# RESEARCH ARTICLE

**Key Points:**
- A copula-based particle filter approach is developed for hydrological data assimilation
- Synthetic experiments and real-case studies demonstrate performance of the CopPF method
- Results suggest that CopPF provides greater opportunities to access new samples which in turn lead to more accurate predictions

**Correspondence to:**
G. H. Huang,
huangg@uregina.ca

# Development of a copula-based particle filter (CopPF) approach for hydrologic data assimilation under consideration of parameter interdependence

**Y. R. Fan[1], G. H. Huang[1,2], B. W. Baetz[3], Y. P. Li[2], and K. Huang[4]**

[1]Institute for Energy, Environment and Sustainable Communities, University of Regina, Regina, Saskatchewan, Canada, [2]Center for Energy, Environment and Ecology Research, UR-BNU, Beijing Normal University, Beijing, China, [3]Department of Civil Engineering, McMaster University, Hamilton, Canada, [4]Faculty of Engineering and Applied Science, University of Regina, Regina, Saskatchewan, Canada

**Abstract** In this study, a copula-based particle filter (CopPF) approach was developed for sequential hydrological data assimilation by considering parameter correlation structures. In CopPF, multivariate copulas are proposed to reflect parameter interdependence before the resampling procedure with new particles then being sampled from the obtained copulas. Such a process can overcome both particle degeneration and sample impoverishment. The applicability of CopPF is illustrated with three case studies using a two-parameter simplified model and two conceptual hydrologic models. The results for the simplified model indicate that model parameters are highly correlated in the data assimilation process, suggesting a demand for full description of their dependence structure. Synthetic experiments on hydrologic data assimilation indicate that CopPF can rejuvenate particle evolution in large spaces and thus achieve good performances with low sample size scenarios. The applicability of CopPF is further illustrated through two real-case studies. It is shown that, compared with traditional particle filter (PF) and particle Markov chain Monte Carlo (PMCMC) approaches, the proposed method can provide more accurate results for both deterministic and probabilistic prediction with a sample size of 100. Furthermore, the sample size would not significantly influence the performance of CopPF. Also, the copula resampling approach dominates parameter evolution in CopPF, with more than 50% of particles sampled by copulas in most sample size scenarios.

## 1. Introduction

Hydrologic models are widely used for many water resource management applications such as water allocation, reservoir operation, risk assessment [e.g., *Mateo et al.*, 2014; *Fan et al.*, 2015a; *Ma et al.*, 2016]. In such models, simplified equations are often used to describe a variety of temporally dynamic and spatially distributed processes in watershed systems. Such descriptions may lead to uncertainties existing in model parameters and structures, which further result in randomness in the resulting model predictions. It is of great importance to deal with various uncertainty sources in hydrologic models in order to provide reliable hydrologic predictions [e.g., *Li et al.*, 2013; *DeChant and Moradkhani*, 2014; *Fan et al.*, 2015b, 2015c; *Yan et al.*, 2015; *Fan et al.*, 2016a].

In the past decade, a large number of approaches have been proposed to quantify uncertainties in hydrologic predictions, among which sequential data assimilation has been widely used for real-time hydrologic prediction. Sequential data assimilation characterizes uncertainties associated with inputs, states, parameters, and outputs of a given model through incorporating available observations of the system, state, or output into prediction process [*Moradkhani et al.*, 2012]. This kind of method continuously assimilates available measurements to update states and parameters to improve model forecasts [*Vrugt et al.*, 2005]. Ensemble Kalman filter (EnKF) and its variants (e.g., ensemble square root filter, iterative ensemble Kalman filter), as well as particle-based filters (e.g., particle filter (PF), particle Markov chain Monte Carlo) are two typical sequential data assimilation categories. The EnKF and its variants use ensembles to approximate state covariance matrices to achieve suboptimal state estimations in which the error statistics are predicted by solving the Fokker-Planck equation using the Monte Carlo method [*Evensen*, 2003; *Shen and Tang*, 2015]. However, the EnKF can achieve good performances with small ensembles; but related studies have claimed
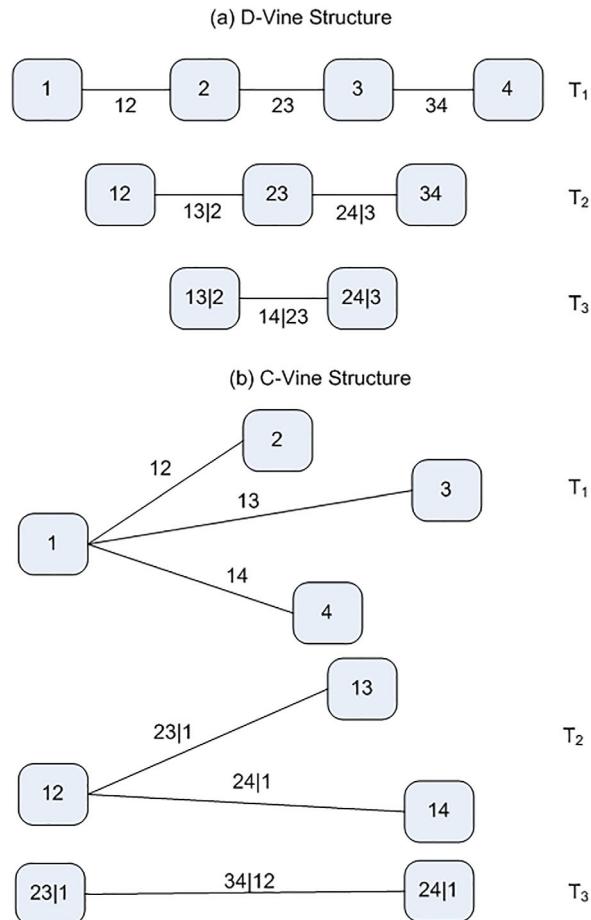
**Figure 1.** Illustration of C-vine and D-vine copula structures for four variables: A C or D-Vine with four variables has three trees and six edges, with each edge associated with a pair-copula.

problems with EnKF resulting from its inherent assumptions [*Moradkhani et al.*, 2005a; *Montzka et al.*, 2011; *Plaza-Guingla et al.*, 2012; *DeChant and Moradkhani*, 2012].

The particle filter (PF) method, as an example of sequential Monte Carlo (SMC) methods, has been suggested to be applicable for those cases violating the assumption of Gaussian errors in the EnKF approach [*Moradkhani et al.*, 2005b; *Weerts and El Serafy*, 2006; *Noh et al.*, 2014; *Plaza-Guingla et al.*, 2013; *Fan et al.*, 2016, 2017]. PF evolves the sample sets of state variables forward using the SMC method to approximate the predictive distributions [*Liu et al.*, 2012]. The PF and its variants update particle weights instead of the original values for state variables, which can reduce numerical instability especially in physically based or process-based models [*Liu et al.*, 2012]. The implementation of PF is based on the sequential importance sampling (SIS) algorithm [*Liu et al.*, 2001; *Moradkhani et al.*, 2012]. However, SIS generally leads to severe deterioration for particles, which means several particles are merely available with significant weight values. Therefore, a resampling procedure is usually used in PF to replace insignificant particles with those with significant weights. Such a procedure can mitigate sample deterioration in PF and maintain the effective sample size. Resampling techniques mainly include residual resampling, multinomial resampling, and systematic resampling

methods. Details for resampling implementation can be found in *Li et al.* [2015]. Resampling algorithms are able to eliminate weight degeneration in PF, but they may lead to sample impoverishment, in which the diversity of particles is small and cannot sufficiently characterize posterior distributional information for states and parameters. Recent studies proposed particle Markov chain Monte Carlo (PMCMC) approaches by combining the strengths of SMC and MCMC to allow more sufficient representation of the posterior distribution, reduce the chance of sample impoverishment, and lead to a more accurate streamflow forecast with manageable ensemble sizes [*Moradkhani et al.*, 2012; *Vrugt et al.*, 2013; *Yan et al.*, 2015; *Yan and Moradkhani*, 2016].

It is common that unknown parameters in hydrologic models are correlated with each other. Studies have demonstrated that parameter correlation can have crucial effects on both parameter estimation and uncertainties in model predictions [*Pohlmann et al.*, 2002; *Lemke et al.*, 2004; *Xu and Gertner*, 2008; *Manache and Melching*, 2008; *Rojas et al.*, 2009; *Pan et al.*, 2011]. Full consideration of parameter interdependence in data assimilation may be an alternative to improve posterior sampling procedure and mitigate particle impoverishment. Therefore, as an extension of previous research, a copula-based particle filter (CopPF) method is proposed in this study. In CopPF, multivariate copulas are introduced into the resampling procedure to reveal the multidimensional correlation structure of model parameters, and the new particles are resampled based on the obtained copulas. Copula approaches have been widely used for multivariate hydrologic modeling [e.g., *Kong et al.*, 2015; *Fan et al.*, 2016b, 2016c; *Xu et al.*, 2017]. Recently *Bárdossy and Hörning* [2016] introduced a copula-based method for Gaussian and non-Gaussian inverse modeling of groundwater flow. Nevertheless, no previous research has been reported to introduce copulas into PF for particle rejuvenation. The developed CopPF approach is the first attempt to improve the resampling procedure in PF by a

**Figure 2.** The flow chart of CopPF.

full characterization of parameter multidimensional correlation structures. The applicability of CopPF is illustrated through two case studies by using a two-parameter regression problem and a hydrologic data assimilation problem.

## 2. Methodology

### 2.1. Particle Filter Method

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

$$x_{t,i} = g(x_{t-1,i}, u_{t,i}, \theta_{t,i}) + \omega_{t,i} \quad (1)$$

where $g$ is a nonlinear function expressing states transition in the hydrologic model; $x_{t,i}$ is a set of sample for model states (expressed as $x_t$); $u_{t,i}$ represents the forcing data; $\theta_{t,i}$ is a sample set of model parameter vector $\theta_t$; and $\omega_{t,i}$ is the process noise in the model prediction. The model output $\tilde{y}_t$ related to real measurements (e.g., streamflow) can be obtained through the measurement operator $h(.)$, subject to model states and parameters as follows:

$$\tilde{y}_{t,i} = h(x_{t,i}, \theta_{t,i}) + v_{t,i} \quad (2)$$

where $\tilde{y}_{t,i}$ is the simulated value for the observation (e.g., streamflow); $v_{t,i}$ indicates the observation noise [*Moradkhani et al.*, 2012].

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.*, 2016a]. Based on equation (1), the transition probability $p(x_t, \theta_t| x_{t-1}, \theta_{t-1})$ can be obtained to describe evolution of model states and parameters, while the likelihood (denoted as $p(y_t|x_t, \theta_t)$) of $y_t$ given estimate for $x_t, \theta_t$ can be obtained from equation (2) [*Shen and Tang*, 2015]. Based on Bayesian law, the posterior of state $x_t$, and parameter $\theta_t$ can be obtained as follows:

$$p(x_t, \theta_t|y_{1:t}) = \frac{p(y_t|x_t, \theta_t)p(x_t, \theta_t|y_{1:t-1})}{\int p(y_t|x_t, \theta_t)p(x_t, \theta_t|y_{1:t-1})dx_t d\theta_t} \quad (3)$$

where $p(x_t, \theta_t| y_{1:t-1})$ represent the prior information. When the model is assumed to be Markovian, the prior distribution can be estimated via the Chapman-Kolmogorov equation expressed as follows [*Moradkhani et al.*, 2012]:

$$p(x_t, \theta_t | y_{1:t-1}) = \int p(x_t, \theta_t | x_{t-1}, \theta_{t-1}) p(x_{t-1}, \theta_{t-1} | y_{1:t-1}) dx_{t-1} d\theta_{t-1} \qquad (4)$$

where $p(x_{t-1}, \theta_{t-1} | y_{1:t-1})$ is the posterior distribution at previous time step. 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, and particle filter (PF) is one of the most attractable methods.

The PF method is a sequential Monte Carlo method that calculates the posterior distribution of states and parameters by a set of random samples. In PF, the posterior distribution of state variable $x_t$ and model parameter $\theta_t$, given observations **y** at previous $t$ steps, can be expressed as a weighted function [*Arulampalam et al.*, 2002; *Moradkhani et al.*, 2005b, 2012]:

$$p(x_t, \theta_t | y_{1:t}) = \sum_{i=1}^{N} w_{t,i}^{+} \delta(x_t - x_{t,i}, \theta_t - \theta_{t,i}) \qquad (5)$$

where $\{x_{t,i}, \theta_{t,i}\}$ is a particle set for model states and parameters with posterior weight values of $w_{t,i}^{+}$; $\delta$ is the Dirac delta function which is equal to 1 when $x_t = x_{t,i}$ and $\theta_t = \theta_{t,i}$, and equal to 0 otherwise; and $N$ is the sample size.

Assuming the system state to be a Markov process, the posterior weight $w_{t,i}^{+}$ in equation (3) can be obtained through a Bayesian recursive expression as follows [*Moradkhani et al.*, 2012; *Yan and Moradkhani*, 2016]:

$$w_{t,i}^{+} = w_{t,i}^{-} \frac{p(y_t | x_{t,i}, \theta_{t,i})}{\sum_{i=1}^{N} w_{t,i}^{-} p(y_t | x_{t,i}, \theta_{t,i})} \qquad (6)$$

where $w_{t,i}^{-}$ is the prior weight for particle $i$. $p(y_t | x_{t,i}, \theta_{t,i})$ can be obtained through the likelihood function $L(y_t | x_{t,i}, \theta_{t,i})$. Equation (6) holds when the transition prior is chosen as the proposal distribution in PF. Moreover, in the data assimilation process through PF, the Gaussian likelihood is widely used in a number of fields [*Moradkhani et al.*, 2005b; *Weerts and El Serafy*, 2006; *Salamon and Feyen*, 2010; *Fan et al.*, 2016a; *Yan and Moradkhani*, 2016], which is expressed as

$$L(y_t | x_{t,i}, \theta_{t,i}) = \frac{1}{\sqrt{(2\pi)^m |R_t|}} \exp\left(-\frac{1}{2R_t} [y_{t,i} - \tilde{y}_{t,i}]^2\right) \qquad (7)$$

In hydrologic data assimilation through PF, a major challenge is the depletion of a particle set, which means that, after a few iterations (time steps), all particles except one are discarded because their importance weights are insignificant [*Doucet et al.*, 2001; *Moradkhani et al.*, 2005b]. To address the above issue, sampling importance resampling (SIR) algorithms are usually applied to eliminate the particles with small importance weights and replace them with particles with large importance weights.

### 2.2. Modeling Multiple Dependence Through Copulas

A copula functions is a multivariate distribution function with uniform margins on the interval [0, 1]. Sklar's Theorem states that any $n$-dimensional distribution function $F$ can be formulated through a copula and its marginal distributions [*Nelsen*, 2006]. Consider one a random vector ($\mathbf{X} = (X_1, X_2, \ldots, X_d)$) with the marginal distributions denoted as $F_1(x_1|\gamma_1), F_2(x_2|\gamma_2), \ldots, F_d(x_d|\gamma_d)$, where $\gamma_1, \gamma_2, \ldots, \gamma_d$ are parameters in probability distributions. The joint probability distribution of $X_1, X_2, \ldots, X_d$ can be expressed as

$$F(x_1, \ldots, x_d | \gamma_1, \ldots, \gamma_d, \theta) = C(F_1(x_1|\gamma_1), \ldots, F_d(x_d|\gamma_d)|\theta) \qquad (8)$$

where $C(.)$ is a copula function and $\theta$ is the parameter in the copula function describing dependence among those correlated variables. If these marginal distributions are continuous, a single copula function $C$ exists, which can be written as [*Nelsen*, 2006; *Aas et al.*, 2009]

$$C(u_1, u_2, \ldots, u_n | \theta) = F(F_{X_1}^{-1}(u_1|\gamma_1), F_{X_2}^{-1}(u_2|\gamma_2), \ldots, F_{X_d}^{-1}(u_d|\gamma_d)) \qquad (9)$$

Based on equations (8) and (9), the corresponding probability density function of random vector **X** can be formulated as [*Nelsen*, 2006; *Aas et al.*, 2009]
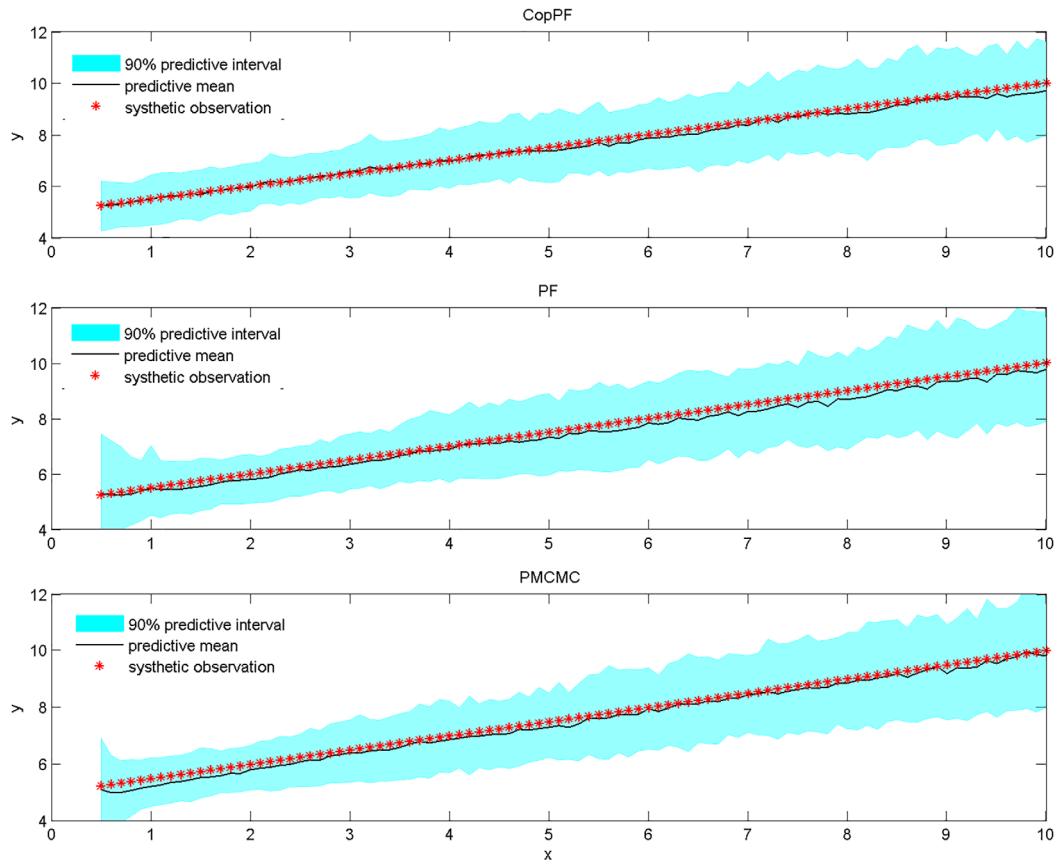
**Figure 3.** Comparison of CopPF, PF, and PMCMC for the toy model. The canyon belt indicates the 90% predictive intervals consisting of the 5% and 90% quantile values.

$$f(x_1, \ldots, x_d|\gamma_1, \ldots, \gamma_d, \theta) = c(F_1(x_1|\gamma_1), \ldots, F_d(x_d|\gamma_d)|\theta) \times f_1(x_1|\gamma_1) \times \ldots \times f_d(x_d|\gamma_d) \qquad (10)$$

where $f_i(x_i|\gamma_i)$ is the probability density function of $X_i$ and $c(.)$ is the density function of copula expressed as [*Nelsen*, 2006; *Aas et al.*, 2009]

$$c(u_1, u_2, \ldots, u_d) = \frac{\partial^d C(u_1, u_2, \ldots, u_d)}{\partial u_1 \ldots \partial u_d} \qquad (11)$$

One of most attractable advantages of copulas is that modeling of the marginal distributions can be conveniently separated from the dependence modeling, leading to flexibility in selection of both the marginal and dependent models [*Brechmann and Schepsmeier*, 2013; *Fan et al.*, 2016b, 2016c]. A great number of bivariate copulas have been developed and investigated for bivariate cases such as Gaussian, Student's *t*, and Archimedean copulas [*Nelsen*, 2006]. However, for random variables larger than three, the available copula families are rather limited and the standard multivariate copulas may not model their interdependence [*Brechmann and Schepsmeier*, 2013].

The pair-copula approach constructs the joint distribution of a *d*-dimensional multivariate random variable through $d(d-1)/2$ bivariate copulas [*Aas et al.*, 2009]. There are a great number of possible pair-copula constructions, among which are the canonical vine (C-vine) and D-vine as two widely used decomposition structures [*Aas et al.*, 2009; *Xiong et al.*, 2015]. Figure 1 shows the structures of C-vine and D-vine for four random variables.

Consider a random vector $(X_1, X_2, \ldots, X_d)$ with $F_k(x_k)$ and $f_k(x_k)$ being the CDF and PDF for random variable $X_k$, the joint PDF

**Table 1.** Performance of CopPF, PF, and PMCMC on the Toy Model

|       | NSE    | RMSE   | CRPS   |
|-------|--------|--------|--------|
| CopPF | 0.9849 | 0.1334 | 0.2034 |
| PF    | 0.9657 | 0.2164 | 0.2389 |
| PMCMC | 0.9839 | 0.1481 | 0.2241 |

$f(x_1, x_2, \ldots, x_d)$, constructed by a D-vine structure, can be expressed as

$$f(x_1, x_2, \ldots, x_d) = \prod_{k=1}^{d} f_k(x_k|\gamma_k) \prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{i,i+j|i+1,\ldots,i+j-1}(F_i(x_i|x_{i+1}, \ldots, x_{i+j-1}),$$

$$F_{i+j}(x_{i+j}|x_{i+1}, \ldots, x_{i+j-1})|\theta_{i,i+j|i+1,\ldots,i+j-1})$$

(12)

where $\gamma_k$ denotes the parameters in marginal distributions and $\theta_{i,i+j|i+1,\ldots,i+j-1}$ represents the parameters in copulas. The pair-copula constructions decompose a multivariate probability density into bivariate copulas, with each pair-copula chosen independently from the others, which allows for an enormous flexibility in dependence modeling [*Brechmann and Schepsmeier*, 2013].

### 2.3. Copula-Based Particle Filter

Resampling steps are essential for particle filters to mitigate the particle degeneracy effect. Resampling is a process in which a new random measure $\chi_t^* = \{x_t^{(i*)}, w_t^{(i*)}\}_{i*=1}^{N*}$ is created from the original random measure $\chi_t = \{x_t^{(i)}, w_t^{(i)}\}_{i=1}^{N}$ and then replaces the original random measure $\chi_t$ [*Li et al.*, 2015]. In this process, particles with large weights are duplicated to replace those particles with small 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]. However, traditional resampling methods may also lead to undesired effects. One of them is sample impoverishment, which means the diversity of the particles is reduced due to the removal of the low-weighted particles [*Gordon et al.*, 1993; *Beadle and Djuric*, 1997; *Bolic et al.*, 2004; *Li et al.*, 2015]. For example, if only a few particles of $\chi_t$ have significant weights, many of the resampled particles will be the same, i.e., the number of different particles in $\chi_t^*$ will be small [*Li et al.*, 2015]. Recently, some research works have investigated sample impoverishment in the resampling process through combining the strengths of sequential Monte Carlo sampling and Markov chain Monte Carlo simulation [*Moradkhani et al.*, 2012; *Vrugt et al.*, 2013; *Yan and Moradkhani*, 2016]. However, the presence of parameter correlation and interaction is not efficiently considered in previous resampling methods, which may reduce the reliability of those algorithms. Consequently, this study aims to propose a copula-based particle filter (CopPF) method by incorporating multivariate copulas in the resampling procedure to reveal the multidimensional correlation structure of model parameters and then rejuvenate particle evolution in a more reasonable way.

In CopPF, the dependence structure of model parameters is quantified through copulas, and the resampled particles are then generated based on the obtained copulas. In general, a hydrologic model has several unknown parameters larger than two, leading to dependence structure of multidimensional. Consequently, the pair-copula method is employed in this study to quantify the multidimensional correlation structure of model parameters. The CopPF method improves upon previous PF methods through resampling new particles based on the joint probability function of the parameters obtained by the pair-copula method. This copula-based resampling process can evolve parameter samples by considering their correlation structure, which can lead to new particles located in the high probability density region. Such new particles are also different from the old one, which can keep diversity of particles,
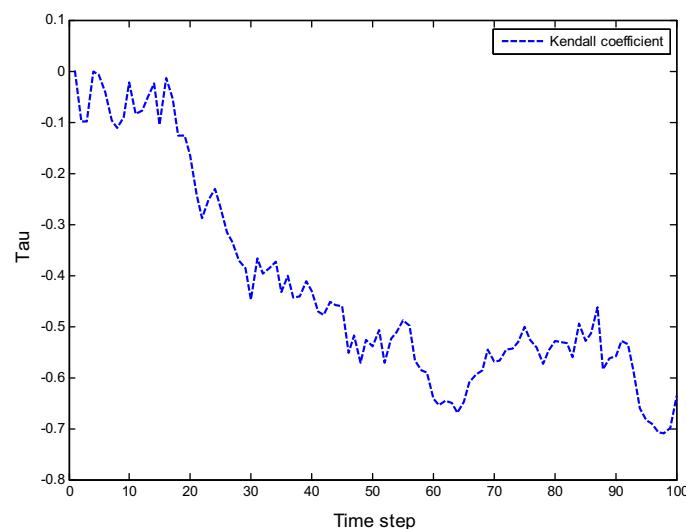


**Figure 4.** The dependence between $\alpha_1$ and $\alpha_2$ during the data assimilation process measured by the Kendall's $\tau$. The results suggest highly correlation between $\alpha_1$ and $\alpha_2$ in the data assimilation process after a short warm-up period ($p < 0.05$).
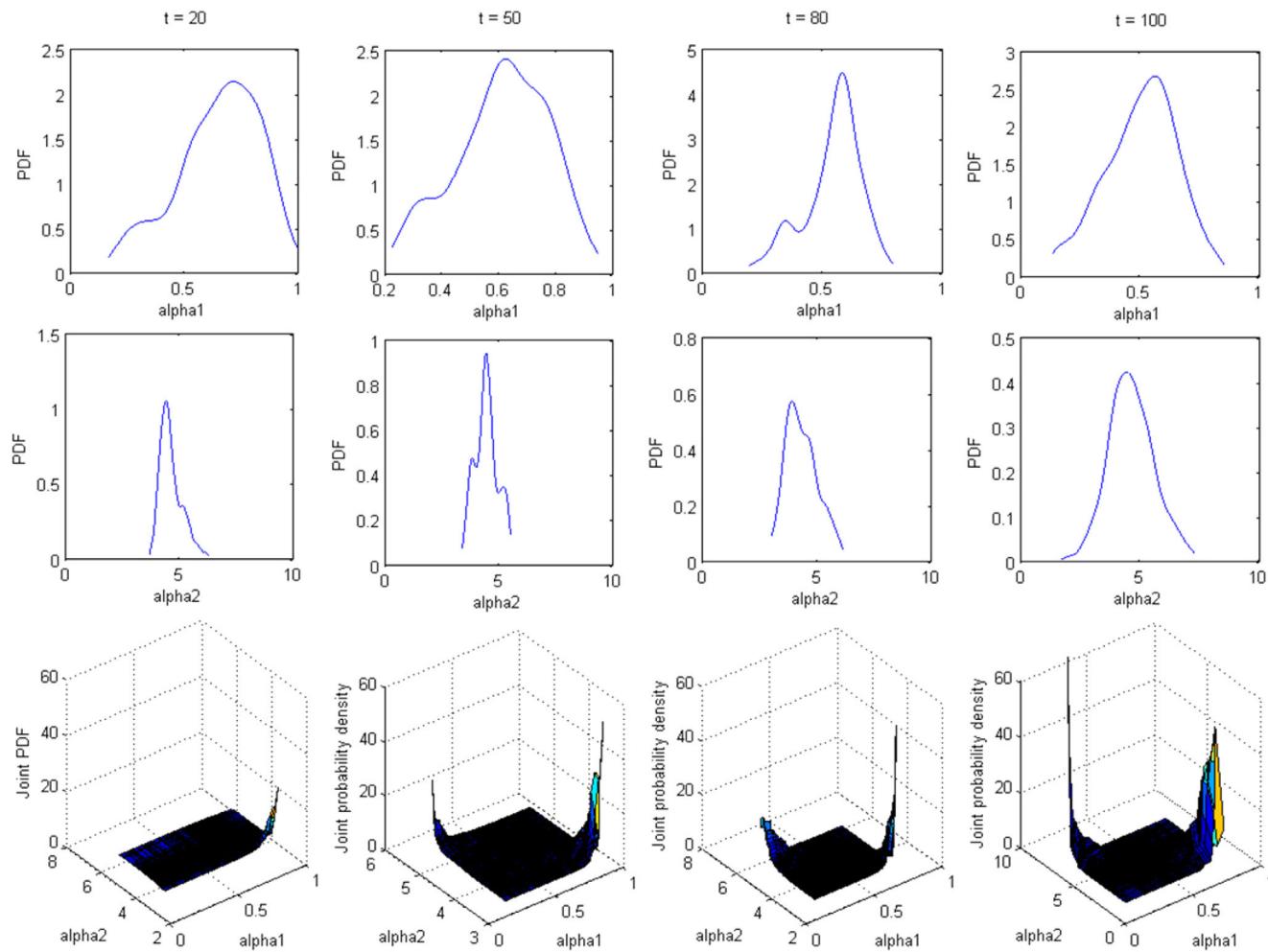
**Figure 5.** Parameter dependence during the data assimilation process. The first row indicates the posterior distribution for $\alpha_1$ under different time step, the second row shows posterior distribution for $\alpha_2$, and all marginal distributions are obtained by kernel smoothing approach. The last row shows the joint probability distribution for $\alpha_1$ and $\alpha_2$, which indicate significant dependence between these two parameters.

mitigate particle impoverishment, and then may lead to better performance than traditional PF-based approaches.

Figure 2 shows the description of the proposed CopPF approach, in which the detailed process is given below.

1. The initial state variables and parameters are sampled from their corresponding uniform distributions:

$$\mathbf{x}_{1,i} \sim \ U\left(\mathbf{x}^L, \mathbf{x}^U\right), i = 1, \ 2, \ \ldots, ne, x \in R^{N_x} \tag{13}$$

$$\boldsymbol{\theta}_{1,i} \sim \ U(\boldsymbol{\theta}^L, \boldsymbol{\theta}^U), i = 1, \ 2, \ \ldots, ne, \ \theta \in R^{N_\theta} \tag{14}$$

where $\boldsymbol{\theta}$ ($\boldsymbol{\theta} = \{\theta^{(1)}, \ldots, \theta^{(d)}\}$) denotes the model parameter vector, $\mathbf{x}$ denotes the state variable vector, and $ne$ means the sample size in data assimilation process. $\mathbf{x}^L$ and $\mathbf{x}^U$ denote the predefined lower and upper bounds of state variable $x$ and $\boldsymbol{\theta}^L$ and $\boldsymbol{\theta}^U$ denote the predefined lower and upper bounds of parameter $\boldsymbol{\theta}$.

2. Assign a prior weight for each particle. In general, the prior weights are assigned uniformly, with each one being equal to $1/ne$.

3. At any time step $t$, the model states for the current step can be forecasted based on the posterior states in step $t - 1$ and the prior parameters in the current step, as expressed by equation (1). The simulated observations are then obtained through equation (2).

4. The particle weights for model states and parameters are updated through equations (6) and (7).
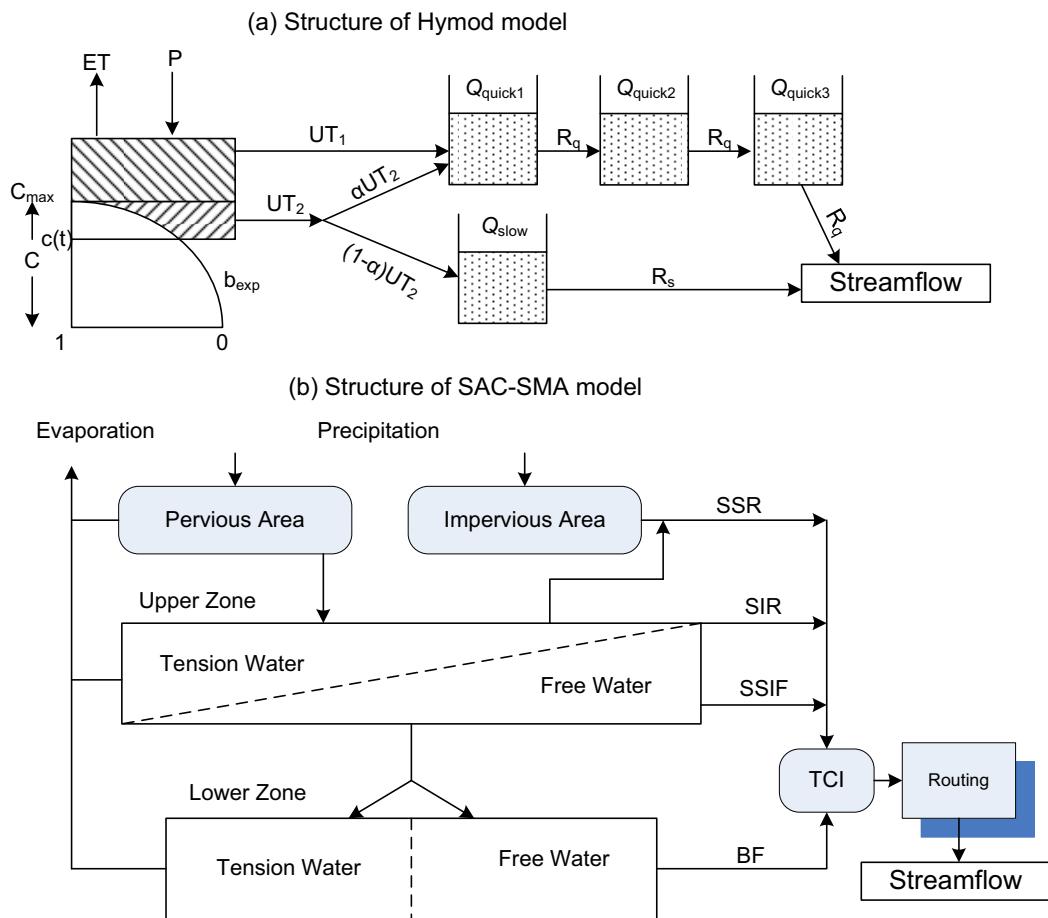
**Figure 6.** Description of Hymod and SAC-SMA models [adopted from *Vrugt et al.*, 2003; *Lakshmi et al.*, 2014].

5. The particles with insignificant weights are removed through a sampling importance resample (SIR) process. In this work, the multinomial resampling scheme proposed by *Moradkhani et al.* [2005b] is employed.

6. Based on the resampled particles $\boldsymbol{\theta}_{t+1-resamp,i}$, the marginal distribution is constructed for each parameter expressed as $\theta^{(i)} \sim G_i(\theta|\gamma_i)$, where $\gamma_i$ are the unknown parameters in marginal distributions. For $\boldsymbol{\theta}_{t+1-resamp,i}$, their distributions can hardly be quantified by some specific distributions (e.g., normal, lognormal) since $\boldsymbol{\theta}_{t+1-resamp,i}$ represent the posterior distributions of model parameters at time step $t+1$. Consequently, the nonparametric kernel density estimator is employed to approximate the distributions of $\boldsymbol{\theta}_{t+1-resamp,i}$. For a set of random sample $\{x_1, x_2, \ldots, x_n\}$ from a distribution with unknown density of $f(x)$, the associated kernel density estimator can be expressed as [*Malec and Schienle*, 2014]

$$\hat{f}(x) = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{x - x_i}{b}\right) \tag{15}$$

where b is a smoothing bandwidth with $b \to 0$ and $nb \to \infty$ and K is a kernel function with its integral equal to 1, i.e., $\int_{-\infty}^{+\infty} K(u)du = 1$.

**Table 2.** 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 | $\alpha$ | [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 |

7. Transform the original parameter values $\boldsymbol{\theta}_{t+1\text{-resamp}}$ to uniform values in [0, 1] by $u^{(i)} = G_i(\theta^{(i)}_{t+1-resamp}|\gamma_i)$, $i = 1, 2, \ldots, d$, where the superscript $i$ denotes the $i$th parameter in the original hydrologic model, and $G_i$ ($i = 1, 2, \ldots, d$) represent the CDF of the sample marginal for $\boldsymbol{\theta}_{t+1\text{-resamp}}$ obtained by the kernel estimation method in Step 6. This step is required because construction of a copula need to its margins on the interval [0, 1]. In this step, the original samples of model parameters are converted to the corresponding probabilities (denoted as $u^{(i)}$) through their estimated probability functions ($G_i$). These probability values are used for determining the multivariate copula function in the next step.

8. Based on $\{u^{(i)}\}$, fit the joint probability density with a D-vine structure expressed as

$$c_{1:d}(\mathbf{u}|\boldsymbol{\eta}) = \prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{i,i+j|i+1,\ldots,i+j-1}(F(u_i|u_{i+1},\ldots,u_{i+j-1}), F(u_{i+j}|u_{i+1},\ldots,u_{i+j-1})|\boldsymbol{\eta}) \tag{16}$$

and the joint density function for parameter $\theta$ can be expressed as

$$f(\boldsymbol{\theta}|\boldsymbol{\eta},\boldsymbol{\gamma}) = c_{1:d}(G_1(\theta^{(1)}|\gamma_1),\ldots,G_d(\theta^{(d)}|\gamma_d)|\boldsymbol{\eta}) \prod_{i=1}^{d} g_i(\theta^{(i)}|\gamma_i) \tag{17}$$

9. Generate new samples for $\tilde{\mathbf{u}}$ through the obtained copula function $c_{1\ldots d}(G_1(\theta|\gamma_1),\ldots,G_d(\theta|\gamma_d)|\boldsymbol{\eta})$, where $\tilde{\mathbf{u}} \in [0, 1]^d$. A number of approaches have been proposed to generate samples from multivariate copulas. In this study, the R package "CDVine" [*Brechmann and Schepsmeier*, 2013] is adopted to generate new samples from a multivariate copula with a D-vine structure.

10. Generation of the original sample value for $\theta^{(i)}$ can be obtained by

$$\theta^{(i)} = G_i^{-1}(\tilde{u}^{(i)}|\gamma_i) \tag{18}$$

11. Generate alternative particles through parameter perturbation by adding small stochastic errors around the sample:

$$\theta^{(i)*}_{t+1,i} = \theta^{(i)}_{t+1-resamp,i} + \tau_{t+1,i}, \ \tau_{t+1,i} \sim N(0, \eta S(\theta^{(i)}_{t+1,i})) \tag{19}$$

where $\eta$ is a hyperparameter indicating the radius near each sample being explored. This hyperparameter is identified artificially [*Moradkhani et al.*, 2005b; *DeChant and Moradkhani*, 2012] and set to be 0.1 in this study. Also, such a parameter can be adjusted in data assimilation process through the variance multiplier approach proposed by *Leisenring and Moradkhani* [2012]. $S(\theta^{(i)}_{t+1,i})$ is the standard deviation of particle values.

12. Based on Steps 6–11, two sets of samples are available for model parameters in the next time step. One set is generated through copula resampling method described by equations (15–18), and the other set is obtained by stochastic perturbation method described by equation (19). The question is which set of samples will be used for the next time step. A data mismatch index is employed to evaluate the performance of the two sets of resampled particles, which is formulated as

$$S(\mathbf{x}_t, \boldsymbol{\theta}_t) = \sum_{i=1}^{ne} (h(\mathbf{x}_{t,i}, \boldsymbol{\theta}_{t,i}) - y_{t,i})^T R_t^{-1} (h(\mathbf{x}_{t,i}, \boldsymbol{\theta}_{t,i}) - y_{t+1,i}) \tag{20}$$

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 for two purposes: (i) it can ensure that the better parameter estimations are chosen in each step and (ii) it can help avoid divergence of the resampled particles in the data assimilation process. Assume that $\boldsymbol{\theta}^{(cop)}_{t+1}$ are the particles generated through copula resampling method and $\boldsymbol{\theta}^*_{t+1-resamp}$ are the new particles generated by the stochastic perturbation method. If $S(\mathbf{x}_{t+1-resamp}, \boldsymbol{\theta}^{(cop)}_{t+1}) \leq S(\mathbf{x}_{t+1-resamp}, \boldsymbol{\theta}^*_{t+1-resamp})$, $\mathbf{x}_{t+1} = \mathbf{x}_{t+1-resamp}$, and $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}^{(cop)}_{t+1}$, otherwise, $\mathbf{x}_{t+1} = \mathbf{x}_{t+1-resamp}$ and $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}^*_{t+1-resamp}$.

13. Let $t = t + 1$ if measurements are still available in the next stage and then return to Step 2; otherwise, stop.

**Table 3.** The Performance of CopPF, PF, and PMCMC in the Synthetic Experiment of Hymod Under Different Sample Size Scenarios

| | | CopPF | | | PF | | | PMCMC | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | NSE | RMSE | CRPS | NSE | RMSE | CRPS | NSE | RMSE | CRPS |
| 50 | Mean | 0.858 | 11.752 | 3.889 | 0.769 | 14.845 | 4.945 | 0.853 | 11.826 | 4.338 |
| | Min | 0.747 | 9.376 | 2.919 | 0.486 | 10.020 | 2.676 | 0.722 | 8.105 | 2.758 |
| | Max | 0.911 | 15.836 | 5.640 | 0.899 | 22.571 | 7.966 | 0.934 | 16.608 | 6.416 |
| 100 | Mean | 0.883 | 10.757 | 3.144 | 0.827 | 12.856 | 3.819 | 0.876 | 10.979 | 3.647 |
| | Min | 0.833 | 9.405 | 2.505 | 0.614 | 9.660 | 2.569 | 0.798 | 8.419 | 2.515 |
| | Max | 0.911 | 12.876 | 4.326 | 0.906 | 19.567 | 7.203 | 0.929 | 14.145 | 6.309 |
| 200 | Mean | 0.889 | 10.456 | 2.792 | 0.870 | 11.289 | 3.080 | 0.885 | 10.649 | 2.885 |
| | Min | 0.841 | 9.123 | 2.380 | 0.786 | 9.786 | 2.523 | 0.864 | 9.648 | 2.493 |
| | Max | 0.916 | 12.541 | 3.436 | 0.903 | 14.581 | 5.237 | 0.906 | 11.614 | 3.376 |
| 300 | Mean | 0.893 | 10.281 | 2.741 | 0.873 | 11.204 | 2.792 | 0.887 | 10.546 | 2.799 |
| | Min | 0.867 | 9.400 | 2.434 | 0.824 | 9.646 | 2.480 | 0.857 | 9.092 | 2.455 |
| | Max | 0.911 | 11.473 | 3.555 | 0.906 | 13.222 | 3.358 | 0.917 | 11.903 | 3.377 |
| 400 | Mean | 0.894 | 10.245 | 2.616 | 0.881 | 10.857 | 2.651 | 0.891 | 10.395 | 2.700 |
| | Min | 0.873 | 9.268 | 2.443 | 0.837 | 9.758 | 2.478 | 0.869 | 9.570 | 2.469 |
| | Max | 0.913 | 11.230 | 2.780 | 0.904 | 12.694 | 3.016 | 0.908 | 11.413 | 3.308 |
| 500 | Mean | 0.897 | 10.097 | 2.587 | 0.883 | 10.730 | 2.687 | 0.892 | 10.322 | 2.697 |
| | Min | 0.884 | 9.413 | 2.414 | 0.854 | 9.328 | 2.436 | 0.863 | 9.114 | 2.482 |
| | Max | 0.911 | 10.712 | 2.758 | 0.912 | 12.035 | 3.182 | 0.916 | 11.644 | 3.064 |

## 3. Case Study I: Simplified Model

A regression model with two unknown parameters, expressed as $y_t = \alpha_1 x + \alpha_2$ is employed to illustrate the CopPF approach. This simplified model is similar to the one proposed by *Vrugt and Sadegh* [2013]. The purpose of this model is to explore the correlation of model parameters during the data assimilation process. Since only two parameters are involved in this simplified model, the joint probability of model parameters can be easily visualized. Synthetic observations are first generated from this model with $x \in [0, 10]$, and $\alpha_1 = 0.5$ and $\alpha_2 = 5.0$. A total number of 100 observations are generated which are denoted as $Y = \{y_1, y_2, \ldots, y_n\}$ for $n = 100$. The proposed CopPF approach is then employed to identify the posterior distributions for $\alpha_1$ and $\alpha_2$ from their initial prior distributions, in which $\alpha_1$ is uniformly distributed within [0, 1] and $\alpha_2$ has a uniform prior of [–10, 10].
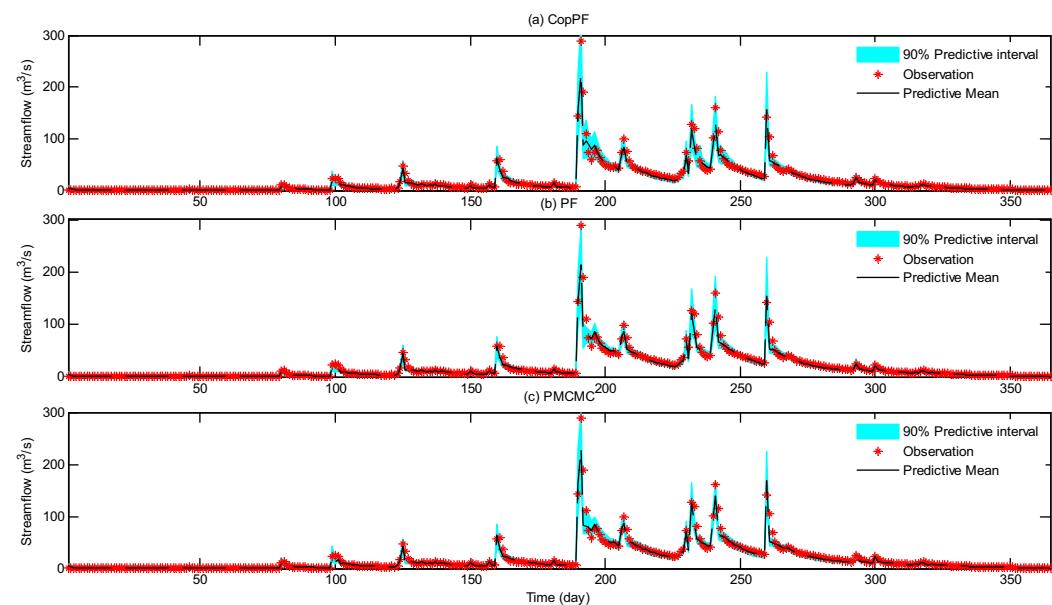


**Figure 7.** Comparison between predictions from different data assimilation schemes and synthetic observations from Hymod: The red stars indicate the synthetic observations while the dashed line shows predictive means. The cyan belt exhibits the 90% predictive intervals consisting of the 5% and 95% quantile values. The results show that most observations are bracketed by the predictive intervals, demonstrating the accuracy of all three approaches.
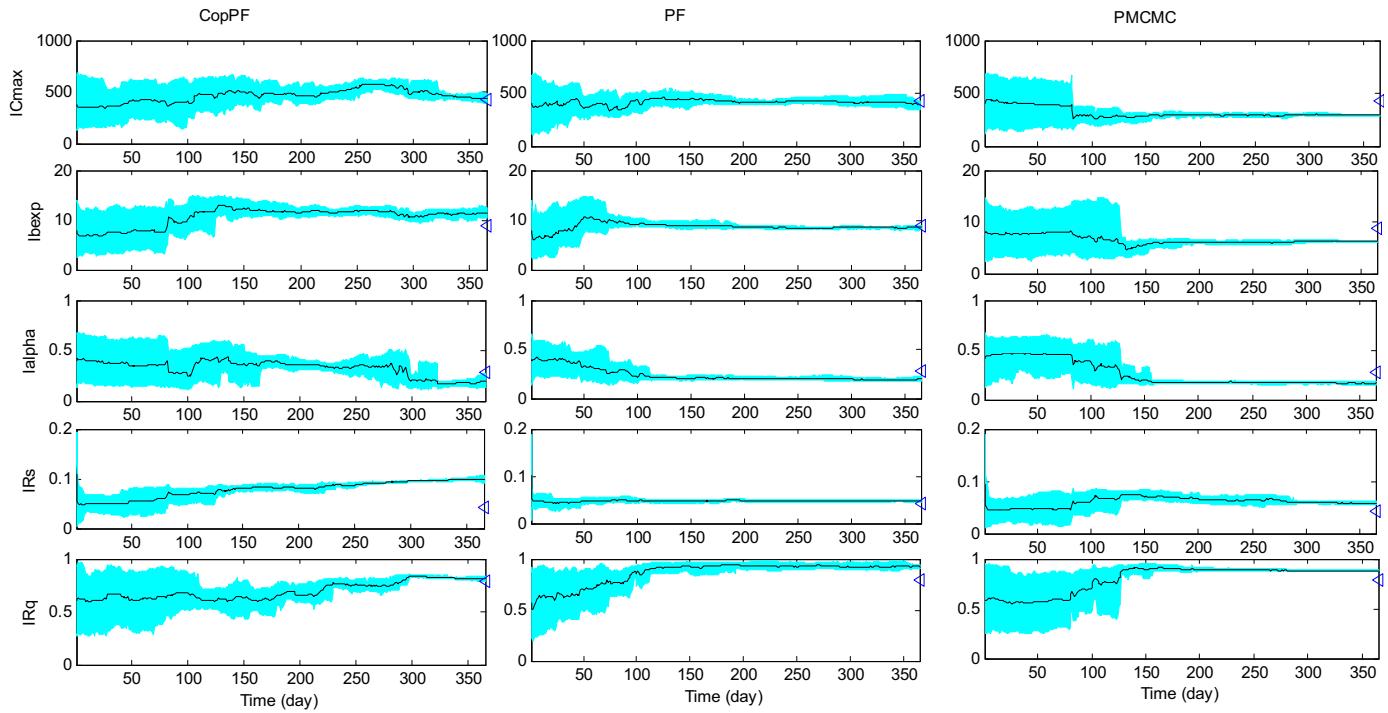
**Figure 8.** Parameter evolution of Hymod in the data assimilation process through CopPF, PF, and PMCMC: The cyan region exhibits the 90% predictive uncertain ranges (bracketed by the 5% and 95% quantiles). The black line denotes the predictive means and the blue triangles at the right-hand side suggest the actual parameter values to generate the synthetic discharge data.

Three indices including root-mean-square error (RMSE), the Nash-Sutcliffe efficiency (NSE) coefficient, and continuous ranked probability score (CRPS) are used for the evaluation of the performance of data assimilation approaches. RMSE and NSE are based on the ensemble average, which are expressed as

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (Qobs_i - Qsim_i)^2} \tag{21}$$

$$NSE = 1 - \frac{\sum_{i=1}^{N} (Qobs_i - Qsim_i)^2}{\sum_{i=1}^{N} (Qobs_i - \overline{Qobs})^2} \tag{22}$$

where $N$ is the total number of streamflow measurements (or predictions), $Qobs_i$ are the observations, $Qsim_i$ are the predictions from the hydrologic model, and $\overline{Qobs}$ is the mean of observations. The CRPS is a measurement of error for probabilistic prediction and defined as the integrated squared difference between the CDF of forecasts and observations [*Murphy and Winkler*, 1987; *Hersbach*, 2000; *Huang et al.*, 2017]:

$$CRPS = \int_{-\infty}^{+\infty} [F^f(x) - F^o(x)]^2 dx \tag{23}$$

where $F^f$ and $F^o$ are CDFs for forecasts and observations, respectively. A small CRPS value indicates a better model performance, with the value of zero suggesting a perfect accuracy for model prediction.

Figure 3 presents the performance of the proposed CopPF approach for the simplified model. Since the model is quite basic, the CopPF can provide accurate predictions. Comparison among the proposed CopPF method, traditional particle filter (PF) method as well as particle Markov chain Monte Carlo (PMCMC) proposed by *Moradkhani et al.* [2012] is also presented in Figure 3. The results suggest that CopPF lead to narrower predictive intervals than PF and PMCMC. Table 1 shows the performance evaluation of CopPF, PF,
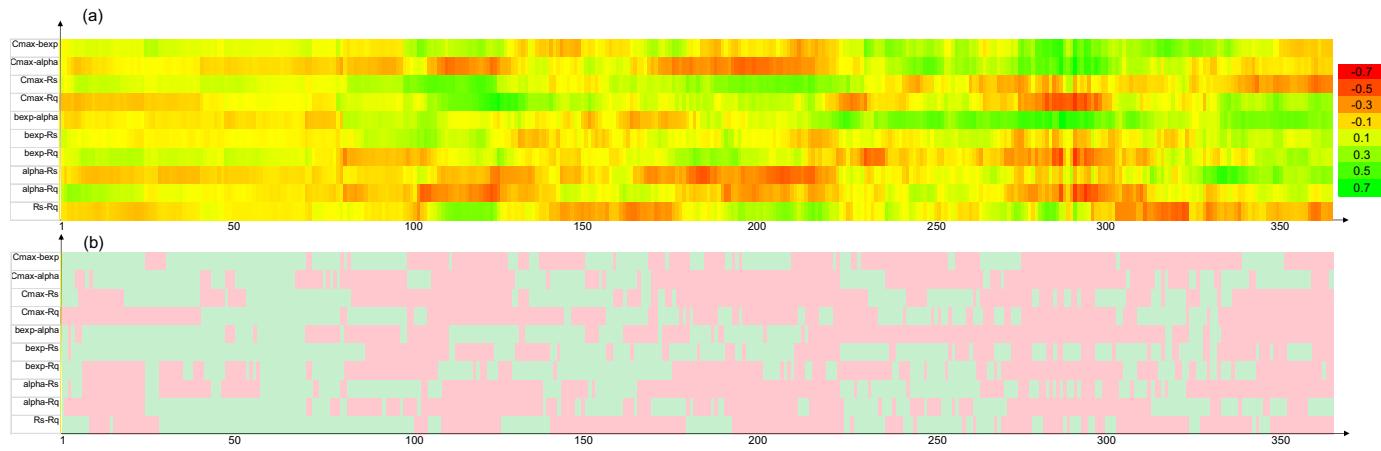
**Figure 9.** Parameter interdependence of Hymod measured through Kendall's τ values. (a) Temporal variations of parameter correlation during the data assimilation process through CopPF. (b) The associated *p*-values, with the light red indicating significant dependence with the *p*-value less than 0.05.
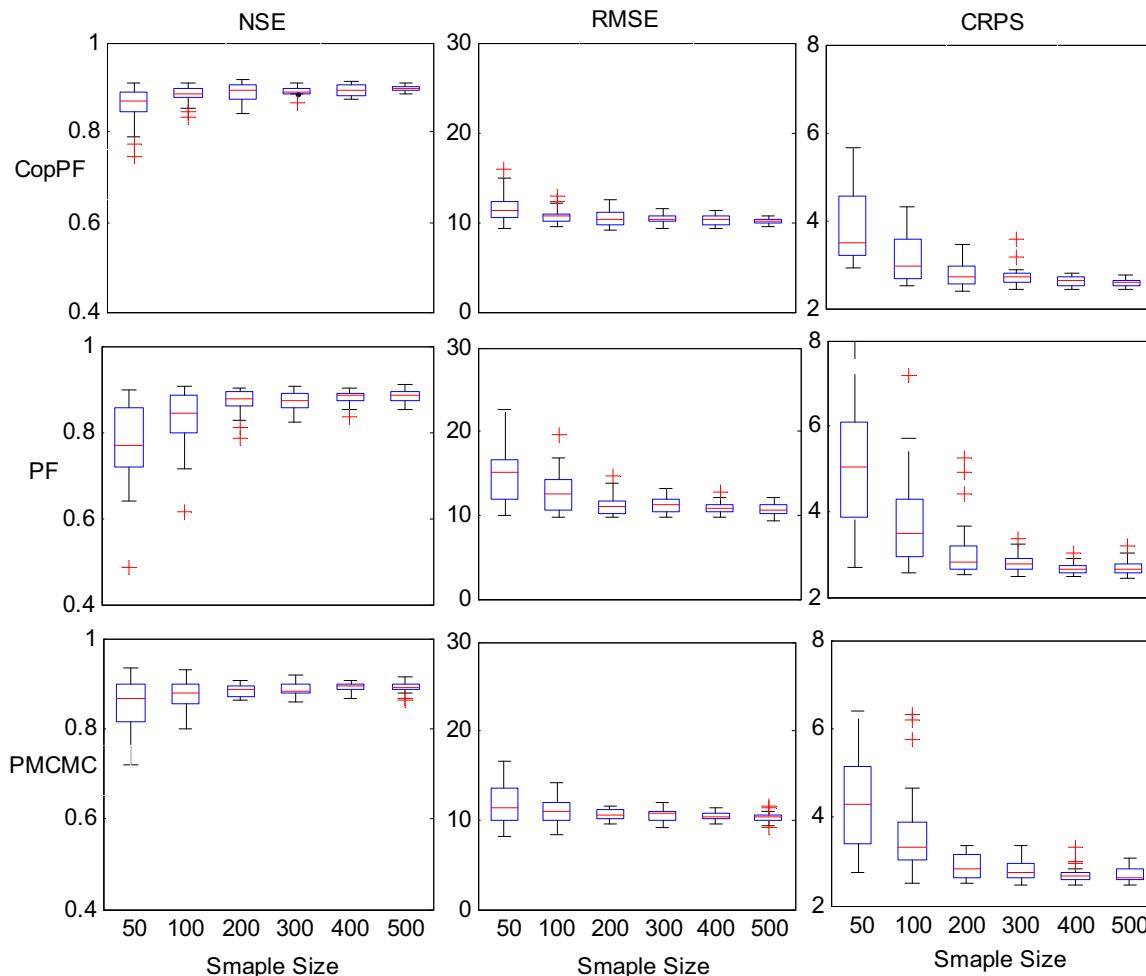


**Figure 10.** The boxplot of NSE, RMSE, and CRPS values for performance of CopPF, PF, PMCMC on Hymod under different sample size scenarios: six sample sizes {50, 100, 200, 300, 400, 500} are selected with 30 replicates run for each sample size scenarios. This boxplot indicates the reliability of the proposed CopPF approach. The CopPF can generally generate more reliable predictions than PF and PMCMC with a limited sample size.
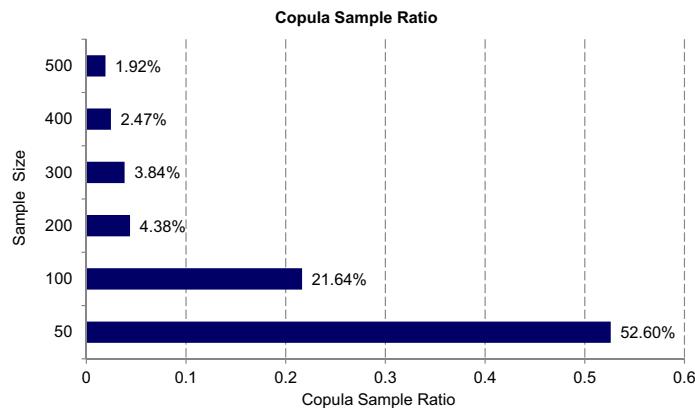
**Figure 11.** The ratio of samples obtained through copulas in the synthetic experiment of Hymod under different sample size scenarios: In each step of CopPF, two sample alternatives are available, namely sample importance resampling (SIR) and copula sample approach. This figure shows that under small sample size scenarios (≤100), a large portion of samples are generated through copulas.

and PMCMC on the toy model, which indicates the good performances of the three methods on such a basic model. However, the developed CopPF approach can achieve a slightly better performance in both deterministic and probabilistic predictions than PF and PMCMC.

To explore the dependence between the two parameters in the model, the values of Kendall's $\tau$ are generated in each data assimilation step, as plotted in Figure 4. The results exhibit that after a short warm-up period, the two parameters show statistically significant $(p < 0.05)$ correlation.

Moreover, such correlation is negative with the associated Kendall's $\tau$ value reaching $-0.7$. This negative correlation can be easily explained from the structure of the simplified model. Both the true values of $\alpha_1$ and $\alpha_2$ in the simplified model are positive, which suggests that, once one parameter is overestimated, the other one would get an underestimated value to balance the associated overestimation resulting from the former parameter. Furthermore, the results from Figure 4 suggest a demand to fully consider the correlated relationship between $\alpha_1$ and $\alpha_2$ to rejuvenate their evolution in a more reliable way.

Figure 5 shows the posterior distributions for $\alpha_1$ and $\alpha_2$, and their joint probability density function obtained by copulas. The above two rows indicate the posterior PDFs for $\alpha_1$ and $\alpha_2$ at different time steps, which are obtained by the kernel smoothing approach. The last row shows the joint PDF for $\alpha_1$ and $\alpha_2$. For this toy model, the most appropriate copula for modeling dependence between $\alpha_1$ and $\alpha_2$ is automatically selected through the R package "CDVine" [*Brechmann and Schepsmeier*, 2013] by using the Akaike information criterion (AIC). For instance, the joint PDF for $\alpha_1$ and $\alpha_2$ at time steps of 20, 50, 80, and 100 are plotted in Figure 4, in which the corresponding copulas are the rotated Joe copula (90°), the rotated BB7 copula (270°), the rotated BB1 copula (270°), and the rotated BB1 copula (270°), respectively.
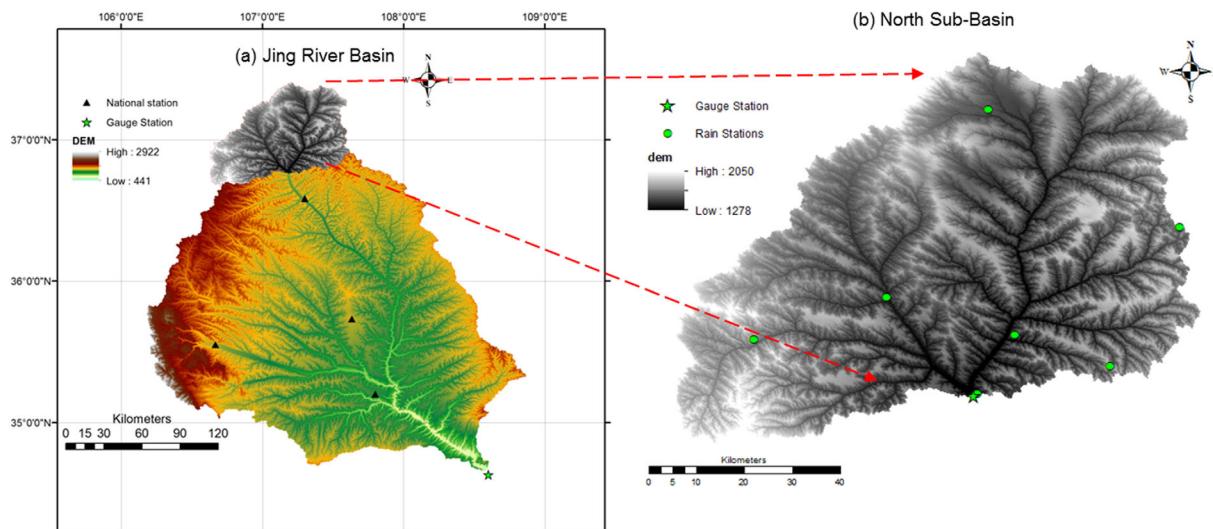


**Figure 12.** The location of the studied catchments: (a) the small catchment is located in the north part of Jing River basin and (b) the Jing River. The black triangles show the location of national meteorological stations used to generate the potential evapotranspiration, the blue cycles indicate the location of rainfall station used to generate areal precipitation, and the blue five-pointed star shows the location of the gauge station.

**Figure 13.** Comparison between predictions from CopPF, PF, PMCMC and synthetic observations from Hymod: the red stars indicate the synthetic observations while the cyan belt exhibits the 90% predictive intervals consisting of the 5% and 95% quantiles from the ensemble. The first 250 iterations are used for spin-up. The results show that the performance of CopPF is best with the NSE, RMSE, and CRPS values of 0.702, 10.782, and 3.159. The three evaluation indices from predictions of PF are 0.602, 12.456, and 3.213, respectively, and the indices of PMCMC predictions are 0.6444, 11.778, and 3.369 respectively. The computation time for CopPF, PF, and PMCMC are about 803, 31, and 50 s, respectively. The copula sample ratio of CopPF is 67.9%.

## 4. Case Study II: Hymod

### 4.1. Synthetic Experiment of Hymod

To further illustrate the applicability of the CopPF approach, CopPF is applied to a hydrologic data assimilation problem through the conceptual model Hymod, as shown in Figure 6a. Hymod is a probability-distributed model running for several time steps (e.g., minute, hour, daily) [*Moore*, 1985]. This model is composed of three components representing water storage and routing process in a catchment, in which a Pareto distribution is employed to reflect soil storage capacity, a slow-flow tank is used to route groundwater flow, and three identical quick flow tanks are used to route surface flow. This model has five unknown parameters, including the maximum storage capacity ($C_{max}$), spatial variability of soil moisture capacity ($b_{exp}$), partitioning factor between the two series of reservoir tanks ($\alpha$), residence time quick flow tank ($R_q$), and residence time of slow-tank ($R_s$), as well as two inputs including precipitation ($P$ (mm/d)) and potential evapotranspiration ($ET$ (mm/d)) [*Vrugt et al.*, 2005; *Remesan et al.*, 2014; *Fan et al.*, 2015b, 2016a]. The structure of Hymod is described in Figure 5. More details about Hymod can be found in a number of sources [*Boyle*, 2000; *Moore*, 2007; *Gharari et al.*, 2013].

A synthetic experiment is initially adopted to test the ability of the CopPF approach. Table 2 provides the "true" model parameters predefined before the synthetic experiment. This table also presents the initial fluctuation ranges for the five parameters before data assimilation. Stochastic perturbations are required in a data assimilation framework to account for the uncertainties in model inputs, parameters, and structures. In the synthetic experiment, uncertainties in the measurement of precipitation, potential evapotranspiration
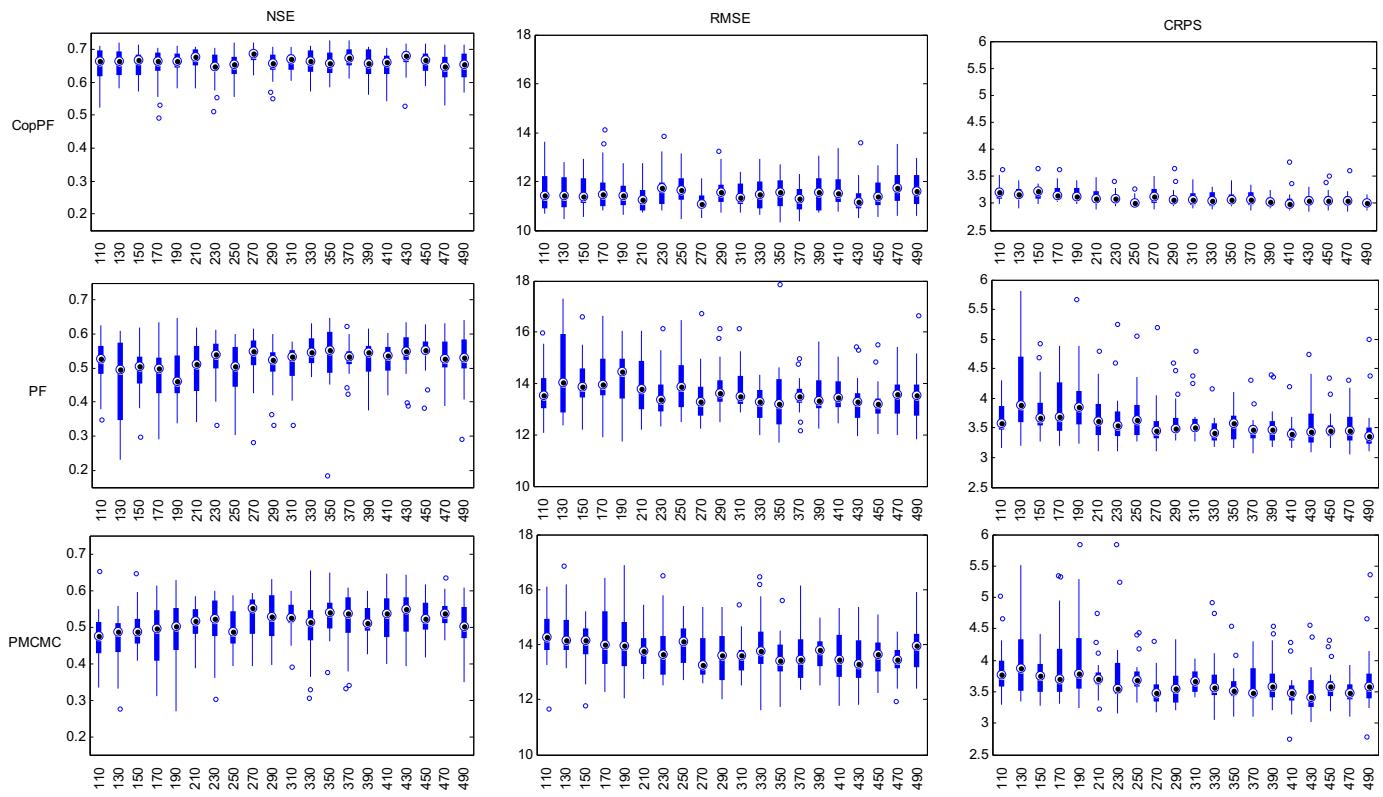
**Figure 14.** Performance comparison among CopPF, PF, and PMCMC on Hymod in different sample size scenarios: the sample size changes from 110 to 490 with an increasing value of 20 in each time. In each sample size scenarios, 20 replicates are run with the first 250 iterations being the spin-up period and the associated minimum, mean, and maximum and quantile values for NSE, RMSE, and CRPS are obtained based on these 20 replicates.

as well as streamflow observations are accounted for by adding random perturbations. For potential evapotranspiration and streamflow observations, Gaussian noise is recommended by a number of sources [e.g., *DeChant and Moradkhani*, 2012; *Moradkhani et al.*, 2012; *Chen et al.*, 2013; *Rasmussen et al.*, 2015]. For precipitation, some previous studies have applied Gaussian noise [e.g., *Rasmussen et al.*, 2015], while others have concluded that lognormal noise may perform better [e.g., *DeChant and Moradkhani*, 2012; *Moradkhani et al.*, 2012]. In this study, the lognormal noise is adopted for the synthetic experiment. The proportionality factors are assumed to be 0.2 for all measurements in the synthetic experiment.

In the synthetic experiment, the initial particle values for model parameters are randomly sampled from their uniform distributions as predefined in Table 2. The sample size for this experiment was initially set to be 100 and 1 year synthetic observations were used for assimilation into the hydrologic prediction process through the proposed CopPF method. Figure 7 shows the comparison of synthetic observations and predictions from CopPF, PF, and PMCMC. The results demonstrate that the 90% predictive intervals (consisting of 5% and 95% quantiles) from the three data assimilation schemes can generally bracket the synthetic observations and the associated predictive mean values can well track the fluctuation in the synthetic discharges. Figure 8 presents temporal evolution of the five parameters in Hymod (i.e., $C_{max}$, $b_{exp}$, $alpha$, $R_s$, $R_q$) in the data assimilation process through CopPF, PF, and PMCMC. The results indicate that all five parameters in Hymod are identified well after a number of data assimilation steps. A discharge series larger than 200 days appears to be sufficient to quantify parameter distributions with limited ranges. As shown in Figure 8, the proposed CopPF can rejuvenate particle evolution in wider fluctuation ranges than those obtained by PF and PMCMC, which suggests the potential of CopPF in mitigating sample impoverishment. However, in all the three data assimilation schemes, some parameters may not converge to their true values (e.g., $R_s$ for CopPF, $R_q$ for PF), and the estimations from PMCMC show apparent bias in this experiment.

In hydrologic models, their parameters are generally interrelated, and thus the optimal value of one parameter is dependent upon the values of other parameters. Similarly, such interrelationships among parameters
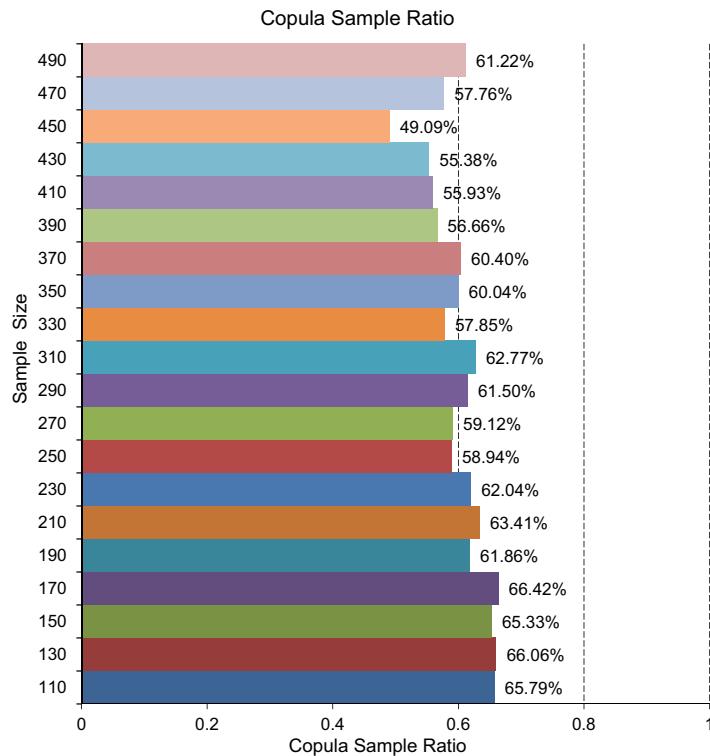
**Figure 15.** The resampling ratio through copula approach for Hymod: in all sample size scenarios, the copula resampling scheme makes a major contribution in the CopPF for the real-case study.

exist in the data assimilation process. Figure 9 presents the interrelationships among different pairs of parameters (denoted as $C_{max}$-$b_{exp}$, $C_{max}$-$\alpha$, $C_{max}$-$R_s$, $C_{max}$-$R_q$, $b_{exp}$-$\alpha$, $b_{exp}$-$R_s$, $b_{exp}$-$R_q$, $\alpha$-$R_s$, $\alpha$-$R_q$, and $R_s$-$R_q$) measured through Kendall $\tau$ values. As shown in Figure 9a, the correlation among different parameters is varied in the data assimilation process. Taking the interdependence of $C_{max}$-$b_{exp}$ as an example, these two parameters are generally positive-correlated in the former data assimilation steps, then develop a negative correlation structure in the middle period, and then return to a positive correlation structure. The correlation among other parameters also shows similar characteristics. This suggests that the five parameters in Hymod may be depending on each other, but the dependence structures of one pair of parameters are var-

ied in data assimilation process. Figure 9b presents the *p*-values associated with the Kendall's $\tau$ in Figure 9a, in which the *p*-value less than 0.05 suggests the dependence of two parameters is significant. The results indicate that the interdependence of the five parameters in Hymod is not always significant. In some data assimilation periods, all the five parameters are independent. However, the developed CopPF can also handle such conditions since resampling procedure through stochastic perturbation (expressed by equation (19)) is also involved in CopPF. When the interdependence among parameters is not significant, the particles would be rejuvenated by the stochastic perturbation method, which would be determined by equation (20). In this experiment, the values of Kendall's $\tau$ for $C_{max}$-$b_{exp}$, $C_{max}$-$\alpha$, $C_{max}$-$R_s$, $C_{max}$-$R_q$, $b_{exp}$-$\alpha$, $b_{exp}$-$R_s$, $b_{exp}$-$R_q$, $\alpha$-$R_s$, $\alpha$-$R_q$, and $R_s$-$R_q$ would be 0.04, $-0.11$, 0.06, $-0.01$, $-0.06$, $-0.01$, 0.09, $-0.07$, $-0.09$, and $-0.06$, respectively, at $t = 62$, which shows independent structures among those parameters with all *p*-values larger than 0.05. At this time step, the particles resampled by the stochastic perturbation method lead to a data mismatch index value of 258.81, while the particles resampled by copulas generate a value of 884.60 for such an index. This indicates that the particles resampled by the stochastic perturbation method would be chosen at this time step.

To further demonstrate the robustness of the proposed method, all the three data assimilation schemes were tested for the synthetic experiment under a number of sample size scenarios. In detail, six sample size scenarios, {50, 100, 200, 300, 400, 500}, were selected, in which 30 replicates were conducted under each sample size scenario. Figure 10 presents the boxplot exhibiting the variations of NSE, RMSE, and CRPS values under different sample size scenarios. It indicates that the performance of CopPF is generally much reliable with limited fluctuating ranges for all the three indices when the sample size is larger than a few hundred values. For instance, as presented in Figure 10, the generated NSE values are approximately larger than 0.8 for all replicates when the sample size is greater than one hundred. For the probabilistic forecasts obtained from CopPF, the associated CRPS would be less than 4.5 for all replicates when the sample size is equal or larger than 100. For PF and PMCMC, they generally provide better forecasts when the sample size increases. However, the performance of PF and PMCMC shows significant variation for low sample size scenarios, as presented in Figure 10.

**Table 4.** Parameter Description of SAC-SMA

| Parameter | Description | Unit | Range | Synthetic True Value |
|---|---|---|---|---|
| UZTWM | Upper zone tension-water maximum storage | mm | [5, 200] | 60.3 |
| UZFWM | Upper zone free-water maximum storage | mm | [5, 120] | 70.3 |
| UZK | Upper zone free water lateral drainage rate | day$^{-1}$ | [0.1, 0.75] | 0.49 |
| PCTIM | Impervious fraction of the watershed area (decimal fraction) | | [0, 0.1] | 0.02 |
| ADIMP | Additional impervious area (decimal fraction) | | [0, 0.5] | 0.3 |
| ZPERC | Maximum percolation rate (dimensionless) | | [5, 500] | 289.0 |
| REXP | Exponent of the percolation equation (dimensionless) | | [1, 7] | 3.4 |
| LZTWM | Lower zone tension water maximum storage | mm | [50, 500] | 282.0 |
| LZFSM | Lower zone supplemental free water maximum storage | mm | [15, 300] | 181.0 |
| LZFPM | Lower zone primary free water maximum storage | mm | [20, 600] | 455.3 |
| LZSK | Lower zone supplemental free water lateral drainage rate | day$^{-1}$ | [0.03, 0.2] | 0.06 |
| LZPK | Lower zone primary free water lateral drainage rate | day$^{-1}$ | [0.001, 0.015] | 0.004 |
| PFREE | Fraction of water percolating from upper zone directly to lower zone free water (decimal fraction) | | [0.001, 0.5] | 0.16 |
| RIVA | Riverside vegetation area (decimal fraction) | | 0.3 | 0.3 |
| SIDE | Ratio of deep recharge to channel base flow (dimensionless) | | 0 | 0 |
| RSERV | Fraction of lower zone free water not transferrable to lower zone tension water (decimal fraction) | | 0 | 0 |

Table 3 gives the minimum, maximum, and mean values of NSE, RMSE and CRPS for CopPF, PF, and PMCMC performance under different sample size scenarios. The results suggest that reliable predictions can be generated through CopPF with a sample size larger than 100. Moreover, optimal performance can be achieved by CopPF with relatively modest sample sizes. As shown in Table 3, CopPF can reach an optimal performance for the synthetic experiment with a sample size over 100, reaching mean values of NSE, RMSE, and CRPS of 0.883, 10.757, and 3.144, respectively. Such a performance will not change significantly as the sample size increases. For instance, the mean NSE values are 0.889, 0.893, 0.894, and 0.897 for the sample size of 200, 300, 400, and 500, respectively. Furthermore, compared with PF and PMCMC, the proposed CopPF can achieve lowest mean values in NSE, RMSE, and CRPS for all sample scenarios, which suggests the best predictions for both deterministic and probabilistic cases.

In CopPF, there are two potential resampling schemes, namely the traditional SIR (i.e., stochastic perturbation expressed by equation (19)) approach and the copula sampling method. At each time step, the particles are resampled through these two schemes and the better sample set is chosen as the posterior estimation for the current stage. The proportion of time steps using the copula sampling scheme is plotted in Figure 11. It is indicated that more than 52% of the time steps (approximately 190 days) used the copula resampling scheme in the synthetic experiment under a sample size of 50. Such a proportion decreases as the sample size increases. In the extreme scenario with a sample size of 500, the copula resampling scheme contributes less than 2% in the data assimilation process. This means that the copula resampling scheme is more applicable for limited sample size scenarios in which serious sample impoverishment occurs. In terms of this synthetic experiment, the SIR scheme is sufficient to keep considerable sample diversity when 500 samples are adopted and only a few steps are needed for the copula resampling procedure.

### 4.2. Real-World Application for a Small-Scale Watershed

The developed CopPF approach is first tested for hydrologic data assimilation in a catchment located in the north part of the Jing River basin in China. Figure 12b shows the location of the studied catchment, which has a drainage area of 4640 km$^2$ and two main tributaries converging together at Hongde station (107.19°E, 36.76°N). In general, the Jing River basin (shown in Figure 12a) has a semiarid and subhumid continental monsoon climate, resulting in considerable temporal-spatial variations in precipitation [*Xu et al.*, 2017]. From the north to south part of the basin, the annual precipitation ranges from 240 to 710 mm, with approximate 50–60% precipitation occurring in the Summer and Fall seasons. In particular, the studied catchment in this case is located in the northern part of the Jing River basin, and the annual precipitation there fluctuates from 240 to 350 mm with a mean annual precipitation of approximately 309 mm. The areal precipitation data for Hymod were interpolated from site precipitation (denoted as P) measurements
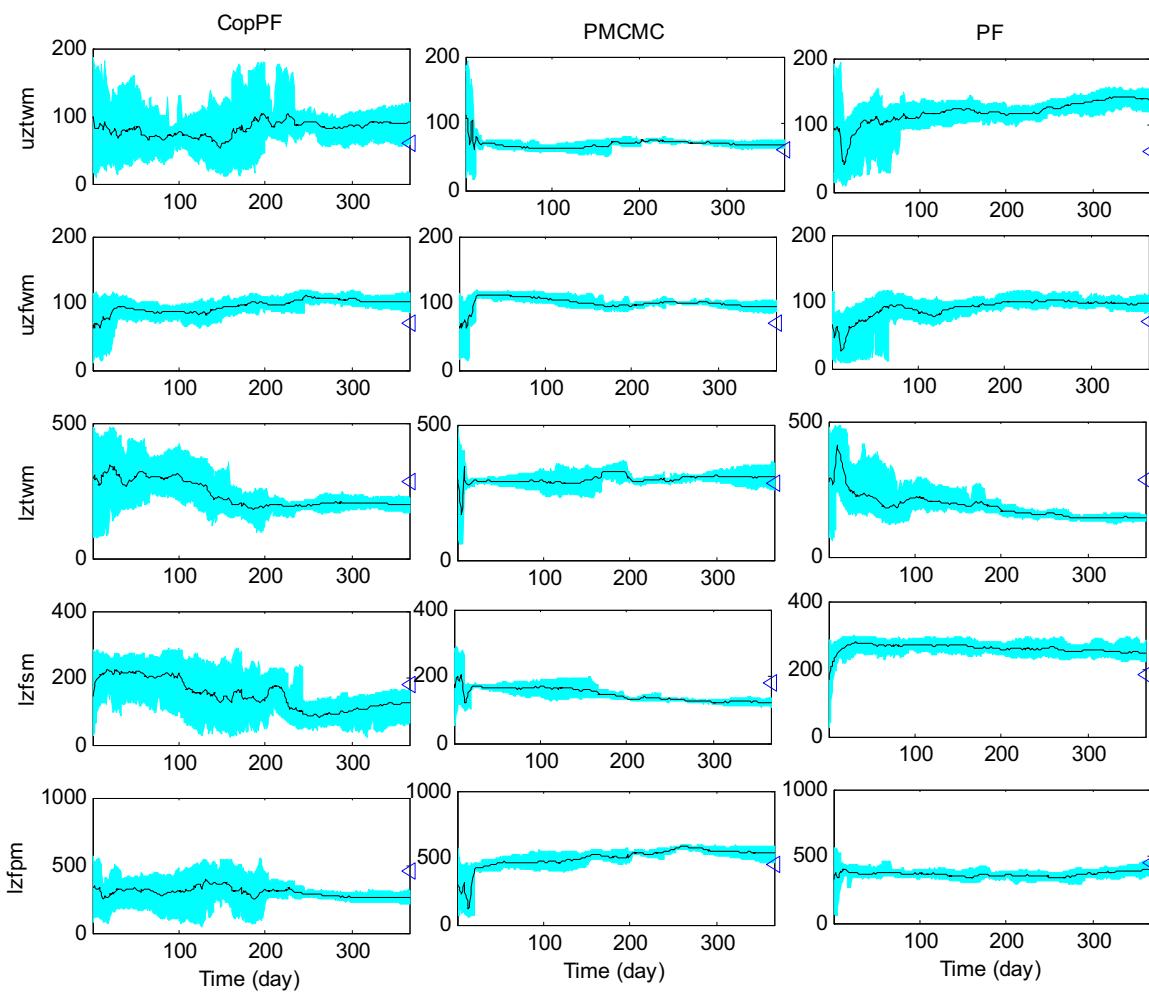
**Figure 16.** Parameter evolution in the data assimilation process through CopPF, PF, and PMCMC for the SAC-SMA model: the cyan region exhibits the 90% predictive uncertain ranges (bracketed by the 5% and 95% quantiles). The black line denotes the predictive means and the blue triangles at the right-hand side suggest the actual parameter values.

distributed over the catchment, and the areal potential evapotranspiration (denoted as ET) data were interpolated from the ET data at the national meteorological stations in the Jing River basin. The streamflow observations at Hongde station were used in this case study.

In this real-world case study, Gaussian noise with a mean value of 0 and 20% relative error was introduced into the potential evapotranspiration to account for uncertainties in this kind of forcing data. Random perturbation was also added to the precipitation, which is assumed to be lognormally distributed with a 20% relative error. Also, Gaussian noise was added to the streamflow observations with a relative error of 20% of the actual observations. The sample size for this case study was set at 100.

Figure 13a presents the comparison between 90% predictive intervals from the CopPF method and real-world streamflow discharge data. It can be observed that the obtained predictive intervals can generally match the observations, except for some underestimations in high-flow periods. Moreover, traditional PF with SIR and particle Markov chain Monte Carlo (PMCMC) methods proposed by *Moradkhani et al.* [2012] are applied to this case study and the results are plotted in Figures 13b and 13c. The results indicate that the predictions from PF and PMCMC can also fit observations well under low flow periods. At high-flow periods, both PF and PMCMC provide underestimations. Even though all three data assimilation schemes provide underestimations at high-flow periods, the proposed CopPF can generally provide better predictions than the PF and PMCMC approaches. To further compare performances of CopPF, PF, and PMCMC, all the three approaches would be run under different sample size scenarios. In detail, the sample size changes from 110 to 490 with an increasing value of 20 in each time and 20 replicates are run for each
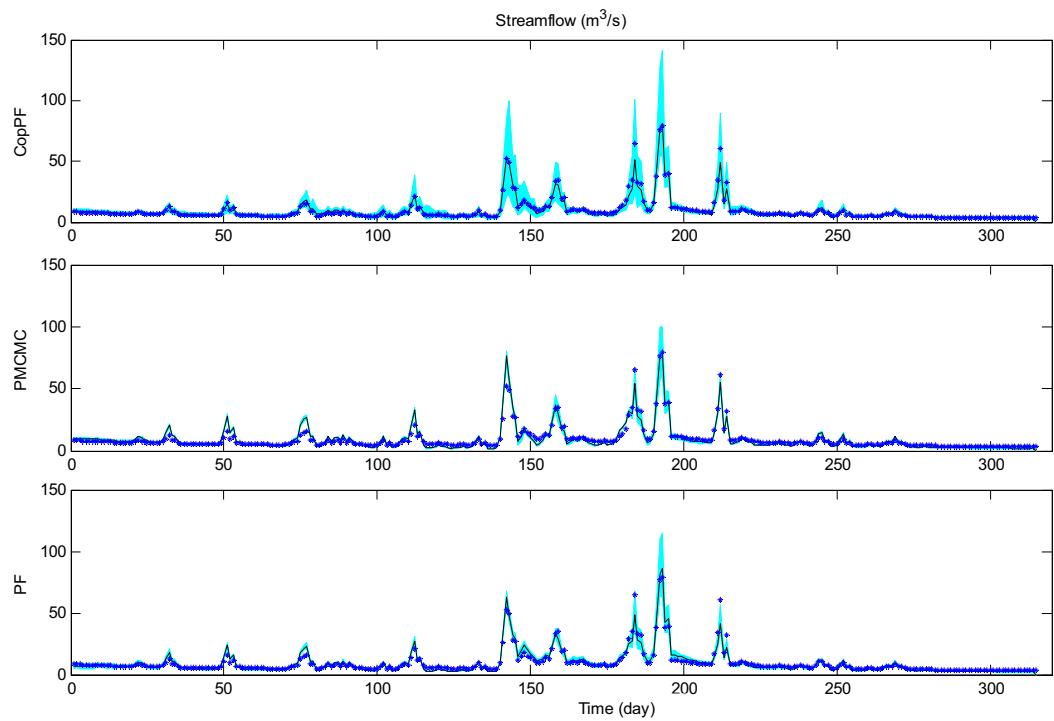
**Figure 17.** Comparison between predictions from different data assimilation schemes and synthetic true observations from SAC-SMA model: The blue stars indicate the true values while the black line shows predictive means. The cyan belt exhibits the 90% predictive intervals consisting of the 5% and 95% quantile values. Such results are obtained by CopPF, PMCMC, and PF with a sample size of 100.

sample scenario with the first 250 steps being the spin-up period. Also, these replicates are run based on similar initial conditions. The corresponding minimum, maximum, and quantile values for NSE, RMSE, and CRPS are generated based on those replicates and plotted in Figure 14. The results show that the performance of CopPF is generally better than PF and PMCMC for nearly all sample size scenarios. Moreover, the CopPF approach will generate more reliable predictions than PF and PMCMC in this case, with the corresponding NSE values larger than 0.5, RMSE values less than 14 and CRPS values less than 4 for all sample size scenarios. In comparison, the performances of PF and PMCMC vary significantly even under large sample size scenarios.

In the previous synthetic experiment, most particles were evolved through the traditional SIR approach with about 53% particles being generated through the copula sampling method for the sample size of 50. In comparison, most particles were resampled through the copula sampling approach in the real-world case study, as presented in Figure 15. In all sample size scenarios, parameter evolution was mainly conducted through the copula resampling algorithm, evolving more than 65% time steps for low sample size scenarios. As the sample size increases, the SIR algorithm makes a greater contribution in CopPF, but the copula resampling approach still dominates parameter evolution. For instance, in those scenarios with sample sizes larger than 400, more than 50% of the parameter evolution steps are still performed by the copula resampling method except for the sample size of 450. Compared with the results from the synthetic experiment, the copula resampling method makes greater contributions in the real-world case study. This phenomenon happens mainly because, for the real-case application, a large discrepancy between particle ensembles and the respective observations may frequently occur. The traditional SIR procedure can merely evolve each parameter in its adjacent area separately, which may not lead to preferred particles with better matching observations. The copula resampling approach infers a new set of parameters by

**Table 5.** Performance of CopPF, PF, and PMCMC on Synthetic Case of SAC-SMA Model

|  | NSE | RMSE | CRPS |
|---|---|---|---|
| CopPF | 0.9721 | 1.6530 | 0.7197 |
| PF | 0.9274 | 2.6672 | 1.6877 |
| PMCMC | 0.8889 | 3.2991 | 1.3755 |

**Table 6.** Performance Comparison of CopPF, PMCMC, and PF on the SAC-SMA Model at Jing River

| Sample Size | Method | NSE | RMSE | CRPS |
|---|---|---|---|---|
| 100 | PF | 0.7082 | 45.8617 | 24.4407 |
| | PMCMC | 0.7137 | 46.2676 | 24.1280 |
| | CopPF | 0.7847 | 40.1186 | 21.4039 |
| 200 | PF | 0.6722 | 48.6064 | 23.1844 |
| | PMCMC | 0.6770 | 49.1448 | 24.5307 |
| | CopPF | 0.7641 | 41.9960 | 19.6568 |
| 300 | PF | 0.6585 | 49.6114 | 24.0295 |
| | PMCMC | 0.7575 | 42.5775 | 20.6561 |
| | CopPF | 0.8158 | 37.1033 | 16.9905 |
| 400 | PF | 0.7326 | 43.8964 | 16.4880 |
| | PMCMC | 0.7825 | 40.3263 | 20.4384 |
| | CopPF | 0.7526 | 43.0046 | 24.0494 |

considering interrelationships among these parameters and offers greater opportunity to access new samples that lead to a better match between predictions and observations.

## 5. Case Study III: Sacramento Soil Moisture Accounting Model (SAC-SMA)

### 5.1. Synthetic Experiment of SAC-SMA

The SAC-SMA model will be further applied for testing the applicability of the proposed CopPF approach. The SAC-SMA model, introduced by *Burnash et al.* [1973], is a physically based conceptual model and is used by the National Weather Service (NWS) for streamflow prediction in U.S. As presented in Figure 6b, this model conceptualizes water storage with two vertical zones: (i) the upper zone accounting for the short-term storage of water in soil and (ii) the lower zone representing the longer-term groundwater storage [*DeChant and Moradkhani*, 2012]. Each zone has free water and tension water components, in which free water is dominantly driven by gravitation forces but may also be depleted by evapotranspiration, percolation, and horizontal flow, while the tension water (slow-flow component) is driven by evapotranspiration and diffusion [*Hirpa*, 2013]. This model has six state variable, including the upper zone tension water content (UZTWC), the upper zone free water content (UZFWC), the lower zone tension water content (LZTWC), the lower zone free primary content (LZFPC), the lower zone free supplemental content (LZFSC), and the total tension water content including impervious area (ADIMC). Sixteen parameters are involved in the SAC-SMA model and Table 4 summarizes the description, variation range as well as the synthetic true value to generate the synthetic streamflow and soil moisture for each parameter. Thirteen parameters are considered for data assimilation with the other three ones fixed at prespecified values according to some literates [*Gan et al.*, 2014; *Wang et al.*, 2016]. Similar to the synthetic experiment of Hymod, Gaussian noise is employed to account for uncertainties in evapotranspiration and the synthetic "true" observations, while lognormal noise is used for precipitation. A proportionality factor value of 0.2 is used for all measurements in the synthetic experiment.

In this synthetic case, the initial particles for the parameters of SAC-SMA are sampled uniformly from their prespecified intervals presented in Table 4. The sample size for this experiment is set to be 100. One year synthetic observations are used for data assimilation with the first 50 step used as the spin-up period. Figure 16 compares temporal evolutions of UZTWM, UZFWM, LZTWM, LZFSM, and LZFPM in the data assimilation process by CopPF, PMCMC, and PF. It can be observed that the CopPF can rejuvenate particle evolution in wider fluctuation ranges than those obtained by PF and PMCMC, which can lead to less opportunities for sample impoverishment. Such a conclusion is consistent with that obtained by the synthetic experiment of Hymod. The comparison between ensemble predictions and synthetic observations is shown in Figure 17. The results conclude that CopPF would produce 90% predictive intervals with larger uncertainties than those from PMCMC and PF. This is mainly due to the wider spread ranges for model parameters as shown in Figure 16. However, as shown in Table 5, the results of NSE, RMSE, and CRPS reveal that the proposed CopPF can provide better results for both probabilistic and deterministic predictions than PMCMC and PF.

### 5.2. Real-World Case for a Large-Scale Watershed

The applicability of the developed CopPF approach is further demonstrated through hydrologic data assimilation in a large-scale watershed by SAC-SMA model. As presented by Figure 12a, the Jing River
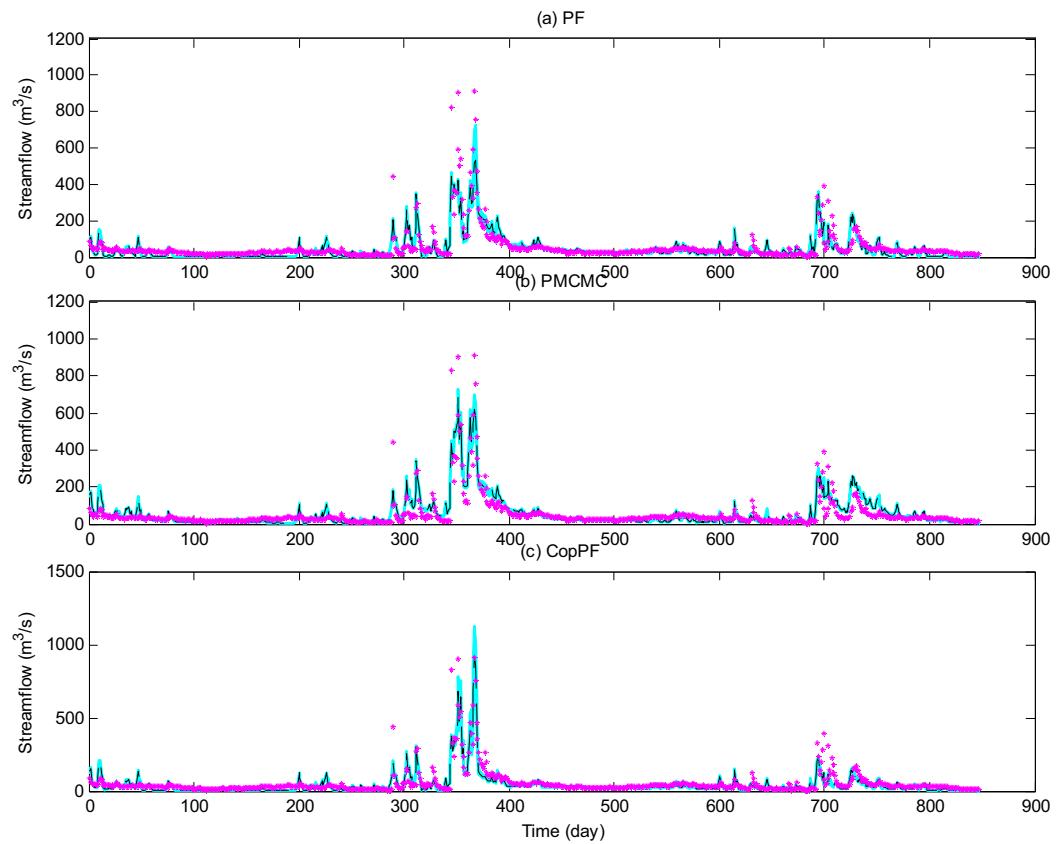
**Figure 18.** Comparison between predictions from different data assimilation schemes and real observations at Zhangjiashan station at Jing River basin: the Fuchsia stars indicate the true values while the black dashed line shows predictive means. The cyan belt exhibits the 90% predictive intervals consisting of the 5% and 95% quantile values. Such results are obtained by CopPF, PMCMC, and PF with a sample size of 100 and the first 250 being the spin-up period. The NSE, RMSE, and CRPS values for PF predictions are 0.708, 45.862, and 24.441. The three indices for PMCMC predictions are 0.714, 46.268, and 24.128. Those evaluation indices for CopPF are 0.785, 40.119, and 21.404. The computation time for PF, PMCMC, and CopPF are 77.4, 115.3, and 1360.7 s. The copula sample ratio of CopPF is 57.1%.

is one of the main tributary of the Weihe River, which is the largest tributary of Yellow River. It is located in the middle of the Loess Plateau (100.9°E–114.55°E and 33.72°N–41.27°N), with a drainage area of 45,421 km$^2$. In this study, the inputs of the SAC-SMA model are obtained based on meteorological measurements at the national stations (the black triangles in Figure 12a), while the streamflow measurements at Zhangjiashan station (108.60°E and 34.63°N) are used for data assimilation by CopPF, PMCMC, and PF approaches. Similar to its catchment in the northern part, Gaussian noise with a mean value of 0 and 20% relative error was introduced into the potential evapotranspiration and streamflow observations to account for their inherent uncertainties in this kind of forcing data, while the precipitation is assumed to be lognormally distributed with a 20% relative error. The sample size for this case study is set at 100 and observations of 3 years are employed with the first 250 iterations used for spin-up of the data assimilation methods.
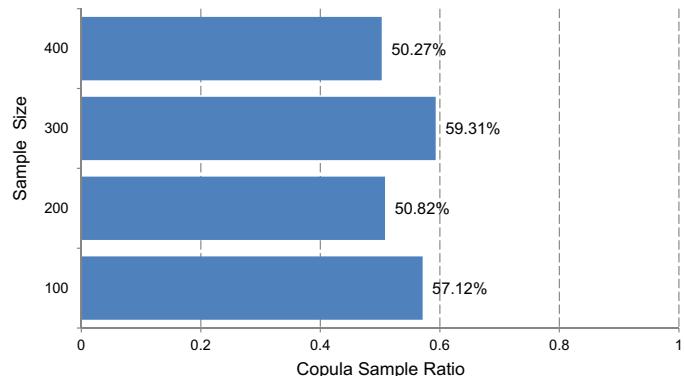


**Figure 19.** The resampling ratio through copula approach for SAC-SMA: in all sample size scenarios, the copula resampling scheme makes a major contribution for hydrological data assimilation in Jing River by SAC-SMA.
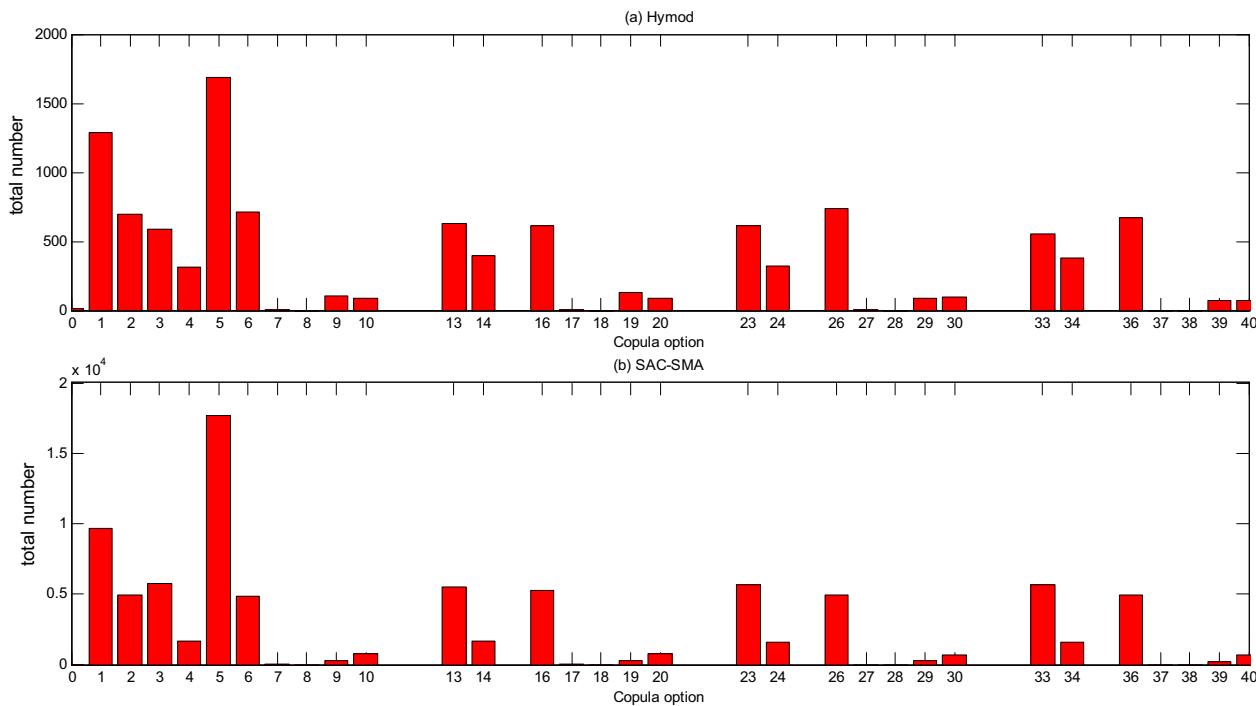
**Figure 20.** Copula options in the data assimilation by Hymod and SAC-SMA hydrological models: the correspondence between xlabel values and copulas can be found in *Brechmann and Schepsmeier* [2013]. In detail, for both Hymod and SAC-SMA models, Gaussian copula (i.e., $x = 1$) and Frank copula (i.e., $x = 5$) are used most frequently.

Figure 18 shows performance of ensemble predictions from PF, PMCMC, and CopPF. It is noticed that for all three approaches provide accurate predictions during the low-flow periods. However, for high-flow periods, both PF and PMCMC generate underestimations, while the proposed CopPF approach can offer better predictions. Moreover, it is observed that 90% predictive intervals from CopPF are wider than those from PF and PMCMC, especially at high-flow periods. This is due to the fact that the wider spreads of parameter posterior distributions from CopPF, as presented in previous synthetic experiments. To further characterize the impacts of sample sizes on those three methods, three more sample scenarios are chosen, including 200, 300, and 400. As shown in Table 6, the proposed CopPF can generate reliable results for both deterministic and probabilistic predictions at low sample size scenarios, and the sample size would not influence the performance of CopPF significantly. Such a conclusion is consistent with the one from the first real case by Hymod. In comparison, larger sample may be required for PF and PMCMC to achieve good estimations. Moreover, the particles are mainly rejuvenated by the copula resampling procedure for data assimilation by the SAC-SMA model for Jing River basin, as presented in Figure 19. This is also consistent with the conclusion for Hymod in the small-scale watershed case.

## 6. Discussion

As illustrated in the case of toy model, the choice of copulas for describing the dependence among parameters may be varied during the data assimilation process. This is also true for the application of CopPF for hydrological data assimilation by Hymod and SAC-SMA models. Figure 20 presents copula options of CopPF on the two real cases. It can be found that the copula choosing patterns are quite similar between Hymod and SAC-SMA model. For both cases, Frank copula (i.e., x = 5) would be chosen most frequently, followed by the Gaussian copula (i.e., x = 1). This may because that the parameters in these two copulas have least restriction. For instance, the parameter in a Frank copula can be any value except zero, while the parameter in a Gaussian copula is described by a specific correlation coefficient varied in [–1, 1]. In comparison the parameter in a Gumbel (i.e., x = 4) should larger than 1, which means that it can only be applicable for positive dependence structures (the parameter $\theta$ in a Gumbel copula can be estimated by $\theta = 1/(1 - \tau)$, where $\tau$ is the Kendall's correlation coefficient).
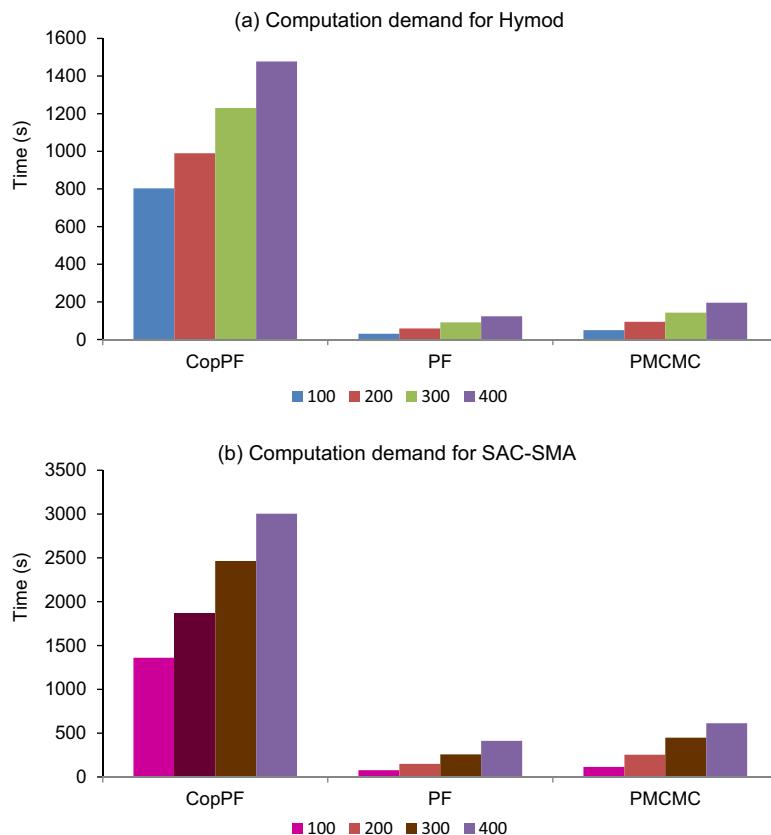
**Figure 21.** Computational demand for different methods on different models.

One major issue of the proposed CopPF approach is its computation demand, since resampling by copulas need additional computation for parameter estimation for both marginal distributions and dependence structures. Such procedures would be extensive time consuming if the best copula combination is determined and then the associated model parameters are estimated by maximum likelihood estimation (MLE) method. For instance, if the above procedures (i.e., include copula selection and parameter estimation by MLE) are applied for Hymod, the computation demand would be about 1908.0 (s) for a sample size of 100, and this would increase to 5993.5 (s) for a sample size of 400. For the SAC-SMA model, much more time is required (about $1.2 \times 10^4$ (s) for a sample size of 100, and $4.2 \times 10^4$ (s) for a sample size of 400) since for SAC-SMA with 13 parameters, 78 bivariate copulas should be identified at each time step. Consequently, we recommend to use prespecified copula in the developed CopPF. The Frank copula is used in this study for three reasons: (i) as presented in Figure 20, Frank copula is used frequently for both Hymod and SAC-SMA models and (ii) Frank copula can model both positive and negative structures; (iii) Frank copula is a kind of Archimedean copula and its parameter can be estimated by inversion of Kendall's $\tau$, which is much faster than MLE. Also, using specific copula is acceptable for CopPF, since particles can be rejuvenated by traditional SIR scheme if samples from the Frank copula are unreasonable. Figure 21 presents the computation demands for CopPF, PF, and PMCMC methods for different sample size scenarios, in which CopPF is implemented with the Frank copula. The results show that CopPF needs more computation time than PF and PMCMC. However, compared with CopPF including copula selection, using a specific copula can significantly reduce time consuming of CopPF, especially for the SAC-SMA model. Even though more computational demand is required for implementation of CopPF, it can still applicable since, as illustrated by the two real-case studies, the developed CopPF can improve particle impoverishment and achieve reliable predictions with low sample sizes. Moreover, the sample size does not influence the performance of CopPF significantly. This is meaningful for those models with limited parameters but extensive computational demand.

## 7. Conclusions

In this study, a copula-based particle filter (CopPF) approach has been proposed to improve performance of particle filter (PF) in hydrological data assimilation by introducing copulas to describe parameter interdependence. In data assimilation through CopPF, two resampling schemes are used for particle rejuvenation, in which one set of samples is resampled by traditional sample importance resampling (SIR) procedure and the other set is generated by copula resampling method. A data mismatch index is then employed to determine which set is used for the posterior estimation for current step. The introduction of copula resampling method can generate new particles which are different from the old ones, and thus increase the diversity of the particles and mitigate sample impoverishment.

To demonstrate the applicability of the developed CopPF method, three kinds of models with different parameter dimensions have been considered for simple regression, as well as data assimilation in both small-scale and large-scale watersheds. Comparison with PF and particle Markov chain Monte Carlo (PMCMC) has also been performed through evaluation indices of root-mean-square error (RMSE), the Nash-Sutcliffe efficiency (NSE) coefficient, and continuous ranked probability score (CRPS). Based on those case studies, some findings can be concluded:

1. In data assimilation, posterior estimations for parameters can be correlated among each other, and for one pair of parameters, their correlation structure would be varied and may be positive or negative in data assimilation process.
2. The proposed CopPF can rejuvenate particle evolution in wider fluctuation ranges than those obtained by PF and PMCMC, and thus lead to less chance for particle degeneracy and impoverishment.
3. The CopPF can provide more accurate results for both deterministic and probabilistic predictions than PF and PMCMC at low sample size scenarios. Moreover, the sample size does not influence the performance of CopPF significantly.
4. The computational time may be an issue for implementation of CopPF. But using prespecified copulas can dramatically reduce the computational demand of CopPF. We recommend to adopting the Frank copula in implementation of CopPF since it can describe both positive and negative correlation structures and has least restriction in its parameter estimation.

This study is the first attempt to incorporate copula methods into sequential hydrologic data assimilation to fully describe parameter multidimensional correlation structures, and then generate new samples based on the knowledge of parameter interdependence. The case studies for one simplified regression and two real-world hydrologic data assimilation problems have demonstrated the significant potential of the proposed CopPF approach. The results conclude that the developed CopPF can provide accurate predictions with low sample size scenarios. However, the impacts of uncertainties in model inputs and streamflow observations may have both individual and joint effects on the performance of the data assimilation schemes. Further research work is required to characterize impacts from uncertainties of inputs and streamflow measurements and reveal the main contributors for impacting CopPF's performance.

## References

Aas, K., C. Czado, A. Frigessi, and H. Bakken (2009), Pair-copula constructions of multiple dependence, *Insurance Math. Econ.*, *44*, 182–198.

Arulampalam, M. S., S. Maskell, N. Gordon, and T. Clapp (2002), A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking, *IEEE Trans. Signal Process.*, *50*, 174–188.

Bárdossy, A., and S. Hörning (2016), Gaussian and non-Gaussian inverse modeling of groundwater flow using copulas and random mixing, *Water Resour. Res.*, *52*, 4504–4526, doi:10.1002/2014WR016820.

Beadle, E. R., and P. M. Djuric (1997), A fast-weighted Bayesian bootstrap filter for nonlinear model state estimation, *IEEE Trans. Aerospace Electron. Syst.*, *33*(1), 338–343.

Bolic, M., P. M. Djuric, and S. Hong (2004), Resampling algorithms for particle filters: A computational complexity perspective, *EURASIP J. Appl. Signal Process.*, *15*, 2267–2277.

Boyle, D. P. (2000), Multicriteria calibration of hydrologic models, PhD dissertation, Dep. of Hydrol. and Water Resour., Univ. of Ariz., Tucson.

Brechmann, E. C., and U. Schepsmeier (2013), Modeling dependence with C- and D-vine copulas: The R package CDVine, *J. Stat. Software*, *52*(3), 1–27.

Burnash, R. J. C., R. L. Ferral, and R. A. McGuire (1973), *A Generalized Streamflow Simulation System: Conceptual Modeling for Digital Computers*, U.S. Dep. of Commer., Natl. Weather Serv., Sacramento, Calif.

Chen, H., D. Yang, Y. Hong, J. J. Gourley, and Y. Zhang (2013), Hydrological data assimilation with the Ensemble Square-Root-Filter: Use of streamflow observations to update model states for real-time flash flood forecasting, *Adv. Water Resour.*, *59*, 209–220.

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

DeChant, C., and H. Moradkhani (2012), Examining the effectiveness and robustness of sequential data assimilation methods for quantification of uncertainty in hydrologic forecasting, *Water Resour. Res.*, *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, *J. Hydrol.*, *519*, 2967–2977.

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

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

Fan, Y. R., G. H. Huang, K. Huang, and B. W. Baetz (2015a), Planning water resources allocation under multiple uncertainties through a generalized fuzzy two-stage stochastic programming method, *IEEE Trans. Fuzzy Syst.*, *23*(5), 1488–1504.

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

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

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

Fan, Y. R., W. W. Huang, G. H. Huang, Y. P. Li, K. Huang, and Z. Li (2016b), Hydrologic risk analysis in the Yangtze River basin through coupling Gaussian mixtures into copulas, *Adv. Water Resour.*, *88*, 170–185.

Fan, Y. R., W. W. Huang, G. H. Huang, K. Huang, Y. P. Li, and X. M. Kong (2016c), Bivariate hydrologic risk analysis based on a coupled entropy-copula method for the Xiangxi River in the Three Gorges Reservoir area, China, *Theor. Appl. Climatol.*, *125*(1), 381–397.

Fan, Y. R., G. H. Huang, B. W. Baetz, Y. P. Li, K. Huang, X. Chen, and M. Gao (2017), Development of integrated approaches for hydrological data assimilation through combination of ensemble Kalman filter and particle filter methods, *J. Hydrol.*, *550*, 412–426.

Gan, Y., Q. Duan, W. Gong, C. Tong, Y. Sun, W. Chu, A. Ye, C. Miao, and Z. Di (2014), A comprehensive evaluation of various sensitivity analysis methods: A case study with a hydrological model, *Environ. Modell. Software*, *51*, 269–285.

Gharari, S., M. Hrachowitz, F. Fenicia, and H. H. G. Savenije (2013), An approach to identify time consistent model parameters: Sub-period calibration, *Hydrol. Earth Syst. Sci.*, *17*, 149–131.

Gordon, N. J., D. J. Salmond, and A. F. M. Smith (1993), Novel approach to nonlinear/non-Gaussian Bayesian state estimation, *IEEE Proc. F: Radar Signal Process.*, *140*(2), 107–113.

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

Hersbach, H. (2000), Decomposition of the continuous ranked probability score for ensemble prediction systems, *Weather Forecasting*, *15*, 559–570.

Hirpa, F. A. (2013), Hydrologic data assimilation for operational streamflow forecasting, PhD dissertation, Univ. of Conn., Storrs, Conn.

Huang, C. C., A. J. Newman, M. P. Clark, A. W. Wood, and X. Zhang (2017), Evaluation of snow data assimilation using the ensemble Kalman filter for seasonal streamflow prediction in the western United States, *Hydrol. Earth Syst. Sci.*, *21*, 635–650.

Kong, X. M., G. H. Huang, Y. R. Fan, and Y. P. Li (2015), Maximum entropy-Gumbel-Hougaard copula method for simulation of monthly streamflow in Xiangxi River, China, *Stochastic Environ. Res. Risk Assess.*, *29*, 833–846.

Lakshmi, V., et al. (2014), *Remote Sensing of the Terrestrial Water Cycle*, John Wiley, Hoboken, N. J.

Leisenring, M., and H. Moradkhani (2012), Analyzing the uncertainty of suspended sediment load prediction using sequential Monte Carlo methods, *J. Hydrol.*, *468–469*, 268–282, doi:10.1016/j.jhydrol.2012.08.049.

Lemke, L. D., L. M. Abriola, and J. R. Lang (2004), Influence of hydraulic property correlation on predicted dense nonaqueous phase liquid source zone architecture, mass recovery and contaminant flux, *Water Resour. Res.*, *40*, W12417, doi:10.1029/2004WR003061.

Li, T., M. Bolic, and P. M. Djuric (2015), Resampling methods for particle filtering: Classification, implementation, and strategies, *IEEE Signal Process. Mag.*, *32*(3), 70–86.

Li, Y., D. Ryu, A. W. Western, and Q. J. Wang (2013), Assimilation of stream discharge for flood forecasting: The benefits of accounting for routing time lags, *Water Resour. Res.*, *49*, 1887–1900, doi:10.1002/wrcr.20169.

Liu, J. S., R. Chen, and T. Logvinenko (2001), A theoretical framework for sequential importance sampling and resampling, in *Sequential Monte Carlo Methods in Practice*, edited by A. Doucet et al., Springer, New York.

Liu, Y., et al. (2012), Advancing data assimilation in operational hydrologic forecasting: Progresses, challenges, and emerging opportunities, *Hydrol. Earth Syst. Sci.*, *16*, 3863–3887.

Ma, M., L. Ren, V. P. Singh, F. Yuan, L. Chen, X. Yang, and Y. Liu (2016), Hydrologic model-based Palmer indices for drought characterization in the Yellow River basin, China, *Stochastic Environ. Res. Risk Assess.*, *30*(5), 1401–1420.

Malec, P., and M. Schienle (2014), Nonparametric kernel density estimation near the boundary, *Comput. Stat. Data Anal.*, *72*, 57–76.

Manache, G., and C. S. Melching (2008), Identification of reliable regression- and correlation-based sensitivity measures for importance ranking of water quality model parameters, *Environ. Modell. Software*, *23*, 549–562.

Mateo, C. M., N. Hanasaki, D. Komori, K. Tanaka, M. Kiguchi, A. Champathong, T. Sukhapunnaphan, D. Yamazaki, and T. Oki (2014), Assessing the impacts of reservoir operation to floodplain inundation by combining hydrological, reservoir management, and hydrodynamic models, *Water Resour. Res.*, *50*, 7245–7266, doi:10.1002/2013WR014845.

Montzka, C., H. Moradkhani, L. Weihermuller, H. J. H. Franssen, M. Canty, and H. Vereecken (2011), Hydraulic parameter estimation by remotely-sensed top soil moisture observations with the particle filter, *J. Hydrol.*, *399*(3–4), 410–421.

Moore, R. J. (1985), The probability-distributed principle and runoff production at point and basin scales, *Hydrol. Sci. J.*, *30*, 273–297.

Moore, R. J. (2007), The PDM rainfall-runoff model, *Hydrol. Earth Syst. Sci.*, *11*(1), 483–499.

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

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

Moradkhani, H., K. L. Hsu, H. Gupta, and S. Sorooshian (2005b), Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter, *Water Resour. Res.*, *41*, W05012, doi:10.1029/2004WR003604.

Murphy, A. H., and R. L. Winkler (1987), A general framework for forecast verification, *Mon. Weather Rev.*, *115*, 1330–1338.

Nelsen, R. B. (2006), *An Introduction to Copulas*, Springer, New York.

Noh, S. J., O. Rakovec, A. H. Weerts, and Y. Tachikawa (2014), On noise specification in data assimilation schemes for improved flood forecasting using distributed hydrological models, *J. Hydrol.*, *519,PartD*, 2707–2721.

Pan, F., J. Zhu, M. Ye, Y. A. Pachepsky, and Y. S. Wu (2011), Sensitivity analysis of unsaturated flow and contaminant transport with corrected parameters, *J. Hydrol.*, *397*(3–4), 238–249.

Plaza-Guingla, D. A., R. De Keyser, G. J. M. De Lannoy, L. Giustarini, P. Matgen, and V. R. N. Pauwels (2012), The importance of parameter resampling for soil moisture data assimilation into hydrologic models using the particle filter, *Hydrol. Earth Syst. Sci.*, *16*(2), 375–390.

Plaza-Guingla, D. A., R. De Keyser, G. J. M. De Lannoy, L. Giustarini, P. Matgen, and V. R. N. Pauwels (2013), Improving particle filters in rainfall-runoff models: Application of the resample-move step and the ensemble Gaussian particle filter, *Water Resour. Res.*, *49*, 4005–4021, doi:10.1002/wrcr.20291.

Pohlmann, K. F., A. E. Hassan, and J. B. Chapman (2002), Modeling density-driven flow and radionuclide transport at an underground nuclear test: Uncertainty analysis and effect of parameter correlation, *Water Resour. Res.*, *38*(5), 1059, doi:10.1029/2001WR001047.

Rasmussen, J., H. Madsen, K. H. Jensen, and J. C. Refsgaard (2015), Data assimilation in integrated hydrological modeling using ensemble Kalman filtering: Evaluating the effect of ensemble size and localization on filter performance, *Hydrol. Earth Syst. Sci.*, *19*, 2999–3013.

Remesan, R., T. Bellerby, and L. Frostick (2014), Hydrological modelling using data from monthly GCMs in a regional catchment, *Hydrol. Processes*, *28*(8), 3241–3263.

Rojas, R., L. Feyen, and A. Dassargues (2009), Sensitivity analysis of prior model probabilities and the value of prior knowledge in assessment of conceptual model uncertainty in groundwater modeling, *Hydrol. Processes*, *23*, 1131–1146.

Salamon, P., and L. Feyen (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 Y. Tang (2015), A modified ensemble Kalman particle filter for non-Gaussian systems with nonlinear measurement functions, *J. Adv. Model. Earth Syst.*, *7*, 50–66, doi:10.1002/2014MS000373.

Vrugt, J. A., and M. Sadegh (2013), Toward diagnostic model calibration and evaluation: Approximate Bayesian computation, *Water Resour. Res.*, *49*, 4335–4345, doi:10.1002/wrcr.20354.

Vrugt J. A., C. J. F. ter Braak, C. G. H. Diks, and G. Schoups (2013), Hydrologic data assimilation using particle Markov chain Monte Carlo simulation: Theory, concepts and applications, *Adv. Water Resour.*, *51*, 457–478.

Vrugt, J. A., C. G. H. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten (2005), Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, *Water Resour. Res.*, *41*, W01017, doi:10.1029/2004WR003059.

Vrugt, J. A., H. V. Gupta, W. Bouten, and S. Sorooshian (2003), A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters, *Water Resour. Res.*, *39*(8), 1201, doi: 10.1029/2002WR001642.

Wang, C., Q. Duan, C. H. Tong, Z. Di, and W. Gong (2016), A GUI platform for uncertainty quantification of complex dynamical models, *Environ. Modell. Software*, *76*, 1–12.

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

Xiong, L. H., C. Jiang, C. Y. Xu, and K. X. Yu (2015), A framework of change-point detection for multivariate hydrological series, *Water Resour. Res.*, *51*, 8198–8217, doi:10.1002/2015WR017677.

Xu, C., and G. Z. Gertner (2008), Uncertainty and sensitivity analysis for models with correlated parameters, *Reliab. Eng. Syst. Safety*, *93*(10), 1563–1573.

Xu, Y., G. H. Huang, and Y. R. Fan (2017), Multivariate flood risk analysis for Wei River, *Stochastic Environ. Res. Risk Assess.*, *31*(1), 225–242.

Yan, H., and H. Moradkhani (2016), Combined assimilation of streamflow and satellite soil moisture with the particle filter and geostatistical modeling, *Adv. Water Resour.*, *94*, 364–378.

Yan, H., C. M. DeChant, and H. Moradkhani (2015), Improving soil moisture profile prediction with the particle filter-Markov chain Monte Carlo method, *IEEE Trans. Geosci. Remote Sens.*, *53*, 6134–6147.

Zhang, Y., D. S. Oliver, Y. Chen, and H. J. Skaug (2014), Data assimilation by use of the iterative ensemble smoother for 2D facies models, *SPE J.*, *20*(1), 169–185.