. (2010). Prospects for the open treatment of uncertainty in environmental research, Prog. Phys. Geog., 34, 75–100, doi:10.1177/0309133309357000.

Chen, H., Yang, D., Hong, Y., Gourley, J.J., Zhang, Y., 2013. Hydrological data assimilation with the Ensemble Square-Root-Filter: use of streamflow observations to update model states for real-time flash flood forecasting. Advance in Water Resources 59, 209-220.

Chen Y., Oliver D.S., (2013). Levenberg–Marquardt forms of the iterative ensemble smoother for efficient history matching and uncertainty quantification. *Computational Geosciences* 17(4), 689-703.

Clark, M. P., Rupp, D. E., Woods, R. A., Zheng, X., Ibbitt, R. P., Slater, A. G., Schmidt, J., and Uddstrom, M. J., (2008). Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model, Adv. Water Resour., 31, 1309–1324, doi:10.1016/j.advwatres.2008.06.005.

DeChant C.M., Moradkhani H., (2012). Examining the effectiveness and robustness of sequential data assimilation methods for quantification of uncertainty in hydrologic forecasting. Water Resources Research, 48, W04518, doi:10.1029/2011WR011011

DeChant C.M., and H. Moradkhani (2014), Toward a Reliable Prediction of Seasonal Forecast Uncertainty: Addressing Model and Initial Condition Uncertainty with Ensemble Data Assimilation and Sequential Bayesian Combination, Journal of Hydrology , 519, 2967-2977, doi: 10.1016/j.jhydrol.2014.05.045.

De Lannoy, G. J. M., Reichle, R. H., Houser, P. R., Pauwels, V. R., and Verhoest, N. E., (2007). Correcting for forecast bias in soil moisture assimilation with the ensemble Kalman filter, Water Resour. Res., 43, W09410, doi:10.1029/2006WR005449.

Doucet, A., N. De Freitas, and N. Gordon (2001), Sequential Monte Carlo Methods in Practice, vol. 1, Springer, N. Y.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. Journal of Geophysical Research: Oceans, 99(C5), 10143-10162.

Evensen, G. (2003), The Ensemble Kalman Filter: theoretical formulation and practical implementation, Ocean Dynamics 53, 343–367.

Fan Y.R., Huang G.H., Baetz B.W., Li Y.P., Huang K., Li Z., Chen X., Xiong L.H., (2016). Parameter uncertainty and temporal dynamics of sensitivity for hydrologic models: A hybrid sequential data assimilation and probabilistic collocation method. *Environmental Modelling & Software* 86, 30-49

Fan Y.R., Huang G.H., Huang K., Baetz B.W., (2015a). Planning Water Resources Allocation under Multiple Uncertainties through A Generalized Fuzzy Two-Stage Stochastic Programming Method. IEEE Transactions on Fuzzy Systems, 23(5), 1488-1504.

Fan Y.R., Huang W.W., Li Y.P., Huang G.H., Huang K., Li Y.P., (2015b). A coupled ensemble filtering and probabilistic collocation approach for uncertainty quantification of hydrological models. Journal of Hydrology, 530, 255-272.

Fan Y.R., Huang W.W., Huang G.H., Huang K., Zhou X., (2015c). A PCM-based stochastic hydrological model for uncertainty quantification in watershed systems. Stochastic Environmental Research and Risk Assessment, 29, 915-927

Frei, M., and H. R. Kunsch (2013), Bridging the ensemble Kalman and particle filters, Biometrika, 100(4), 781–800

Fearnhead, P. and Clifford, P.: On-line inference for hidden Markov models via particle filters, J. R. Stat. Soc. B Met., 65, 887–899, 2003.

Gordon, N.J., Salmond, D.J., Smith, A.F.M., 1993. Novel approach to nonlinear/nonGaussian Bayesian state estimation. IEEE Proceedings F: Radar Signal Process. 140 (2), 107e113.

Gu Y., Oliver D.S., (2007). An iterative ensemble Kalman filter for multiphase fluid flow data assimilation. SPE Journal, 12(4), 438–446

Han X.., Li X., (2008). An evaluation of the nonlinear/non-Gaussian filters for the sequential data assimilation. Remote Sensing of Environment, 112, 1434-1449.

Kavetski, D., Kuczera, G., and Franks, S. W. (2006a). Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, Water Resour. Res., 42, W03407, doi:10.1029/2005WR00436.

Kavetski, D., Kuczera, G., and Franks, S. W., (2006b) Bayesian analysis of input uncertainty in hydrological modeling: 2. Application,Water Resour. Res., 42, W03408, doi:10.1029/2005WR004376.

Kong X.M., Huang G.H., Fan Y.R., Li Y.P., (2015). Maximum entropy-Gumbel-Hougaard copula method for simulation of monthly streamflow in Xiangxi river, China. Stochastic Environmental Research and Risk Assessment 29, 833-846.

Leisenring, M., & Moradkhani, H. (2012). Analyzing the uncertainty of suspended sediment load prediction using sequential data assimilation. Journal of Hydrology, 468, 268-282

Li T., Bolic M., Djuric P.M., (2015) Resampling Methods for Particle Filtering: Classification, implementation, and strategies. IEEE Signal Processing Magazine, 32(3), 70-86

Li Z., Huang G.H., Fan Y.R., Xu J.L., Hydrologic Risk Analysis for Nonstationary Streamflow Records under Uncertainty. Journal of Environmental Informatics 26 (1), 41-51.

Liu, J. S. and Chen, R.: Sequential Monte Carlo methods for dynamic systems, J. Am. Stat. Assoc., 93, 1032–1044, 1998.

Liu Y., Weerts A.H., Clark M., Hendricks Franssen H.-J., Kumar S., Moradkhani H., Seo D.-J., Schwanenberg D., Smith P., van Dijk A.I.J.M., van Velzen N., He M., Lee H., Noh S.J., Rakovec O., Restrepo P., (2012). Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities. Hydrology and Earth System Sciences, 16, 3863-3887.

Madadgar, S. and H. Moradkhani (2014). Improved Bayesian Multi-modeling: Integration of Copulas and Bayesian Model Averaging. Water Resources Research, 50, 9586-9603, doi: 10.1002/2014WR015965.

Montanari A., Brath A., (2004), A stochastic approach for assessing the uncertainty of rainfall-runoff simulations. Water Resources Research, 40, W01106

Moor, R.J., 1985. The probability-distributed principle and runoff production at point and basin scales. Hydrological Science Journal 30, 273-297.

Moor, R.J., 2007. The PDM rainfall-runoff model. Hydrology and Earth Systems Science 11 (1), 483-499.

Moradkhani, H. (2008). Hydrologic remote sensing and land surface data assimilation. Sensors, 8(5), 2986-3004.

Moradkhani, H., S. Sorooshian, H. V. Gupta, and P. Houser (2005a), Dual state – parameter estimation of hydrologic models using ensemble Kalman filter. Advances in Water Resources, 28, 135 – 147.

Moradkhani H., Dechant C.M., Sorooshian S., (2012). Evolution of ensemble data assimilation for uncertainty quantification using the particle filter-Markov chain Monte Carlo method, Water Resources Research, 48, W12520, doi:10.1029/2012WR012144.

Moradkhani, H., Hsu, K. L., Gupta, H., & Sorooshian, S. (2005b). Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter. Water Resources Research, 41(5)

Pappenberger, F. and Beven, K. J., (2006). Ignorance is bliss: Or seven reasons not to use uncertainty analysis, Water Resour. Res., 42, W05302, doi:10.1029/2005WR004820, 2006.

Parrish, M., H. Moradkhani, and C.M. DeChant (2012). Towards Reduction of Model Uncertainty: Integration of Bayesian Model Averaging and Data Assimilation, Water Resources Research, 48, W03519, doi:10.1029/2011WR011116.

Pathiraja, S., L. Marshall, A. Sharma, and H. Moradkhani (2016), Detecting non-stationary hydrologic model parameters in a paired catchment system using Data Assimilation, Advances in Water Resources, 94, 103-119, doi:10.1016/j.advwatres.2016.04.021.

Pathiraja, S., L. Marshall, A. Sharma, and H. Moradkhani (2016), Hydrologic Modeling in Dynamic catchments: A Data Assimilation Approach, Water Resources Research, doi: 10.1002/2015WR017192

Plaza-Guingla D. A., De Keyser R., De Lannoy G. J. M., Giustarini L., Matgen P., and Pauwels V. R. N., (2013). Improving particle filters in rainfall-runoff models: Application of the resample-move step and the ensemble Gaussian particle filter, *Water Resource Research*, 49, doi:10.1002/wrcr.20291.

Rasmussen J., Madsen H., Jensen K.H., Refsgaard J.C., (2015). Data assimilation in integrated hydrological modeling using ensemble Kalman filtering: evaluating the effect of ensemble size and localization on filter performance. Hydrology and Earth System Sciences, 19, 2999-3013.

Reichle R., Mclaughlin D., Entekhabi D., (2002). Hydrologic data assimilation with the ensemble Kalman filter. Monthly Weather Review, 130(1), 103-114.

Rezaie, J. and Eidsvik, J. (2012). Shrinked $(1 - \alpha)$ ensemble Kalman filter and α Gaussian mixture filter. Computational Geosciences, 16:837–852.

Schaake, J., Franz, K., Bradley, A., and Buizza, R., (2006). The Hydrologic Ensemble Prediction EXperiment (HEPEX), Hydrol. Earth Syst. Sci. Discuss., 3, 3321–3332, doi:10.5194/hessd-3-3321-2006.

Salamon, P. and Feyen, L. (2010). Disentangling uncertainties in distributed hydrological modeling using multiplicative error models and sequential data assimilation, Water Resour. Res., 46, W12501, doi:10.1029/2009WR009022.

Shen Z., and Tang Y., (2015). A modified ensemble Kalman particle filter for non-Gaussian systems with nonlinear measurement functions. Journal of Advances in Modeling Earth Systems, 07, doi:10.1002/2014MS000373.

Shi, Y., K. J. Davis, F. Zhang, C. J. Duffy, and X. Yu (2014), Parameter estimation of a physically based land surface hydrologic model using the ensemble Kalman filter: A synthetic experiment. Water Resources Research, 50, 1-19, doi:10.1002/2013WR014070

Snyder, C., Bengtsson, T, Bickel, P., and Anderson, J.: Obstacles to high-dimensional particle filtering, Mon. Weather Rev., 136, 4629–4640, 2008.

Su, C.H., Ryu, D., Crow, W.T., Western, A.W., 2014. Beyond triple collocation: applications to soil moisture monitoring. J. Geophys. Res. Atmos. 119, 6419–6439. http://dx.doi.org/10.1002/2013JD021043.

Vrugt, Jasper A., ter Braak, Cajo J.F., Diks Cees G.H., Schoups, Gerrit, (2013). Hydrologic data assimilation using particle Markov chain Monte Carlo simulation: Theory, concepts and applications. Advances in Water Resources, 51, 457-478.

Vrugt J.A., Diks C.G.H., Gupta H.V., Bouten W., Verstraten J.M., (2005). Improved treatment of uncertainty in hydrologic modelling: Combining the strengths of global optimization and data assimilation. Water Resources Research, 41, W01017.

Weerts A.H., EI Serafy G.Y.H., (2006). Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models. Water Resources Research, 42, W09403. doi:10.1029/2005WR004093

Xie X., Zhang D., (2013). A partitioned update scheme for state-parameter estimation of distributed hydrologic models based on the ensemble Kalman filter. Water Resources Research, 49, 7530-7365

Yan, H., DeChant, C.M., Moradkhani, H., 2015. Improving soil moisture profile prediction with the particle filter-Markov chain Monte Carlo method. IEEE Transactions on Geoscience and Remote Sensing. 53, 6134–6147. http://dx.doi.org/10.1109/TGRS.2015.2432067.

Yan, H., C.M., Moradkhani, H., 2016. Combined assimilation of streamflow and satellite soil moisture with the particle filter and geostatistical modeling. Advances in Water Resources, 94, 364–378

Zhang Y., Oliver D.S., Chen Y., Skaug H.J., (2014). Data Assimilation by Use of the Iterative Ensemble Smoother for 2D Facies Models, SPE Journal, 20(1), 169–185.

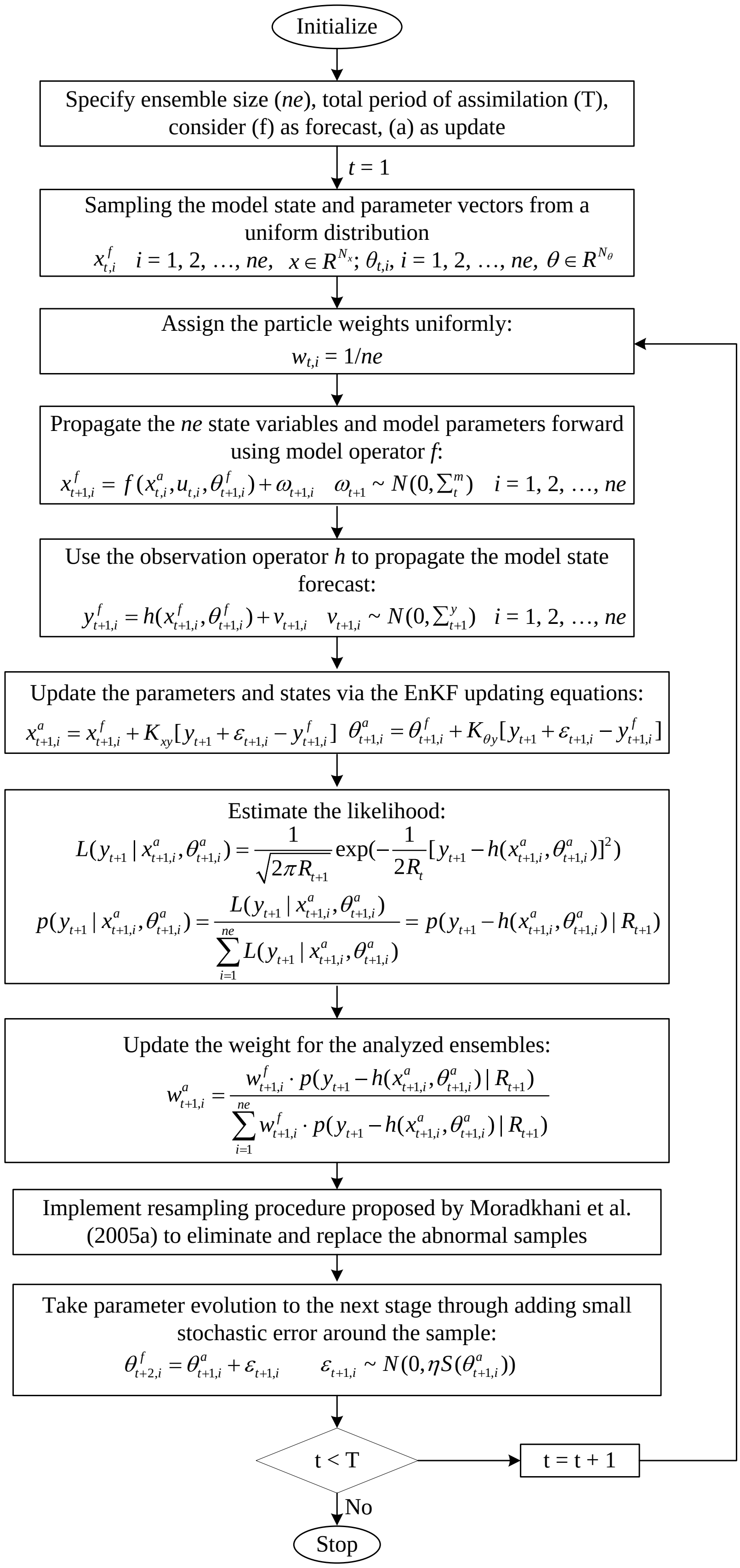
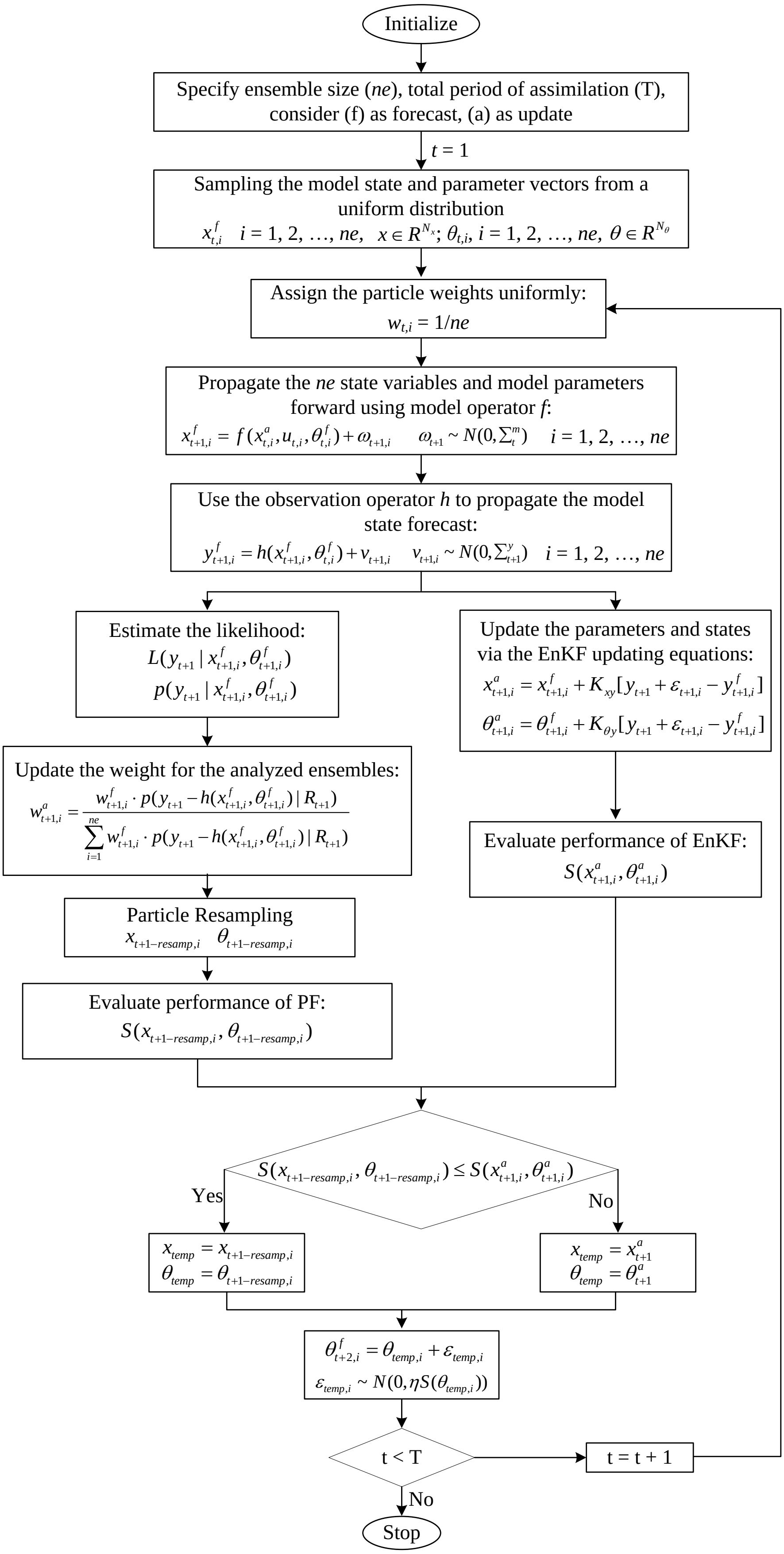


Figure 1. The flow chart of CEnPF

892



893 Figure 2. The flow chart of PEnPF

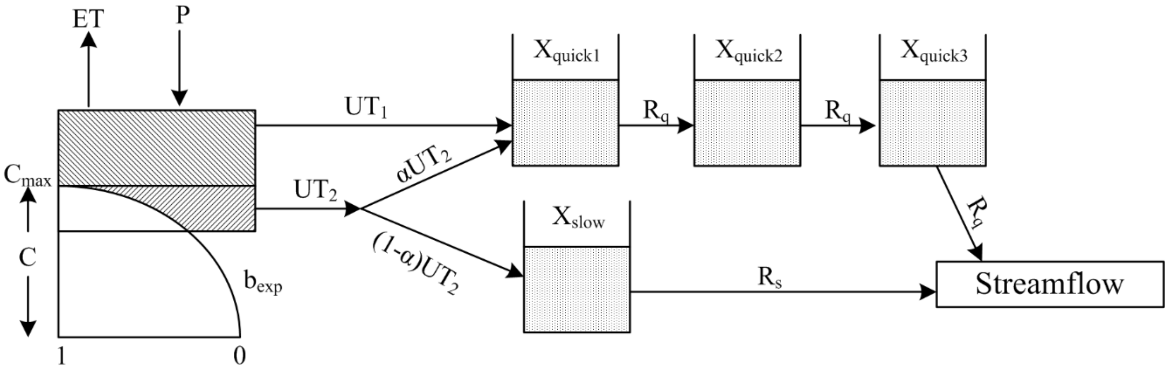


Figure 3 Description of Hymod

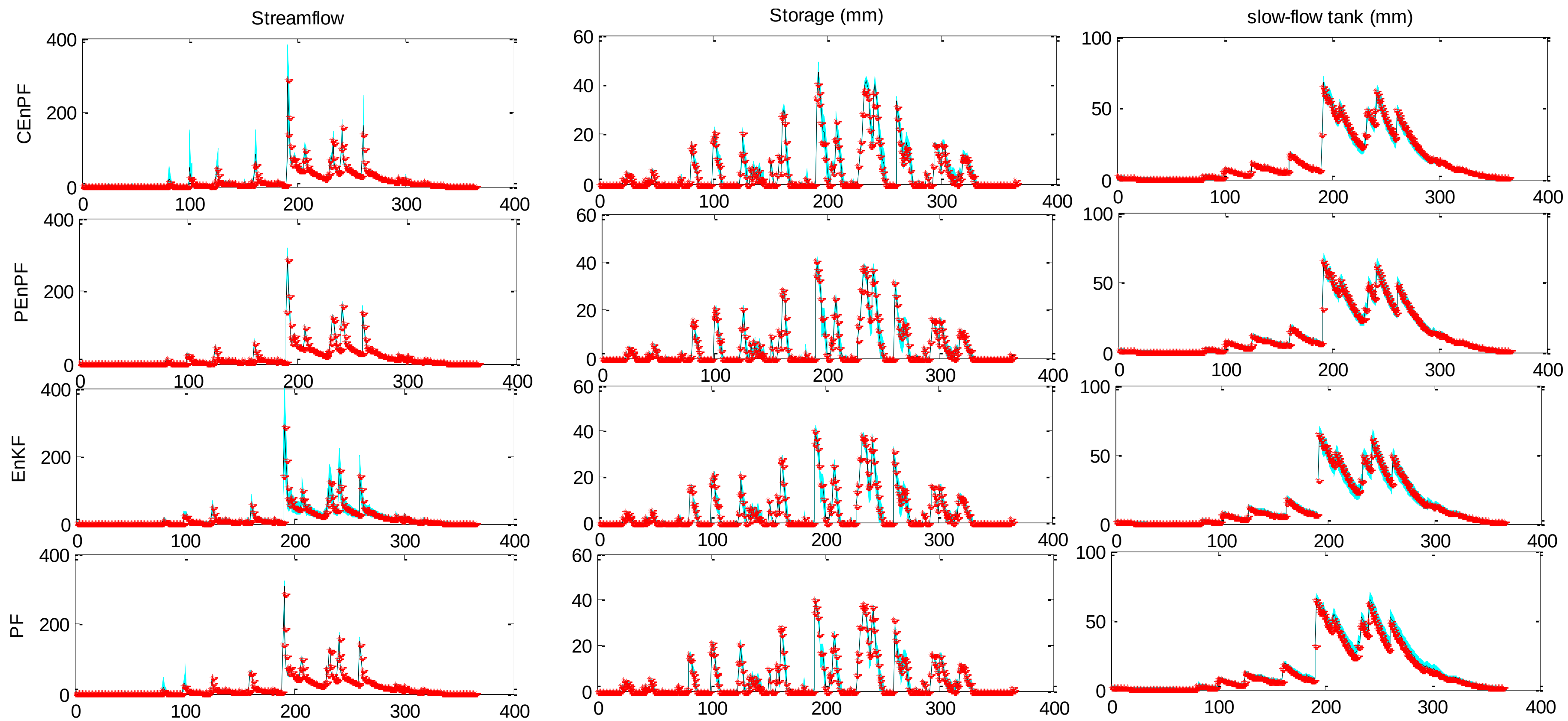


Figure 4: Comparison between ensemble predictions and synthetically generated true discharge: Four methods are used including EnKF, PF, CEnPF and PEnPF. The cyan region indicates the 90% predictive intervals, the red stars denote the synthetic observations, and the black line indicates the predictive mean values.

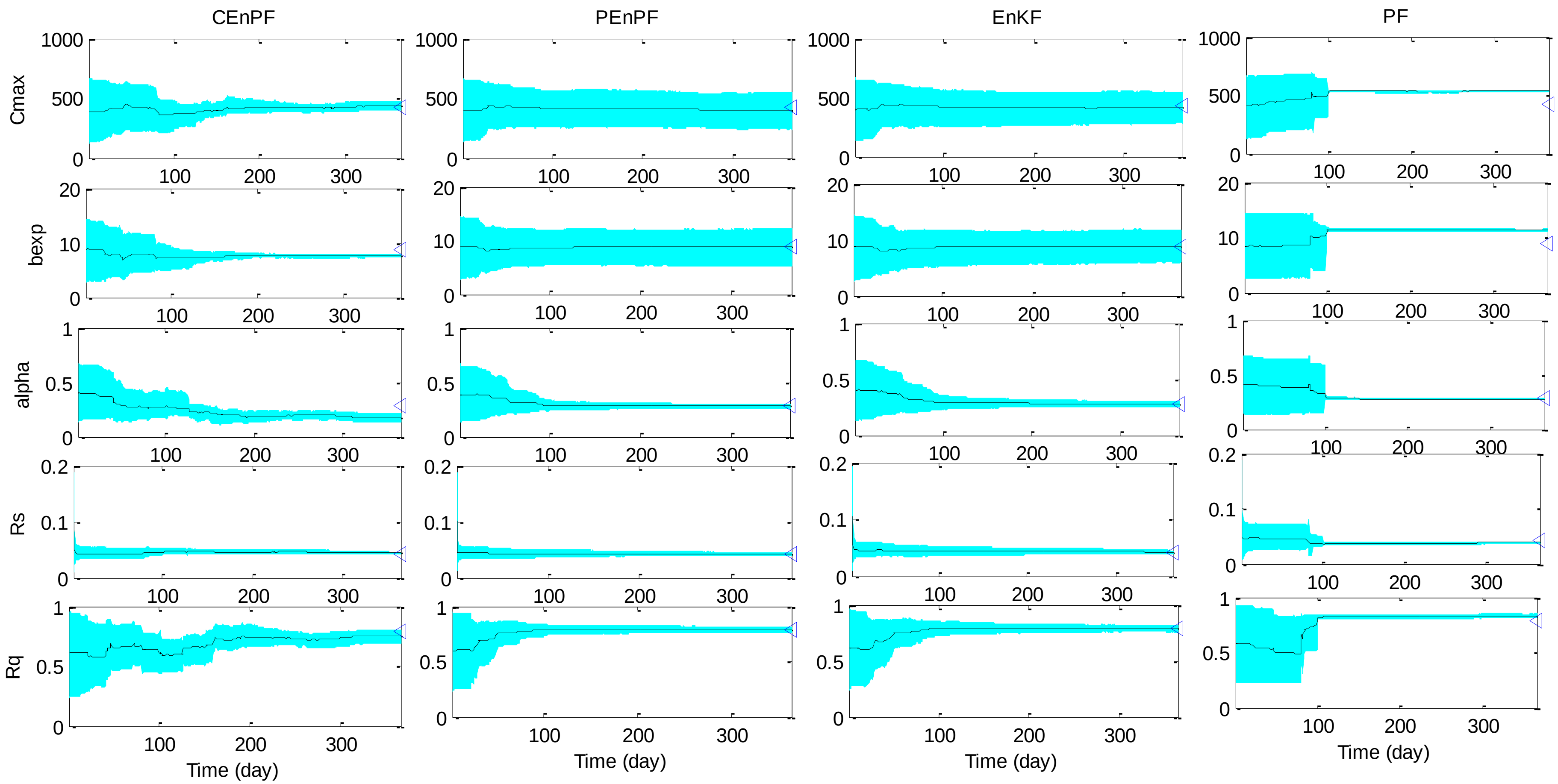


Figure 5: Convergence of the parameter distributions for the EnKF, PF, CEnPF and PEnPF for the synthetic experiments: The cyan region indicates the 90% intervals, the black line denotes the mean values, and the triangle is the predefined parameter value.

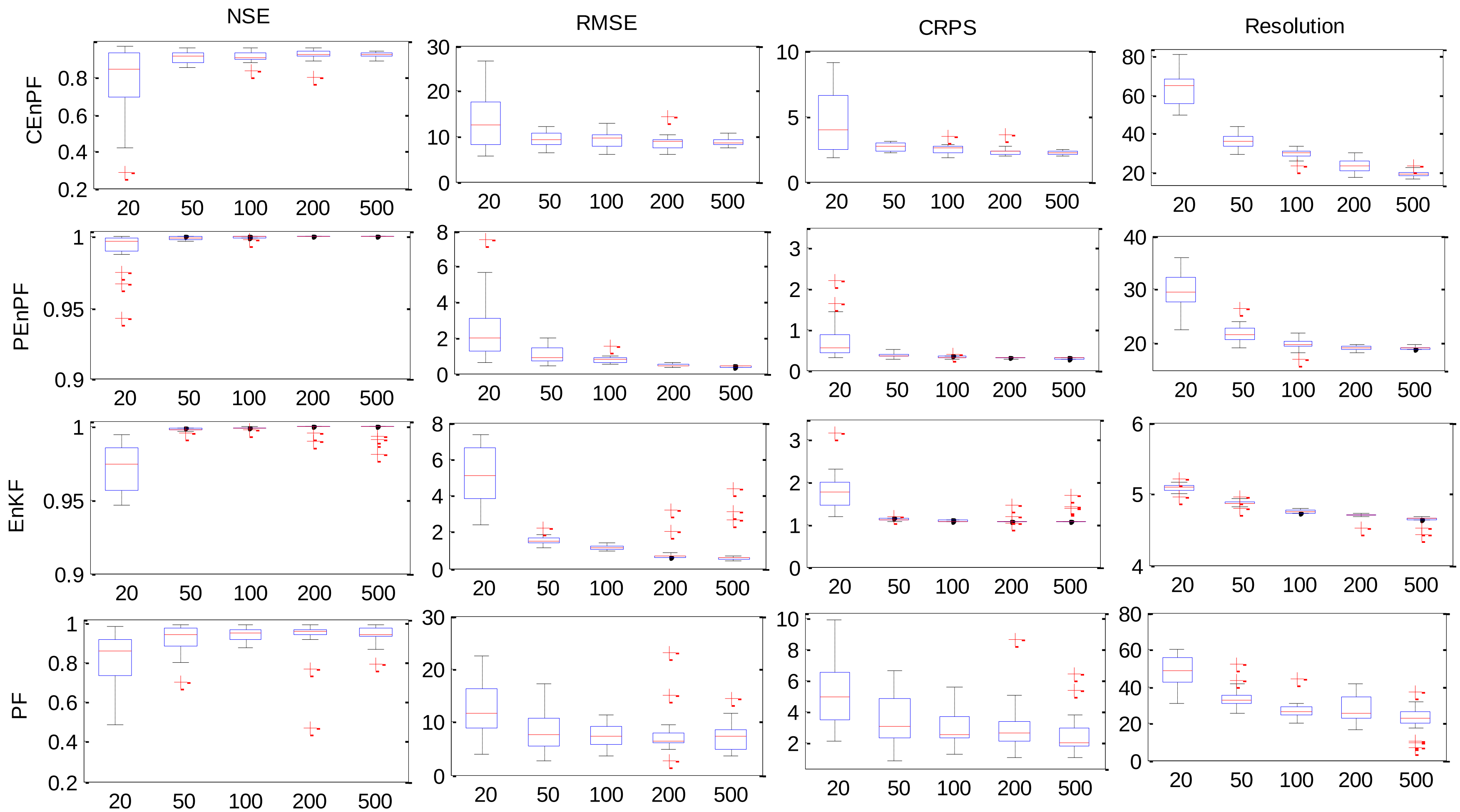


Figure 6. Performance comparison among EnKF, PF, CEnPF and PEnPF through a boxplot: The results show that all four methods will perform better with an increase in sample size. Generally, the PEnPF performs best than the other in both deterministic and probabilistic predictions, followed by EnKF, CEnPF and PF, if they are evaluated through NSE, RMSE and CRPS. However, the EnKF produces predictions with a lower resolution thn PEnPF.

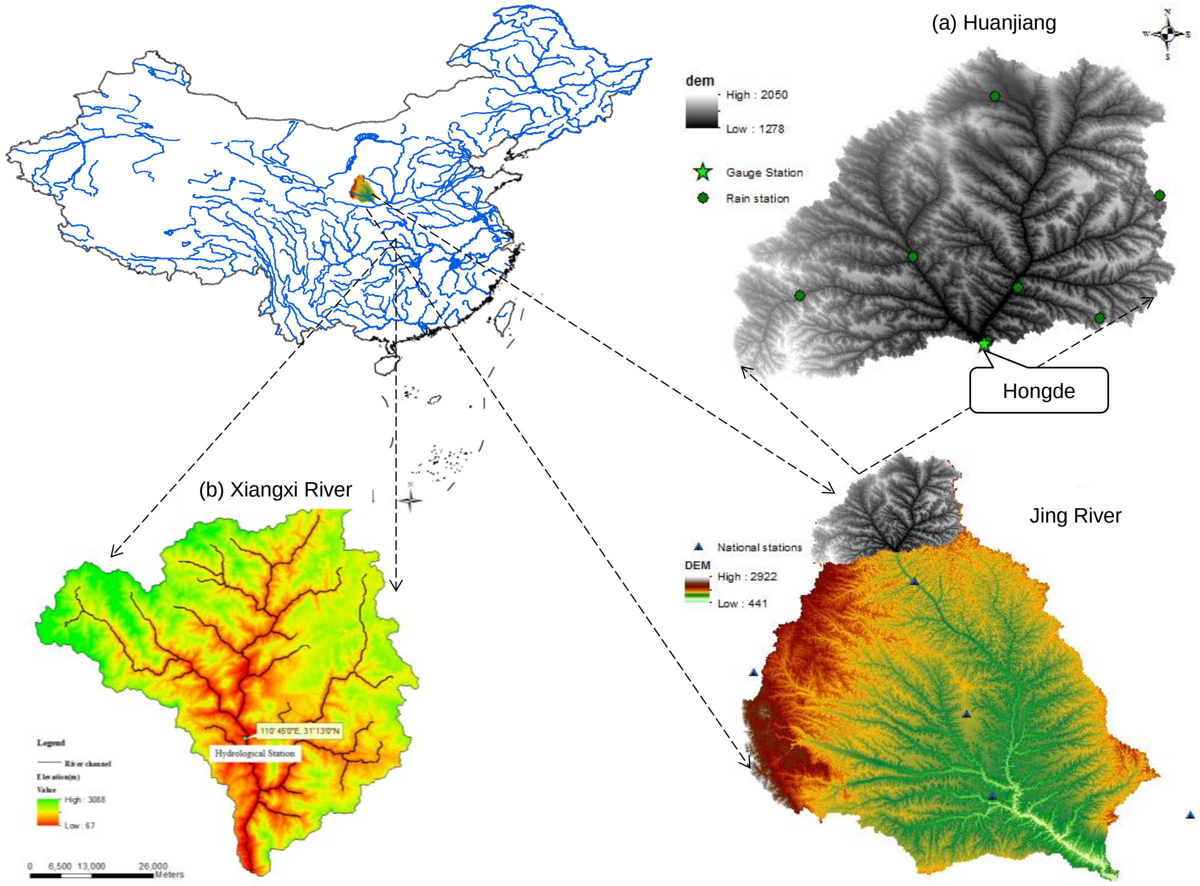


Figure 7. The location of the studied watersheds. Two watersheds are used to demonstrate the applicability of the proposed data assimilation schemes. One watershed named Huanjiang, located in the the north part of Jing River. Precipitation data from the seven rain stations in this catchment are used to generate the areal precipitation in the studied sub-catchment. The potential evapotranspiration (PE) are interpolated based on the PE results at the five national meteorological stations. The streamflow observations at Hongde station are used to evaluate the performance of the proposed methods. For the Xiangxi river watershed, meteorological and streamflow observations at Xingshan ($31^0$13'N, $110^0$45'E) station will be used.

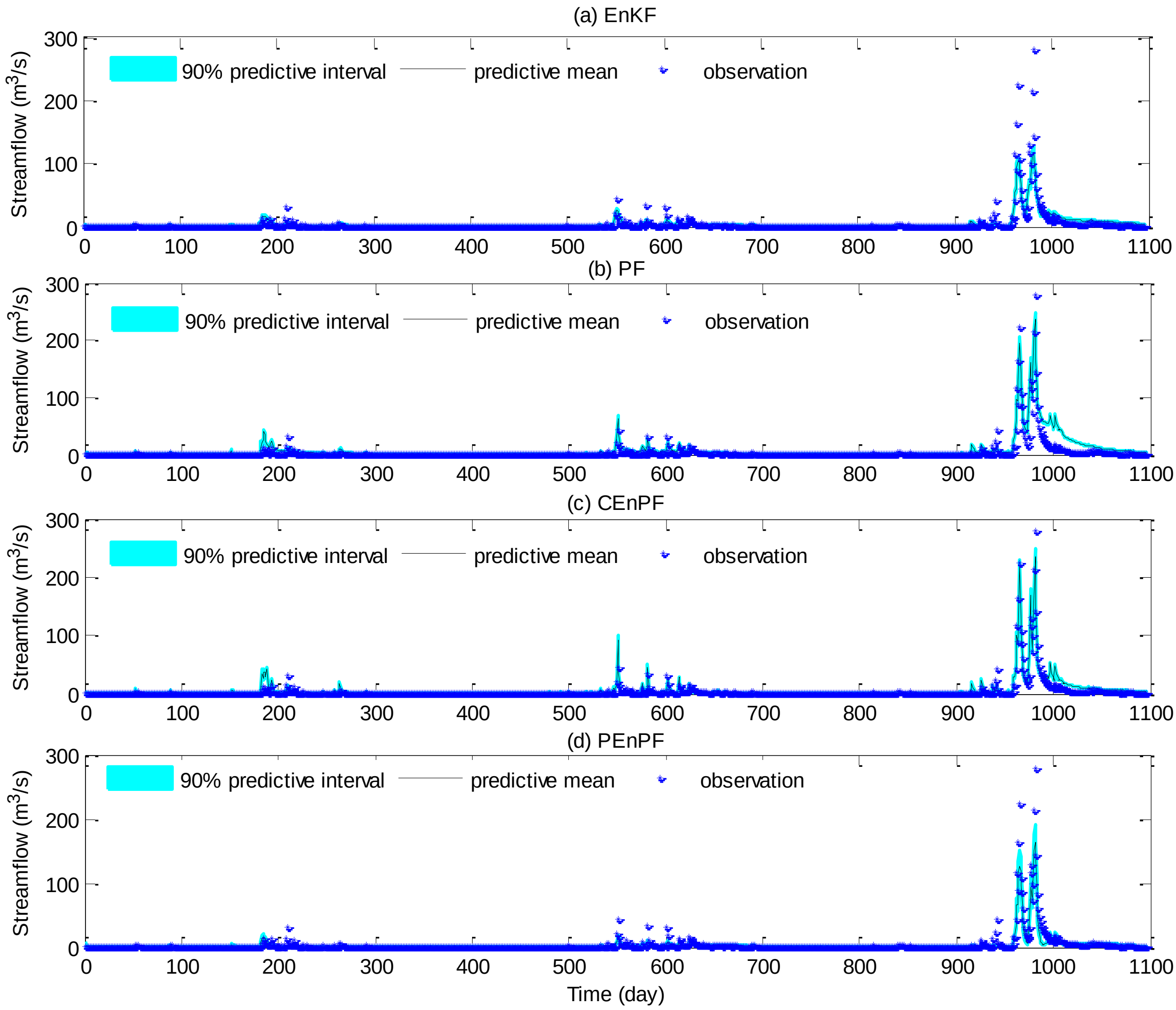


Figure 8. Comparison between the predication intervals and observations for Huanjiang river through different data assimilation schemes: (a) EnKF, (b) PF, (c) CEnPF, (d) PEnPF.

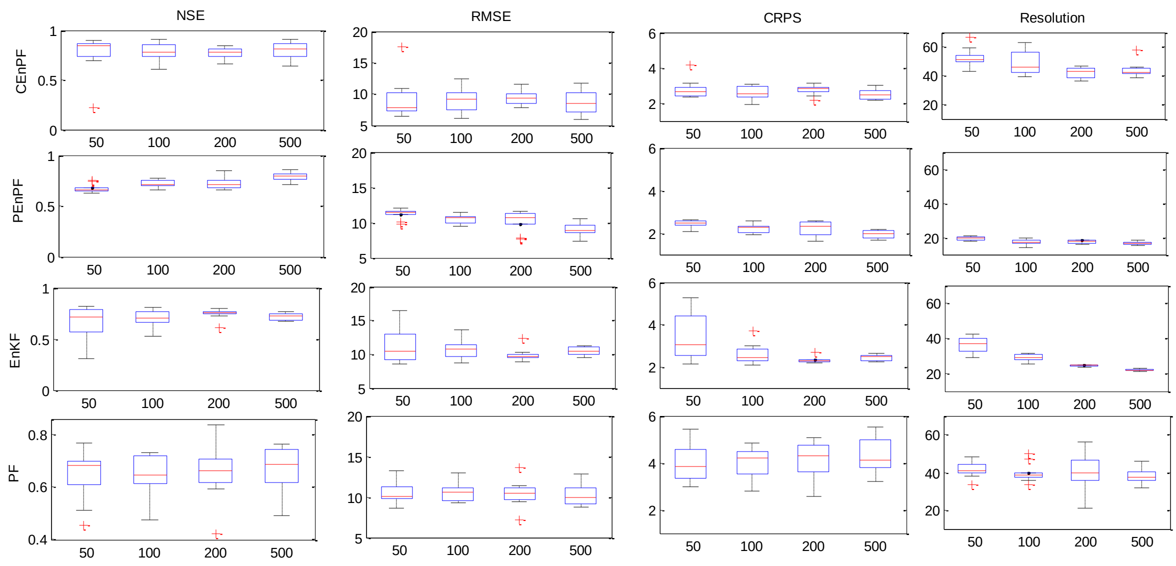


Figure 9. Performance comparison among different data assimilation schemes by using NSE, RMSE, CRPS and Resolution

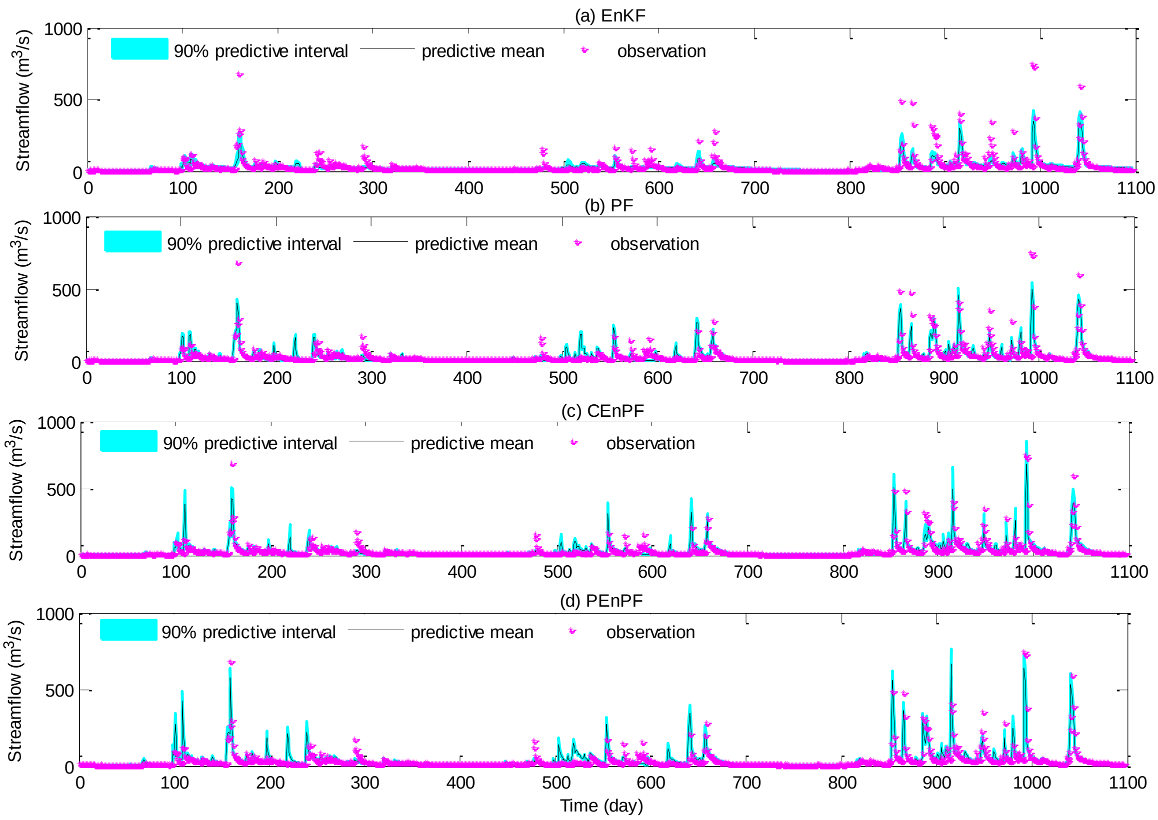


Figure 10. Comparison between the predication intervals and observations for Xiangxi river through different data assimilation schemes: (a) EnKF, (b) PF, (c) CEnPF, (d) PEnPF.

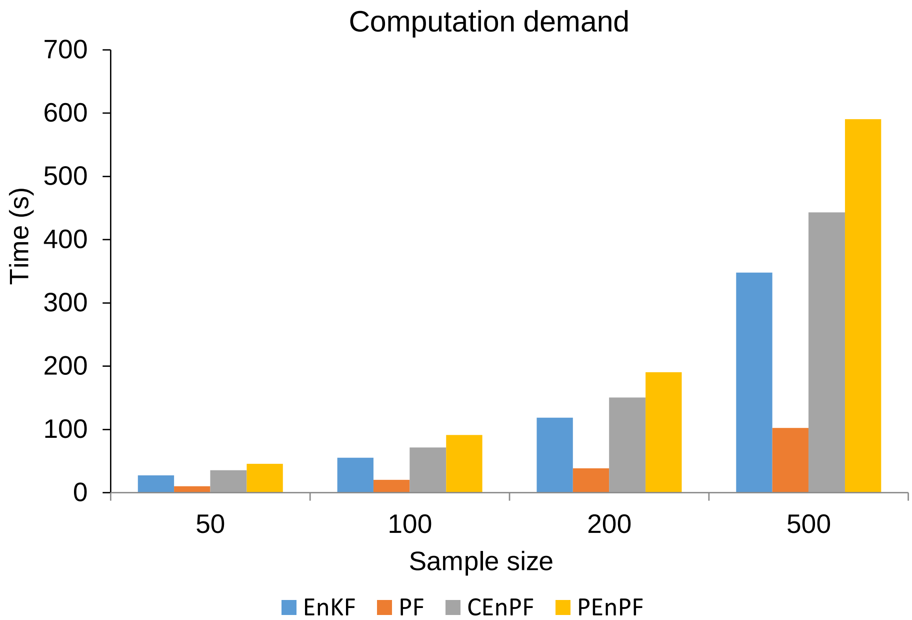


Figure 11. Computation demand for EnKF, PF, CEnPF and PEnPF under different sample size scenarios

Table 1. The predefined true values (used in synthetic experiment), initial fluctuating ranges of Hymod parameters

Description	Parameter	Range	Synthetic true value
Maximum storage capacity of watershed	C_{max} (mm)	[100, 700]	428.18
Spatial variability of soil moisture capacity	b_{exp}	[2, 15]	8.79
Factor distributing flow to the quick-flow tank	α	[0.10, 0.70]	0.28
Residence time of the slow-flow tank	R_s (1/day)	[0.001, 0.20]	0.042
Residence time of the quick-flow tank	R_q (1/day)	[0.2, 0.99]	0.79

Table 2. the location of rain gauge stations in Huanjiang river basin

Name	Longitude	Latitude
Ganjipan	107.22	37.30
Fanxue	107.58	37.08
Shancheng	107.03	36.95
Gengwan	107.27	36.88
Honglaochi	106.78	36.87
Siheyuan	107.45	36.82
Hongde	107.20	36.77

Table 3 Locations of National meteorological stations in Jing river basin

Name	Longitude	Latiude
Changwu	107.80	35.20
Xifengzhen	107.63	35.73
Guyuan	106.27	36.00
Huanxian	107.30	36.58
Tongchuan	109.07	35.08

Table 4. The NSE coefficient between the ensemble predictions and real observations in Huanjiang river.

		50	100	200	500
CEnPF	Mean	0.7548	0.7803	0.7736	0.8007
	Min	0.2174	0.6047	0.6620	0.6429
	Max	0.8943	0.9044	0.8464	0.9109
PEnPF	Mean	0.6739	0.7175	0.7294	0.7899
	Min	0.6249	0.6613	0.6563	0.7137
	Max	0.7555	0.7702	0.8471	0.8607
EnKF	Mean	0.6532	0.6907	0.7448	0.7181
	Min	0.3035	0.5223	0.6134	0.6738
	Max	0.8140	0.8056	0.7977	0.7667
PF	Mean	0.6470	0.6458	0.6509	0.6660
	Min	0.4521	0.4721	0.4176	0.4885
	Max	0.7656	0.7318	0.8383	0.7633

Table 5. The RMSE values between the ensemble predictions and real observations in Huanjiang river.

		50	100	200	500
CEnPF	Mean	9.2789	9.0914	9.3338	8.6391
	Min	6.4205	6.1079	7.7408	5.8972
	Max	17.4726	12.4186	11.4827	11.8033
PEnPF	Mean	11.2552	10.4769	10.1872	9.0089
	Min	9.7672	9.4682	7.7224	7.3720
	Max	12.0960	11.4950	11.5790	10.5680
EnKF	Mean	11.3404	10.8787	9.9398	10.4714
	Min	8.5184	8.7083	8.8827	9.5404
	Max	16.4840	13.6516	12.2815	11.2803
PF	Mean	10.5186	10.5716	10.4382	10.2374
	Min	8.6479	9.2499	7.1836	8.6903
	Max	13.2215	12.9784	13.6322	12.7747

Table 6. The CRPS values between the ensemble predictions and real observations in Huanjiang river.

		50	100	200	500
CEnPF	Mean	2.7980	2.5831	2.7709	2.5238
	Min	2.3589	1.9576	2.1644	2.1624
	Max	4.1678	3.0720	3.1563	3.0222
PEnPF	Mean	2.4414	2.2300	2.2268	1.9614
	Min	2.0791	1.9265	1.6249	1.6750
	Max	2.6434	2.5651	2.5963	2.1885
EnKF	Mean	3.3559	2.5764	2.3244	2.4289
	Min	2.1443	2.0683	2.2054	2.2345
	Max	5.2723	3.7094	2.7044	2.6382
PF	Mean	3.9765	4.0262	4.1305	4.2854
	Min	2.9877	2.7904	2.5652	3.2007
	Max	5.4238	4.8530	5.0780	5.5043

Table 7. The Resolution between the ensemble predictions and real observations in Huanjiang river.

		50	100	200	500
CEnPF	Mean	52.4690	48.8849	42.4754	43.7232
	Min	43.2976	39.0868	36.1500	38.7363
	Max	66.7200	62.8025	46.6733	57.6743
PEnPF	Mean	19.4104	17.2911	17.6186	16.6493
	Min	17.5940	14.0080	16.0280	15.0580
	Max	20.9610	19.4370	18.6260	18.6290
EnKF	Mean	35.9948	29.0739	24.6598	21.9759
	Min	28.9328	25.4233	23.6961	21.0699
	Max	42.5571	31.6062	25.1039	22.7798
PF	Mean	41.5654	39.6750	39.8738	38.5949
	Min	33.4221	33.5924	21.0764	31.9602
	Max	48.1742	49.7531	55.9405	45.8325

Table 8. Comparison of different data assimilation approaches at Xingxi River

		NSE	RMSE	CRPS	Resolution
50	EnKF	0.5553	43.9565	15.2674	23.5072
	PF	0.6837	36.4071	19.0750	32.4610
	CEnPF	0.6951	36.3942	18.4432	39.8297
	PEnPF	0.7294	33.6750	21.2260	24.2767
100	EnKF	0.6014	41.6133	14.1384	21.8007
	PF	0.7338	34.0062	18.5035	23.0801
	CEnPF	0.7127	35.3301	17.1706	24.2102
	PEnPF	0.7166	35.0884	21.0474	12.9162
200	EnKF	0.6110	41.1089	13.8818	20.8912
	PF	0.7163	34.4767	19.5430	19.4740
	CEnPF	0.6725	37.7190	17.6068	21.2002
	PEnPF	0.7465	33.1868	16.8556	21.7079
500	EnKF	0.5231	45.5183	14.8714	22.2468
	PF	0.6786	36.6998	18.6901	22.2949
	CEnPF	0.7530	32.7555	15.8585	20.2561
	PEnPF	0.7403	32.9869	15.7859	24.3501