https://doi.org/10.1016/j.jhydrol.2017.09.052

Title Page:

Examining dynamic interactions among experimental factors influencing hydrologic data assimilation with the ensemble Kalman filter

Author names and affiliations: S. Wang[1,*], G.H. Huang[2], B.W. Baetz[3], X.M. Cai[4], B.C. Ancell[5], Y.R. Fan[2]

[1]Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, China

[2]Institute for Energy, Environment and Sustainable Communities, University of Regina, Regina, Saskatchewan, Canada

[3]Department of Civil Engineering, McMaster University, Hamilton, Ontario, Canada

[4]Ven Te Chow Hydrosystems Laboratory, Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

[5]Department of Geosciences, Texas Tech University, Lubbock, Texas, USA

[*]Corresponding author. Phone: (852) 5937-3808; Email: shuo.s.wang@polyu.edu.hk, sshuo.wwang@gmail.com

© 2017. This manuscript version is made available under the CC-BY-NC-ND 4.0 license https://creativecommons.org/licenses/by-nc-nd/4.0/

SUMMARY

The ensemble Kalman filter (EnKF) is recognized as a powerful data assimilation technique that generates an ensemble of model variables through stochastic perturbations of forcing data and observations. However, relatively little guidance exists with regard to the proper specification of the magnitude of the perturbation and the ensemble size, posing a significant challenge in optimally implementing the EnKF. This paper presents a robust data assimilation system (RDAS), in which a multi-factorial design of the EnKF experiments is first proposed for hydrologic ensemble predictions. A multi-way analysis of variance is then used to examine potential interactions among factors affecting the EnKF experiments, achieving optimality of the RDAS with maximized performance of hydrologic predictions. The RDAS is applied to the Xiangxi River watershed which is the most representative watershed in China's Three Gorges Reservoir region to demonstrate its validity and applicability. Results reveal that the pairwise interaction between perturbed precipitation and streamflow observations has the most significant impact on the performance of the EnKF system, and their interactions vary dynamically across different settings of the ensemble size and the evapotranspiration perturbation. In addition, the interactions among experimental factors vary greatly in magnitude and direction depending on different statistical metrics for model evaluation including the Nash–Sutcliffe efficiency and the Box–Cox transformed root-mean-square error. It is thus necessary to test various evaluation metrics in order to enhance the robustness of hydrologic prediction systems.

Keywords: Ensemble Kalman filter; Data assimilation; Hydrologic ensemble prediction; Interaction; Streamflow; Uncertainty

1. Introduction

Since hydrologic models are mathematical representations of complex watershed processes, uncertainty is pervasive throughout hydrologic predictions (DeChant and Moradkhani, 2014). Uncertainty in hydrologic predictions originates from various sources, including the descriptions of boundary and initial conditions, the errors in model forcing data, difficulty in obtaining accurate parameter estimates, and model structural deficiencies (Ajami et al., 2007). Therefore, efficient quantification and reduction of uncertainty are necessary to provide reliable hydrologic predictions (Moradkhani et al., 2012; Wang et al., 2015c).

Over the past few decades, tremendous efforts have been made in the development and application of sequential data assimilation techniques for explicitly dealing with various sources of uncertainty in hydrologic modeling (Weerts and El Serafy, 2006; Liu and Gupta, 2007; Ryu et al., 2009; Gharamti et al., 2013; Panzeri et al., 2014; Randrianasolo et al., 2014; Khan and Valeo, 2016; Wang et al., 2017). Sequential data assimilation techniques continuously update model states when new observations become available to improve the forecast accuracy (Vrugt et al., 2005). The Kalman filter (KF) is the most commonly used sequential data assimilation technique, which was developed in the 1960s for optimal control of linear dynamic systems (Kalman, 1960). For nonlinear dynamics, the extended Kalman filter (EKF) can be used, which linearizes the error covariance equation by using a tangent linear operator. However, EKF produces unstable results when the nonlinearity in dynamic systems is strong and requires considerable computational effort due to the error covariance propagation (Evensen, 1992). As a result, the ensemble Kalman filter (EnKF) was introduced by Evensen (1994). The EnKF takes advantage of the Monte Carlo method to approximate the error covariance evolution equation used in the EKF, which is capable of providing the forecast error estimate with a significantly lower computational cost and without any closure problem in the error covariance evolution equation.

Due to its attractive features of real-time adjustment and efficient implementation, the EnKF has been extensively used for recursive estimation of hydrologic model parameters and state variables (Xie and Zhang, 2010; Cammalleri and Ciraolo, 2012; DeChant and Moradkhani, 2012; Rafieeinasab et al., 2014; Gharamti et al., 2015; Liu et al., 2016; Pathiraja et al., 2016b). For example, Moradkhani et al. (2005) proposed a dual state-parameter estimation approach based on the EnKF for sequential estimation of both parameters and state variables of a hydrologic model. Wang et al. (2009) proposed a constrained EnKF framework for simultaneous state estimation and sequential parameter learning in hydrologic modeling, in which the naive method, the projection and accept/reject methods were used to deal with inequality constraints. Samuel et al. (2014) evaluated the variations of streamflow and soil moisture by using the EnKF with dual state-parameter estimation for streamflow assimilation, soil moisture assimilation, and combined assimilation of streamflow and soil moisture. Pathiraja et al. (2016a) investigated the potential for data assimilation by using the EnKF to detect known temporal patterns in model parameters from streamflow observations. The EnKF is recognized as a powerful data assimilation technique that generates an ensemble of model variables through stochastic perturbations of forcing data and observations (inputs and outputs). Thus, identification of perturbation factors and selection of the ensemble size are key features of the EnKF (Moradkhani et al., 2005). However, relatively little guidance exists in literature with regard to the proper specification of the magnitude of the perturbation and the ensemble size, posing a significant challenge in optimally implementing the EnKF (Crow and Loon, 2006).

Since inappropriate specification of factors affecting the EnKF experiment can degrade the performance of the data assimilation system, sensitivity experiments are often carried out for identifying error parameters and estimating the ensemble size (Clark et al., 2008; Sun et al., 2009; McMillan et al., 2013; Rasmussen et al., 2015). However, the sensitivity analysis experiments are limited in determining optimal settings of the EnKF applied to a particular problem. Thus, Yin et al. (2015) used a series of mathematical derivations to derive the optimal ensemble size of the EnKF used for a sequential soil moisture data assimilation system. In the EnKF, stochastic perturbations account for uncertainties in model parameters, inputs, and outputs. Specification of perturbation factors is a key feature of the EnKF, which plays a crucial role in the performance of sequential data assimilation experiments (Clark et al., 2008).

As a recursive scheme for estimating state variables and model parameters, the experimental factors involved in the EnKF are actually correlated with each other, and their interactions have a remarkable influence on the behavior of nonlinear dynamic systems. For example, many of the highly sensitive factors may provide redundant and misleading information regarding the variability of response variables since their sensitivities may be correlated with those of the other factors. As a result, failure to account for potential interactions among experimental factors can degrade the performance of the EnKF system (Crow and Loon, 2006; Thiboult and Anctil, 2015). It is thus necessary to examine the interactions among experimental factors and to quantify their contributions to the variation in model responses in order to maximize the predictive performance.

In this paper, we develop a robust data assimilation system (RDAS) with a factorial experimental design framework to enhance the effectiveness and robustness of the EnKF for hydrologic ensemble predictions. A multi-factorial EnKF method will be proposed by combining the strengths of multivariate hypothesis testing and sequential data assimilation techniques. In the RDAS, the EnKF will be carried out under various combinations of factors with different scenarios, leading to a diverse set of EnKF experiments. The multi-way analysis of variance (ANOVA) will then be used to uncover dynamic interactions among factors involved in the EnKF experiments, which provides meaningful insights for advancing the understanding of the sequential data assimilation process and maximizing the EnKF performance. The RDAS will be applied to predict daily streamflow in the Xiangxi River watershed in China since daily streamflow predictions play a key role in flood risk assessment and management.

This paper is organized as follows. Section 2 introduces the framework of the proposed RDAS for hydrologic ensemble predictions. Section 3 provides details on the study area and the experimental setup. Section 4 presents a systematic analysis of multi-factorial EnKF experiments along with a thorough discussion of interactions among experimental factors affecting the performance of the EnKF system. Finally, conclusions are drawn in Section 5.

2. Development of robust data assimilation system

The RDAS takes into account potential interactions among experimental factors influencing the EnKF data assimilation and quantifies their joint effects on the EnKF performance. An overview of the steps involved within the RDAS framework is provided as follows: 1) selection of the EnKF experimental factors, 2) factorial design of the EnKF experiments, 3) execution of ensemble data assimilation experiments, 4) multi-way ANOVA, 5) examination of dynamic interactions among experimental factors, 6) quantification of the joint effects of experimental factors on the EnKF performance, and 7) hydrologic ensemble predictions. The aforementioned steps can be categorized into four parts: EnKF, multi-factorial ANOVA, multi-factorial EnKF, and selection of statistical metrics for model evaluation.

2.1. Ensemble Kalman filter

The EnKF is a sequential data assimilation technique that makes use of Monte Carlo integration methods to approximate the error covariance matrix by a stochastic ensemble of model realizations (Evensen, 2003). In contrast to the extended Kalman filter (EKF), the EnKF represents the error covariance evolution through a set of model realizations rather than an explicit mathematical expression, which is particularly useful for nonlinear dynamic models. As a result, the ensemble of model states is integrated forward in time to predict error statistics (DeChant and Moradkhani, 2012). The model forecast can be made through the EnKF as follows:

$$x_{i,t+1}^{-} = f\left(x_{i,t}^{+}, u_{i,t+1}, \theta_{i,t+1}^{-}\right) + \omega_{i,t+1}, \qquad \omega_{i,t+1} \sim N\left(0, \Sigma_{t+1}^{m}\right), \tag{1}$$

where i and t denote the ensemble number and the time step, respectively; $x_{i,t}^{+}$ and $x_{i,t+1}^{-}$ represent the posterior model states at the previous time step and the predicted model states at the current time step, respectively; f represents the forward model that propagates the system states from time t to t+1 in response to model inputs $u_{i,t+1}$ and parameters $\theta_{i,t+1}$; and $\omega_{i,t+1}$ represent the model errors that follow a Gaussian distribution with zero mean and covariance $\sum_{t+1}^{m}$. As for the recursive parameter estimation through the EnKF, it is assumed that model parameters are perturbed by a small random noise in order to maintain diversity in posterior parameters:

$$\theta_{i,t+1}^{-} = \theta_{i,t}^{+} + \tau S(\theta_{i,t}^{-}), \qquad (2)$$

where τ is a small tuning parameter which was set to 0.01 in this study, and $S(\theta_{i,t}^{-})$ is the standard deviation of the prior parameter distribution at the previous time step (DeChant and Moradkhani, 2012).

Prior to the update of model states and parameters, predictions can be made by:

$$y_{i,t+1}^{-} = h\left(x_{i,t+1}^{-}, \theta_{i,t+1}^{-}\right), \qquad (3)$$

where $y_{i,t+1}^{-}$ is the prediction, and h is the operator that relates state variables and parameters to measured variables (streamflow) and yields the expected value of the prediction given model states and parameters. After predictions are obtained, the posterior states and parameters are estimated with the Kalman update equations as follows:

$$x_{i,t+1}^{+} = x_{i,t+1}^{-} + K_{xy}\left[y_{t+1} + \varepsilon_{i,t+1} - y_{i,t+1}^{-}\right], \qquad \varepsilon_{i,t+1} \sim N\left(0, \Sigma_{t+1}^{y}\right), \qquad (4)$$

$$\theta_{i,t+1}^{+} = \theta_{i,t+1}^{-} + K_{\theta y}\left[y_{t+1} + \varepsilon_{i,t+1} - y_{i,t+1}^{-}\right], \qquad \varepsilon_{i,t+1} \sim N\left(0, \Sigma_{t+1}^{y}\right), \qquad (5)$$

where y_{t+1} is the observed value, $\varepsilon_{i,t+1}$ is the observation error which is assumed to be Gaussian and independent of mode error ω_{t+1} (Moradkhani et al., 2005), and K_{xy} and $K_{\theta y}$ represent the Kalman gains for states and parameters, respectively:

$$K_{t+1}^{x} = \Sigma_{t+1}^{xy}\left[\Sigma_{t+1}^{yy} + \Sigma_{t+1}^{y}\right]^{-1}, \tag{6}$$

$$K_{t+1}^{\theta} = \Sigma_{t+1}^{\theta y}\left[\Sigma_{t+1}^{yy} + \Sigma_{t+1}^{y}\right]^{-1}. \tag{7}$$

where Σ_{t+1}^{xy} is the cross covariance of ensembles of state variables $x_{i,t+1}^{-}$ with predicted observations $y_{i,t+1}^{-}$, $\Sigma_{t+1}^{\theta y}$ is the cross covariance of ensembles of model parameters $\theta_{i,t+1}^{-}$ with predicted observations $y_{i,t+1}^{-}$, Σ_{t+1}^{yy} is the variance of predicted observations, and Σ_{t+1}^{y} is the observation error variance (Dechant and Moradkhani, 2011).

Although critical issues for the EnKF data assimilation have been introduced in the literature, little effort has been made to explicitly examine the potential interactions among experimental factors affecting the EnKF data assimilation, including the ensemble size (i) and random perturbations to model parameters (θ), inputs (u), and outputs (y). Multi-factorial ANOVA is thus proposed to reveal the latent interactions among experimental factors and their joint effects on the EnKF performance.

2.2. Multi-factorial analysis of variance

Multi-factorial ANOVA is a powerful tool for examining the effects of multiple factor variables and their interactions on response variables by conducting hypothesis tests with the F-statistic. The null hypothesis assumes that the group means for all response variables are equal while the alternative hypothesis states that at least one mean is different (Montgomery and Runger, 2013).

In a factorial experiment, all possible combinations of the levels of factors are investigated. For example, if there are a levels of factor A, b levels of factor B, and c levels of factor C, there will be a total of $abcn$ observations in a complete factorial experiment with n replicates. ANOVA is derived from the partitioning of total variability into various components due to different sources of variation. The ANOVA model for such a factorial experiment can thus be expressed as:

$$y_{ijkl} = \mu + \tau_i + \beta_j + \gamma_k + (\tau\beta)_{ij} + (\tau\gamma)_{ik} + (\beta\gamma)_{jk} + (\tau\beta\gamma)_{ijk} + \varepsilon_{ijkl} \quad \begin{cases} i = 1, 2, \ldots a \\ j = 1, 2, \ldots b \\ k = 1, 2, \ldots c \\ l = 1, 2, \ldots n \end{cases}. \tag{8}$$

where μ is the overall mean effect, τ_i is the effect of the ith level of factor A, β_i is the effect of the jth level of factor B, γ_k is the effect of the kth level of factor C, $(\tau\beta)_{ij}$ is the effect of the interaction between factors A and B, $(\tau\gamma)_{ik}$ is the effect of the interaction between factors A and C, $(\beta\gamma)_{jk}$ is the effect of the interaction between factors B and C, $(\tau\beta\gamma)_{ijk}$ is the effect of the interaction between factors A, B, and C, and ε_{ijkl} is the random error component.

The ANOVA model contains three main effects, three two-factor interactions, a three-factor interaction, and an error term. The effects are defined as deviations from the overall mean (Montgomery, 2000). The F-statistic can then be used to test the statistical significance for each of the factors as well as their interactions (Wang et al., 2015b). The multi-factorial ANOVA approach is useful in testing differences between two or more means by analyzing variances from multiple sources (Wu and Hamada, 2009; Shen et al., 2012; Liu et al., 2016; Wang et al., 2015a, 2016a; Zeng et al., 2016), and thus it can be used to reveal the potential interactions among experimental factors involved in the EnKF experiments.

2.3. Multi-factorial ensemble Kalman filter

To enhance the effectiveness and robustness of sequential data assimilation techniques, a multi-factorial EnKF method is proposed by merging the strengths of factorial ANOVA and the EnKF. The multi-factorial EnKF method is capable not only of uncovering interactions among perturbation factors influencing the EnKF, but also of quantifying and reducing uncertainties in hydrologic ensemble predictions.

To tackle various sources of uncertainty including input and output measurement errors as well as parameter uncertainty, stochastic perturbations were used in the EnKF by adding noise to the forcing data (precipitation and evapotranspiration) and streamflow observations in this study. Four experimental factors were thus taken into account in the data assimilation experiment, including the ensemble size (ES), precipitation noise (PN), evapotranspiration noise (EN), and observation noise (ON) added to the EnKF. Each factor had three levels (scenarios) of interest based on . In fact, the EnKF experimental factors are correlated with each other in the data assimilation experiment, and the magnitude and direction of pairwise interactions between experimental factors vary along with different settings of other factors, leading to nonlinear dynamics of interacting factors. It is thus necessary to explore dynamic interactions among perturbation factors in order to maximize the performance of the EnKF system.

As shown in Fig. 1, a 3^4 factorial experimental design can be constructed, in which the EnKF data assimilation experiment is conducted under each combination of experimental factors, leading to hydrologic ensemble predictions. In the 3^4 system of factorial designs, there are 3^4 factorial combinations with $3^4 - 1$ degrees of freedom between them. If there are n replicated experiments, there will be $n3^4 - 1$ total degrees of freedom and $3^4(n - 1)$ degrees of freedom for error. These factorial combinations allow sums of squares to be computed for four main effects, six two-factor interactions, three

three-factor interactions, and one four-factor interaction (Montgomery, 2000; Wang et al., 2016b).

After the sums of squares for effects are computed through the ANOVA model, the *F* test can be performed to reveal the statistical significance for each of the four factors and their interactions. As a result, the pairwise interactions between the EnKF factors can be calculated by:

$$F_{ES\times PN}=\frac{SS_{ES\times PN}/4}{SS_E/3^4(n-1)},\quad F_{ES\times EN}=\frac{SS_{ES\times EN}/4}{SS_E/3^4(n-1)},\quad F_{ES\times ON}=\frac{SS_{ES\times ON}/4}{SS_E/3^4(n-1)},\tag{9}$$

$$F_{PN\times EN}=\frac{SS_{PN\times EN}/4}{SS_E/3^4(n-1)},\quad F_{PN\times ON}=\frac{SS_{PN\times ON}/4}{SS_E/3^4(n-1)},\quad F_{EN\times ON}=\frac{SS_{EN\times ON}/4}{SS_E/3^4(n-1)}.\tag{10}$$

where $SS_{ES\times PN}$, $SS_{ES\times EN}$, $SS_{ES\times ON}$, $SS_{PN\times EN}$, $SS_{PN\times ON}$, $SS_{EN\times ON}$, and SS_E represent the sum of squares for the pairwise interactions between the ensemble size and the precipitation noise, those between the ensemble size and the evapotranspiration noise, those between the ensemble size and the observation noise, those between the precipitation noise and the evapotranspiration noise, those between the precipitation noise and the observation noise, those between the evapotranspiration noise and the observation noise, and the error component, respectively. The factorial data assimilation experiment with the EnKF is useful for maximizing the performance of hydrologic ensemble predictions through robustly examining dynamic interactions among experimental factors.

Place Fig. 1 here

2.4. Statistical metrics for model evaluation

Two statistical metrics, the Nash–Sutcliffe efficiency (NSE) and the Box–Cox transformed root-mean-square error (TRMSE), are applied to evaluate prediction errors. The NSE is commonly used to emphasize the predictive capacity of high flows due to the use of squared residuals, and can be defined as:

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{T}(Q_{s,t} - Q_{o,t})^2}{\sum_{t=1}^{T}(Q_{o,t} - \bar{Q}_o)^2}. \tag{11}$$

where T is the number of time steps, $Q_{o,t}$ is the observed discharge at time t, $Q_{s,t}$ is the predicted discharge at time t, and $\bar{Q}_o$ is the mean of observed discharges. NSE ranges from $-\infty$ to 1 (Nash and Sutcliffe, 1970). The TRMSE emphasizes low-flow prediction errors, and can be defined as:

$$\mathrm{TRMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{Q}_{s,t} - \hat{Q}_{o,t})^2}, \tag{12}$$

$$\hat{Q} = \frac{(1+Q)^{\lambda} - 1}{\lambda}. \tag{13}$$

where $\hat{Q}_{s,t}$ is the transformed predicted discharge at time t, $\hat{Q}_{o,t}$ is the transformed observed discharge at time t, and $\hat{Q}$ represents the Box-Cox transformation of discharge Q (Box and Cox, 1964), where $\lambda = 0.3$ as recommend by Misirli et al. (2013).

In this paper, both NSE and TRMSE metrics were used to examine the performance of hydrologic ensemble predictions through the EnKF data assimilation experiments. In addition, parameter identification and interaction detection were performed based on different evaluation metrics, enhancing the robustness of hydrologic data assimilation.

3. Experimental setup

3.1. Site and model descriptions

The robust data assimilation system (RDAS) is applied to predict daily streamflow in the Xiangxi River watershed. As shown in Fig. 2, the Xiangxi River is the largest tributary of the Three Gorges Reservoir (TGR) in Central China's Hubei Province. The Xiangxi River watershed with a total area of 3,099 km^2 lies in the subtropical region, and experiences a typical continental monsoon climate with substantial temperature variations in Spring and concentrated rainfalls in Summer. The weather is rainy in Autumn and snowy in Winter. It is the most representative watershed in the TGR region in terms of topographic properties, runoff volumes, and economic conditions (Han et al., 2014). In this study, a total of three years of meteorological and hydrological data from January 1994 to December 1996 were used for assimilating daily streamflow in the Xiangxi River watershed. The three years of data were selected because there were continuous observational data and relatively little human interference in the natural river flow during the period of time, ensuring the data quality and maximizing the reliability of hydrologic predictions.

Place Fig. 2 here

Data assimilation experiments with the EnKF were undertaken by using HyMOD which is a well-known rainfall-runoff model with a daily time step (Moore, 1985). The runoff production in HyMOD is characterized as a rainfall excess process, and the runoff is determined according to a probability-distributed storage capacity model (Moore, 2007; Bulygina and Gupta, 2011; Young, 2013). The catchment is considered as a finite number of points, and each of them has a certain soil moisture capacity denoted as c [L]. Due to spatial variability such as soil type and depth within the catchment, the variability of soil moisture capacities can be characterized by a cumulative distribution function (CDF) defined as:

$$F(c) = 1 - \left(1 - \frac{c}{C_{\max}}\right)^{b_{\exp}} \qquad 0 \le c \le C_{\max}\,. \tag{14}$$

where $C_{\max}$ [L] is the maximum soil moisture capacity, and $b_{\exp}$ [-] is the degree of spatial variability in soil moisture capacities and affects the shape of the CDF. The CDF indicates the probability of occurrence of a specific soil moisture capacity across the catchment. The HyMOD model partitions excess rainfall into surface and subsurface soil moisture storage, denoted as S [L], through a partitioning factor β [-]. The surface storage is characterized by three quick-flow tanks (S_1, S_2, and S_3), and the subsurface storage is represented by a single slow-flow tank (S_s). The residence time of slow- and quick-flow tanks are represented as R_s and R_q [T], respectively. The generated streamflow is the addition of discharges from slow- and quick-flow tanks. The input data of daily precipitation P [mm/d] and potential evapotranspiration ET [mm/d] are used to drive the HyMOD model.

HyMOD is characterized by five state variables (S, S_s, S_1, S_2, and S_3) and five parameters ($C_{\max}$, $b_{\exp}$, β, R_s, and R_q). The initial ranges of model parameters are given in

Table 1. To properly assess the performance of the RDAS, a predefined set of "true" model parameters and the observed forcing data including daily precipitation and potential evapotranspiration were used to generate synthetic streamflow observations. The EnKF was then performed to assimilate the synthetic streamflow observations, and the convergence of model parameters to the "true" values was evaluated accordingly. Since uncertainty inevitably exists in the forcing data and streamflow observations in practice, stochastic perturbations are employed by adding noise to the forcing data and observations in order to account for various sources of uncertainty, leading to an ensemble of model variables. As a result, specification of perturbation factors and the ensemble size is a key feature of the EnKF, which plays an important role in the performance of hydrologic ensemble predictions.

Place Table 1 here

3.2. Factorial design of ensemble data assimilation experiments

To develop the RDAS for hydrologic ensemble predictions in the Xiangxi River watershed, a 3^4 full factorial design that involved four factors with each having three levels was first constructed. As shown in Fig. 3, experimental factors represent the ensemble size and random perturbations added to precipitation, potential evapotranspiration, and streamflow observation in the EnKF data assimilation experiment. The precipitation data were log-normally perturbed with three relative errors of 10%, 20%, and 30%, while potential evapotranspiration and streamflow observation were normally perturbed with the error scenarios of 10%, 20%, and 30%, respectively. In addition, three ensemble sizes of

30, 50, and 80 were taken into account in the factorial experimental design. As a result, the 3^4 factorial design contained 81 combinations of the four factors under three scenarios.

The EnKF data assimilation experiment was conducted under each factorial combination, leading to hydrologic ensemble predictions. Such a factorial design of ensemble data assimilation experiments is able to reveal meaningful implications for maximizing the EnKF performance through examining potential interactions among experimental factors. In addition, the factorial experimental design is useful for enhancing the effectiveness and robustness of sequential data assimilation methods as well as for quantifying and reducing uncertainty in hydrologic predictions.

Place Fig. 3 here
----

4. Results and discussion

4.1. Examination of interactions among experimental factors affecting hydrologic data assimilation

According to the 3^4 factorial design, a total of 81 NSE and TRMSE values were obtained through the EnKF data assimilation experiments under different combinations of factor settings (as shown in Fig. 4). The multi-factorial ANOVA was then performed to examine the interactions among factors affecting the performance of hydrologic predictions. To check the Gaussian assumption of the residual distribution for the multi-factorial ANOVA, Fig. 5 presents the normal probability plots of residuals for the NSE and TRMSE metrics, respectively. As the resulting plot is approximately linear, this means that the Gaussian assumption of the residual distribution is valid. The Shapiro–Wilk test was also performed as a rigorous statistical checking for normality of residuals. The results with *P*-values being greater than 0.05 verify that residuals are normally distributed.

Place Figs. 4 and 5 here

Fig. 6 depicts the response desirability in terms of maximum NSE and minimum TRMSE produced in different regions of the plane defined by pairs of experimental factors, where each region of the plane represents a different combination of the levels of two factors. The response desirability indicates which levels of experimental factors produce the most desirable predicted response on NSE and TRMSE, and its values range from 0.0 for an undesirable response to 1.0 for a highly desirable response. As shown in Fig. 6(a), the pairwise interactions between the ensemble size and other factors involved in EnKF data assimilation experiments tend to produce relatively high desirability values of the response on NSE when the ensemble size is set to 30 and 80. Likewise, the pairwise interaction between the precipitation noise and other factors would generate a relatively high response desirability when the settings of the precipitation noise are 10% and 30%. In addition, when the setting of the evapotranspiration noise is 20%, its interactions with other factors would lead to a high response desirability.

Different from the desirability values of the response on NSE, when the setting of the ensemble size is 50, its interactions with other factors tend to produce relatively large desirability values of the response on TRMSE, as shown in Fig. 6(b). Our findings reveal dynamic interactions between experimental factors affecting the EnKF data assimilation and their contributions to the variation of overall desirability values of responses in terms of different statistical metrics for model evaluation. These findings are useful for identifying the optimal settings of experimental factors so as to produce related responses with the highest overall desirability, thus maximizing the performance of hydrologic predictions.

Place Fig. 6 here

Fig. 7 shows the *F*- and *t*-values derived from the factorial ANOVA for all pairwise interactions between the EnKF experimental factors. Results reveal that the interaction between the precipitation noise and the streamflow observation noise has the largest impact on the predictive performance in terms of NSE and TRMSE. To further examine the nonlinear relationships between the settings of correlated factors and the resulting predictive performance, Fig. 8 presents the fitted surfaces of NSE and TRMSE for the most significant pairwise interactions between the precipitation noise and the streamflow observation noise. The fitted surface of NSE reveals that, when the precipitation noise is at its low (10%) and medium (20%) levels, there is an increasing trend in the NSE value when changing the observation noise from its low (10%) and high (30%) levels to its medium level (20%). When the precipitation noise is at its high level (30%), the NSE value increases across the high, medium, and low levels of the observation noise, and thus the maximum value of NSE would be obtained when the precipitation noise is at its high level and the observation noise is at its low level.

As for the fitted surface of TRMSE, when the precipitation noise is at its low level, the TRMSE value decreases across the low, medium, and high levels of the observation noise, and thus the minimum value of TRMSE would be obtained when the precipitation noise is at its low level and the observation noise is at its high level. Results reveal that the pairwise interaction between the precipitation noise and the streamflow observation noise added to the EnKF system varies greatly in magnitude and direction depending on different statistical metrics for model evaluation. It is thus necessary to investigate interactions among experimental factors influencing the EnKF data assimilation based on various evaluation metrics so as to maximize the predictive performance.

Place Figs. 7 and 8 here

To further explore multi-factor interactions in hydrologic data assimilation with the EnKF, Fig. 9 shows the variations of marginal means of NSE and TRMSE with 95% confidence intervals for all three-way interactions. Our findings reveal complex interactions among multiple experimental factors affecting the EnKF data assimilation and their contributions to the accuracy of hydrologic predictions. For example, Fig. 9(c) shows a considerable difference in the variation of NSE associated with the three levels (scenarios) of the streamflow observation noise over the levels of the precipitation noise, collapsed across the levels of the evapotranspiration noise, implying that the interaction between the settings of the precipitation noise and the streamflow observation noise varies significantly depending on the settings of the evapotranspiration noise. Such a multi-way interaction analysis reveals that the maximum value of NSE would be obtained when the settings of precipitation noise, evapotranspiration noise, and streamflow observation noise are 30%, 20%, and 10%, respectively.

As shown in Fig. 9(f), when the setting of the evapotranspiration noise is 20%, increasing the streamflow observation noise leads to a decreasing value of TRMSE at the low level (10%) of the precipitation noise but an increasing value of TRMSE at the high level (30%) of the precipitation noise. The minimum value of TRMSE would be obtained when the settings for precipitation noise, evapotranspiration noise, and streamflow observation noise are 10%, 20%, and 30%, respectively. Our findings indicate that the magnitude and direction of interactions among experimental factors vary dynamically in the EnKF data assimilation process. It is thus necessary to explore dynamic interactions among perturbed errors added to the EnKF and reveal their contributions to the performance of hydrologic predictions, advancing our understanding of the sequential data assimilation process.

Place Fig. 9 here

4.2. Ensemble streamflow predictions through robust data assimilation system

Based on the factorial design and analysis of the EnKF data assimilation experiments, the optimal combinations of factor settings can be identified for different evaluation metrics.

To maximize the performance of the EnKF system in terms of NSE, the optimal settings of the ensemble size and the perturbed errors added to precipitation, evapotranspiration, and streamflow observation data would be 80, 30%, 20%, and 10%, respectively. To minimize the TRMSE value, the optimal settings of the corresponding factors would be 50, 10%, 20%, and 30%, respectively. The EnKF data assimilation system with maximized performance can then be employed for recursive parameter estimation and streamflow predictions by using the HyMOD model.

Fig. 10 depicts the time evolution of model parameters with 90% confidence intervals derived from streamflow assimilation based on the NSE metric over a period of three years from January 1994 to December 1996. It is indicated that all parameters are seen to converge toward the "true" values defined as: C_{max} = 657 mm, b_{exp} = 5.54, β = 0.72, R_{s} = 0.15 d, and R_{q} = 0.70 d. The maximum storage capacity of the watershed denoted by C_{max} is the most identifiable parameter because C_{max} shows the fastest convergence with the smallest uncertainty bound. In contrast, the slow-flow tank parameter R_{s} is less identifiable than the others as R_{s} shows the slowest convergence. This is because the maximum storage capacity of the watershed is strongly correlated with the streamflow observation, whereas the slow-flow tank has the minimum contribution to the volume of generated streamflow. When TRMSE is used as the evaluation metric for streamflow assimilation, there will be a different recursive pattern of parameter evolution.

As shown in Fig. 11, all parameters converge toward the "true" values given as: C_{max} = 123 mm, b_{exp} = 5.47, β = 0.52, R_{s} = 0.04 d, and R_{q} = 0.47 d. Results reveal that parameter estimates vary significantly depending on different statistical metrics for model evaluation. For example, although C_{max} is still the most identifiable parameter with the smallest uncertainty bound, it converges to a much lower “true” value of 123 mm as compared with the value of 657 mm derived for the NSE metric. This is because the NSE metric emphasizes the predictive capacity of high flows due to the use of squared residuals while the TRMSE metric emphasizes low-flow prediction errors, leading to different parameter sets that achieve optimum predictive performance. It is thus necessary to perform an uncertainty assessment of model parameters and predictions based on different statistical metrics for model evaluation, enhancing our understanding of catchment behavior.

In practical applications, the optimal settings of the EnKF data assimilation system and the derived parameter set should be applied depending on the prediction of different hydrologic events. For example, the settings of the EnKF system and the parameter set derived based on the NSE metric can be used to predict flood events since the NSE emphasizes the predictive accuracy of high flows. In addition, the multi-objective optimization techniques can be used to achieve the best compromise between different evaluation metrics, which will provide meaningful guidance on the settings of the EnKF system and the parameter set used in practice.

Place Figs. 10 and 11 here

Fig. 12 presents a comparison between assimilated daily streamflows and observations over a period of three years from January 1994 to December 1996 in the Xiangxi River watershed. Results of streamflow predictions with 90% confidence intervals are derived from the ensemble of model outputs at each time step. Generally, there is good agreement between simulated and observed streamflow time series based on both NSE and TRMSE metrics, indicating that the RDAS is able to reasonably capture the rainfall-runoff relationship in the Xiangxi River watershed. Nevertheless, some of the high-flow events do not lie within the uncertainty bounds of streamflow predictions since the attribution of uncertainties to parameter estimates only is insufficient. Future studies will be undertaken to take into account other sources of uncertainty, such as those attributable to model structural errors in the RDAS.

Place Fig. 12 here

----

5. Conclusions

In this study, we developed a robust data assimilation system (RDAS) for hydrologic ensemble predictions based on a multi-factorial experimental design. The factorial design and analysis of hydrologic data assimilation experiments can examine interactions among the EnKF factors including the ensemble size and random perturbations added to precipitation, potential evapotranspiration, and streamflow observation data. Thus, the RDAS is useful to enhance the robustness of the EnKF data assimilation for quantification and reduction of uncertainty in hydrologic predictions. The RDAS was applied to predict daily streamflow in the Xiangxi River watershed located in China's Hubei Province. The NSE and TRMSE metrics were used to represent different hydrologic characteristics for sensitivity analysis, parameter estimation, and streamflow prediction.

Our results uncover that the pairwise interaction between perturbed precipitation and streamflow observations has the most significant impact on the performance of the EnKF system applied to the Xiangxi River watershed based on both NSE and TRMSE metrics, and their interactions vary greatly in magnitude and direction across different settings of the ensemble size and the evapotranspiration perturbation. Investigating single factors (e.g., the optimal ensemble size) with limited sensitivity experiments is thus inappropriate; instead, a systematic and comprehensive analysis of multi-way interactions among experimental factors is crucial for reaching the maximum efficiency of the EnKF data assimilation system. Our findings are useful for advancing the understanding of the sequential data assimilation process and for maximizing the performance of hydrologic predictions through identifying the optimal combinations of factor settings.

The EnKF data assimilation system with maximized performance was then employed for recursive parameter estimation and streamflow predictions by using the HyMOD model. Our findings reveal that identification of model parameters is conditional on different statistical metrics for model evaluation since there is a different recursive pattern of parameter evolution for the NSE and TRMSE metrics. The maximum storage capacity of the watershed is the most identifiable parameter with the fastest convergence and the smallest uncertainty bound in this case, but there are varying tracks of recursively estimating the corresponding time-evolving posterior distribution based on the NSE and TRMSE metrics. It is thus necessary to test various evaluation metrics in order to enhance the robustness of hydrologic prediction systems, especially when performing sensitivity analysis and parameter identification. By comparing assimilated daily streamflows against observations over a period of three years, the RDAS is shown to be capable of capturing well the rainfall-runoff relationship of the Xiangxi River watershed.

The RDAS is useful for advancing our understanding of the nonlinear dynamics of interacting factors influencing the EnKF data assimilation experiment, which has a strong potential to strengthen our capability in providing hydrometeorological forecasting. The RDAS is not restricted to hydrologic ensemble predictions; instead, the proposed framework of the RDAS is applicable to various ensemble prediction experiments with sequential data assimilation.

Acknowledgments

This research was supported by the Hong Kong Polytechnic University start-up grant (1-ZE8S). The authors would like to express thanks to the editor and three anonymous reviewers for their constructive comments and suggestions.

References

Ajami, N.K., Duan, Q., Sorooshian, S., 2007. An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. Water Resour. Res. 43 (1), W01403.

Box, G.E.P., Cox, D.R., 1964. An analysis of transformations. J. R. Stat. Soc. Series B Stat. Methodol. 26 (2), 211−252.

Bulygina, N., Gupta, H., 2011. Correcting the mathematical structure of a hydrological model via Bayesian data assimilation. Water Resour. Res. 47 (5), W05514.

Cammalleri, C., Ciraolo, G., 2012. State and parameter update in a coupled energy/hydrologic balance model using ensemble Kalman filtering. J. Hydrol. 416−417, 171−181.

Clark, M.P., Rupp, D.E., Woods, R.A., Zheng, X., Ibbitt, R.P., Slater, A.G., Schmidt, J., Uddstrom, M.J., 2008. Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. Adv. Water Resour. 31 (10), 1309−1324.

Crow, W.T., Loon, E.V., 2006. Impact of incorrect model error assumptions on the sequential assimilation of remotely sensed surface soil moisture. J. Hydrometeorol. 7 (3), 421−432.

Dechant, C., Moradkhani, H., 2011. Radiance data assimilation for operational snow and streamflow forecasting. Adv. Water Resour. 34 (3), 351−364.

DeChant, C.M., Moradkhani, H., 2012. Examining the effectiveness and robustness of sequential data assimilation methods for quantification of uncertainty in hydrologic forecasting. Water Resour. Res. 48 (4), W04518.

DeChant, C.M., Moradkhani, H., 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.

Evensen, G., 1992. Using the extended Kalman filter with a multilayer quasi-geostrophic ocean model. J. Geophys. Res.−Oceans 97 (C11), 17905−17924.

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

Evensen, G., 2003. The Ensemble Kalman Filter: theoretical formulation and practical implementation. Ocean Dynam. 53 (4), 343−367.

Gharamti, M. E., Hoteit, I., Valstar, J., 2013. Dual states estimation of a subsurface flow-transport coupled model using ensemble Kalman filtering. Adv. Water Resour. 60, 75−88.

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

Han, J., Huang, G., Zhang, H., Li, Z., Li, Y., 2014. Heterogeneous precipitation and streamflow trends in the Xiangxi River watershed, 1961–2010. J. Hydrol. Eng. 19 (6), 1247–1258.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. J. Basic Eng. 82 (1), 35−45.

Khan, U.T., Valeo, C., 2016. Short-term peak flow rate prediction and flood risk assessment using fuzzy linear regression. J. Environ. Inform. 28 (2), 71−89.
Liu, D., Mishra, A.K., Yu, Z., 2016. Evaluating uncertainties in multi-layer soil moisture estimation with support vector machines and ensemble Kalman filtering. J. Hydrol. 538, 243−255.
Liu, X.M., Huang, G.H., Wang, S., Fan, Y.R., 2016. Water resources management under uncertainty: Factorial multi-stage stochastic program with chance constraints. Stoch. Env. Res. Risk. A. 30, 945–957.
Liu, Y., Gupta, H.V., 2007. Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. Water Resour. Res. 43 (7), W07401.
McMillan, H.K., Hreinsson, E.Ö., Clark, M.P., Singh, S.K., Zammit, C., Uddstrom, M.J., 2013. Operational hydrological data assimilation with the recursive ensemble Kalman filter. Hydrol. Earth Syst. Sci. 17, 21−38.
Misirli, F., Gupta, H.V., Sorooshian, S., Thiemann, M., 2013. Bayesian Recursive Estimation of Parameter and Output Uncertainty for Watershed Models. in Calibration of Watershed Models, edited, pp. 113−124, American Geophysical Union.
Montgomery, D., 2000. Design and Analysis of Experiments, 5th Edition. John Wiley & Sons, New York.
Montgomery, D.C., Runger, G.C., 2013. Applied Statistics and Probability for Engineers. Sixth Edition ed., John Wiley & Sons, New York.
Moore, R.J., 1985. The probability-distributed principle and runoff production at point and basin scales. Hydrolog. Sci. J. 30 (2), 273−297.
Moore, R.J., 2007. The PDM rainfall-runoff model. Hydrol. Earth Syst. Sci. 11 (1), 483–499.
Moradkhani, H., Sorooshian, S., Gupta, H.V., Houser, P.R., 2005. Dual state–parameter estimation of hydrological models using ensemble Kalman filter. Adv. Water Resour. 28 (2), 135−147.
Moradkhani, H., DeChant, C.M., Sorooshian, S., 2012. Evolution of ensemble data assimilation for uncertainty quantification using the particle filter-Markov chain Monte Carlo method. Water Resour. Res. 48 (12), W12520.
Nash, J. E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. J. Hydrol. 10 (3), 282−290.
Panzeri, M., Riva, M., Guadagnini, A., Neuman, S.P., 2014. Comparison of Ensemble Kalman Filter groundwater-data assimilation methods based on stochastic moment equations and Monte Carlo simulation. Adv. Water Resour. 66, 8−18.
Pathiraja, S., Marshall, L., Sharma, A., Moradkhani, H., 2016a. Hydrologic modeling in dynamic catchments: A data assimilation approach. Water Resour. Res. 52 (5), 3350−3372.
Pathiraja, S., Marshall, L., Sharma, A., Moradkhani, H., 2016b. Detecting non-stationary hydrologic model parameters in a paired catchment system using data assimilation. Adv. Water Resour. 94, 103−119.
Rafieeinasab, A., Seo, D.-J., Lee, H., Kim, S., 2014. Comparative evaluation of maximum likelihood ensemble filter and ensemble Kalman filter for real-time assimilation of streamflow data into operational hydrologic models, J. Hydrol. 519, 2663−2675.
Randrianasolo, A., Thirel, G., Ramos, M.H., Martin, E., 2014. Impact of streamflow data

assimilation and length of the verification period on the quality of short-term ensemble hydrologic forecasts. J. Hydrol. 519, 2676−2691.

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

Ryu, D., Crow, W.T., Zhan, X., Jackson, T.J., 2009. Correcting unintended perturbation biases in hydrologic data assimilation. J. Hydrometeorol. 10, 734−750.

Samuel, J., Coulibaly, P., Dumedah, G., Moradkhani, H., 2014. Assessing model state and forecasts variation in hydrologic data assimilation. J. Hydrol. 513, 127−141.

Shen, Y.S., Tan, Z.F., Shen, X.L., Bai, J.J., Li, Q.Z.,Wang, S., 2012. Study of energy saving and emission reduction based on the OLAP multi-indicator relational model. J. Environ. Inform. 20 (2), 115−122.

Sun, A.Y., Morris, A., Mohanty, S., 2009. Comparison of deterministic ensemble Kalman filters for assimilating hydrogeological data. Adv. Water Resour. 32 (2), 280−292.

Thiboult, A., Anctil, F., 2015. On the difficulty to optimally implement the Ensemble Kalman filter: An experiment based on many hydrological models and catchments. J. Hydrol. 529, 1147−1160.

Vrugt, J. A., Diks, C.G.H., Gupta, H.V., Bouten, W., Verstraten, J.M., 2005. Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. Water Resour. Res. 41 (1), W01017.

Wang, D., Chen, Y., Cai, X., 2009. State and parameter estimation of hydrologic models using the constrained ensemble Kalman filter. Water Resour. Res. 45 (11), W11416.

Wang, S., Huang, G.H., Baetz, B.W., 2015a. An inexact probabilistic–possibilistic optimization framework for flood management in a hybrid uncertain environment. IEEE Trans. Fuzzy Sys. 23, 897–908.

Wang, S., Huang, G.H., Baetz, B.W., Ancell, B.C., 2017. Towards robust quantification and reduction of uncertainty in hydrologic predictions: Integration of particle Markov chain Monte Carlo and factorial polynomial chaos expansion. J. Hydrol. 548, 484–497.

Wang, S., Huang, G.H., Baetz, B.W., Huang, W., 2015b. A polynomial chaos ensemble hydrologic prediction system for efficient parameter inference and robust uncertainty assessment. J. Hydrol. 530, 716–733.

Wang, S., Huang, G.H., Baetz, B.W., Huang, W., 2016a. Probabilistic inference coupled with possibilistic reasoning for robust estimation of hydrologic parameters and piecewise characterization of interactive uncertainties. J. Hydrometeorol. 17, 1243−1260.

Wang, S., Huang, G.H., Zhou, Y., 2016b. A fractional-factorial probabilistic-possibilistic optimization framework for planning water resources management systems with multi-level parametric interactions. J. Environ. Manage. 172, 97-106.

Wang, S., Huang, G.H., Huang, W., Fan, Y.R., Li, Z., 2015c. A fractional factorial probabilistic collocation method for uncertainty propagation of hydrologic model parameters in a reduced dimensional space. J. Hydrol. 529 (3), 1129–1146.

Weerts, A.H., El Serafy, G.Y.H., 2006. Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models. Water Resour. Res.

42 (9), W09403.
Wu, C.F.J., Hamada, M.S., 2009. Experiments: Planning, Analysis, and Optimization. John Wiley & Sons, New Jersey.
Xie, X., Zhang, D., 2010. Data assimilation for distributed hydrological catchment modeling via ensemble Kalman filter. Adv. Water Resour. 33 (6), 678−690.
Yin, J., Zhan, X., Zheng, Y., Hain, C.R., Liu, J., Fang, L., 2015. Optimal ensemble size of ensemble Kalman filter in sequential soil moisture data assimilation. Geophys. Res. Lett. 42 (16), 6710−6715.
Young, P.C., 2013. Hypothetico-inductive data-based mechanistic modeling of hydrological systems. Water Resour. Res. 49 (2), 915−935.
Zeng, X.T., Li, Y.P., Huang, G.H., Liu, J., 2016. Modeling water trading under uncertainty for supporting water resources management in an arid region. J. Water Res. Pl.-ASCE 142 (2), 04015058.