Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: http://ees.elsevier.com

Research papers

Development of clustered polynomial chaos expansion model for stochastic hydrology prediction

F. Wang [a], G.H. Huang [a,b,*], Y. Fan [c], Y.P. Li [a]

[a] State Key Joint Laboratory of Environmental Simulation and Pollution Control, School of Environment, Beijing Normal University, Beijing 100875, China
[b] Center for Energy, Environment and Ecology Research, UR-BNU, School of Environment, Beijing Normal University, Beijing 100875, China
[c] Department of Civil and Environmental Engineering, Brunel University London, Uxbridge, Middlesex, UB8 3PH, United Kingdom

ARTICLE INFO

This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Jie Chen, Associate Editor

Keywords
Stochastic projection
Polynomial Chaos Expansion
Stepwise Cluster Analysis
Dynamic sensitivity
Multilevel Factorial Analysis

ABSTRACT

This study introduced a clustered polynomial chaos expansion (CPCE) model to reveal random propagation and dynamic sensitivity of uncertainty parameters in hydrologic prediction. In the CPCE model, the random characteristics of the streamflow simulations resulting from parameter uncertainties are characterized through the polynomial chaotic expansion (PCE) model based on the probabilistic collocation method. At the same time, a multivariate discrete non-functional relationship between PCE coefficients and hydrological model inputs is established based on stepwise cluster analysis. Therefore, compared with traditional PCE method, the developed CPCE model cannot only reflect uncertainty propagation in stochastic hydrological simulation, but also have the capability of random forecasting. Moreover, the dynamic sensitivities of model parameters are investigated through the multilevel factorial analyses. The developed approach was applied for streamflow forecasting for the Ruihe watershed, China. Results showed that with effective quantification for the random characteristics of hydrological processes, the CPCE model can directly predict runoff series and generate the associated probability distributions at different time periods. The dynamic sensitivity analysis indicates that the maximum soil moisture capacity within the catchment plays a key role in the accuracy of the low-flow forecasting, while the degree of spatial variability in soil moisture capacities has a remarkable impact on the accuracy of the high-flow forecasting in the studied watershed.

1. Introduction

The hydrologic system is random in nature; in other words, its behaviors change with the time consistent with the law of probability as well as the sequential relationship between the occurrences of the system (Chow and Kareliotis, 1970). Conventional hydrological models (e.g., distributed model, conceptual hydrological model) do not closely represent the natural stochastic processes, which may produce results that can hardly match the behaviors of the hydrologic system accurately (Vinogradov et al., 2011). That is, the random features of the system can hardly be well reflected in the traditional modelling process (Lindenschmidt and Rokaya, 2019). Such an overlook of the system randomness would further lead to unreliable hydrological predictions, limiting the applicability of hydrological models to many real-world water resources issues (Khaiter and Erechtchoukova, 2019; Lu et al., 2017). Thus, great efforts are desired to reveal these stochastic features and analyze their impacts on resulting predictions in the hydrologic system (Papalexiou et al., 2011; Wang et al., 2020a)

Previously, there were many studies in developing mathematical models for revealing the stochastic features of hydrological processes. For instance, Chow and Kareliotis (1970) treated the watershed as the stochastic system and represented the system components, including precipitation, evapotranspiration, storage and runoff, through time series models. In terms of research progress in mathematical modeling research, Singh and Woolhiser (2002) reported a comprehensive review and provided a short synopsis of used models. Given the inherent complexity of watershed system, recent studies have utilized more advanced stochastic methods, such as Bayesian analysis (Kavetski et al., 2006) and data assimilation (Fan et al., 2017a), for uncertainty

* Corresponding author at: Institute for Energy, Environment and Sustainable Communities, University of Regina, Regina, Saskatchewan, Canada.
E-mail address: huang@iseis.org (G.H. Huang)

https://doi.org/10.1016/j.jhydrol.2021.126022
Received 3 August 2020; Received in revised form 6 November 2020; Accepted 18 January 2021
Available online xxx
0022-1694/

quantification of hydrological predictions. In particular, to alleviate the computational burden encountered in stochastic simulation, the surrogate modeling techniques have gained much attention in hydrological modeling and water resource management (Razavi et al., 2012; Rui et al., 2013). A surrogate model can be considered as "a simple surrogate of a complex model" which describes the relationship between model parameters and model outputs (e.g. streamflow) (Chowdhury, 2019; Wang et al., 2014). Some widely used surrogate model methods entail polynomial chaos expansion (PCE) (Hu et al., 2019; Wang et al., 2015), artificial neural networks (ANN) (Yan and Minsker, 2006), Gaussian processes regression (GPR) (Marrel et al., 2008), support vector machine (SVM) (Zhang et al., 2009), and sparse polynomial chaos(SPC) (Wang et al., 2020b). Razavi et al. (2012) reported a comprehensive review for research progress in surrogate models in the water resources field. It has shown that a surrogate modeling, when applied correctly, is able to perform as well as the original model. Recently, surrogate modeling-based uncertainty quantification methods have been proposed in hydrological simulations. For example, Fan et al. (2015a) investigated the applicability of second- and third-order PCE in quantifying uncertainties of the conceptual hydrologic predictions. Fan et al. (2015b) combined the capabilities of the ensemble Kalman filter and the probabilistic collocation method to quantify the uncertainty of hydrologic models. Meng and Li (2018) assessed the model uncertainties quantitatively through running PCE surrogate model rather than solving governing equations or performing the simulator repetitively. Many studies reported that the surrogate model based optimization is an efficient way to reduce the computational burden of parameter optimization of a hydrological model (Gong et al., 2016; Gou et al., 2020; Wang et al., 2020b).

However, the surrogate models may not represent the studied system accurately when strong nonlinearity/irregularity presents in the modeling responses (Meng and Li, 2018). Moreover, these methods have been applied mostly to uncertainty quantification of the modeling processes and the application of stochastic approaches in hydrological predictions has barely begun. The application of stochastic approaches in forecasting processes need more research (Ghaith and Li, 2020; Xiang et al., 2019). Besides, the coefficients of PCE surrogate models are mainly quantified based on observed model responses (i.e., streamflow) in previous research works, which prevent the PCE model from offering hydrological predictions. Therefore, as an extension of previous studies, the objective of this study is to develop a clustered polynomial chaos expansion (CPCE) model for revealing the stochastic features in hydrologic predictions. The objective entails: (i) characterization of the random characteristics in hydrological process through the PCE model based on the probabilistic collocation method, (ii) establishment of relationship between the PCE model parameters and the input variables based on stepwise cluster analyses (SCA), (iii) demonstration of the predictability of the CPCE model for the Ruihe watershed, China, and (iv) investigation of the dynamic sensitivity of model parameters on probabilistic predictions through multilevel factorial analyses.

2. Development the clustered polynomial chaos expansion (CPCE) model

Fig. 1 presents the specific framework of clustered polynomial chaos expansion (CPCE) model. In detail, stochastic hydrological analysis is carried out based on the PCE and the probabilistic collocation method according to the probability characteristics of the model parameter $\boldsymbol{\theta}$. The distribution of parameter $\boldsymbol{\theta}$ is firstly transformed into a standard normal distribution by the gaussian anormophosis method, and a PCE model based on Hermite polynomial is established. Through the probabilistic collocation method, a series of regression models are established to obtain the coefficients of PCE. Then, a stepwise cluster analysis is adopted to characterize the discrete relationship between the meteorological data (e.g., precipitation, temperature, wind speed, and evapotranspiration) and PCE coefficients. Finally, based on the established cluster tree and the PCE model, the probabilistic daily flow can be obtained in the forecasting period. The alogical inference graph is presented in Fig. 1(b) to illustrate the procedures for determining PCE coefficients in the forecasting period. The model output y (e.g., streamflow) of the hydrological model (e.g., Hymod) is determined by forcing data × (e.g., precipitation, potential evapotranspiration), parameter vector $\boldsymbol{\theta}$ (e.g., maximum soil moisture capacity), and state vector β (e.g., soil moisture storage), which can be conceptualized as:

$$y = f(x,\beta,\theta). \tag{1}$$

Since the state variable is continuously updated in daily iteration process, its specific value depends on the previous climatic (e.g., precipitation, tempretural, potential evapotranspiration) and hydrological inputs (Chen et al., 2011; Ghaith et al., 2020; Wang et al., 2018; Zhong et al., 2016). Therefore, the state variable can be assumed as a function of the previous forcing data (e.g., multi-day cumulative precipitation):

$$\beta_t \approx g\left(x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4} \ldots \middle| \theta\right) \tag{2}$$

where t is the time. The PCE surrogate model describes the relationship between hydrological model parameter and streamflow (i.e., model output), which can be conceptualized as:

$$y = h(a,\theta) \tag{3}$$

where a are coefficients of PCE. Combining Eqs. (1), (2), and (3), when the model parameters are the same, the PCE coefficients can be expressed as a function of the previous climate:

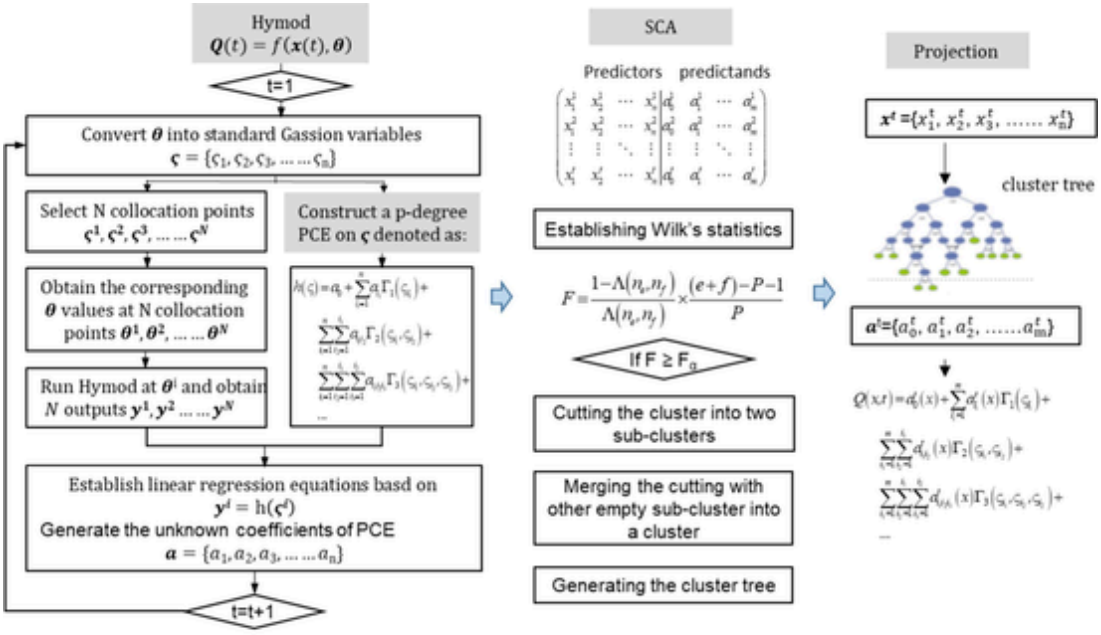
$$f\left(x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4} \ldots |\theta\right) \approx h\left(a_t|\theta\right) \tag{4}$$

2.1. Polynomial chaos expression (PCE)

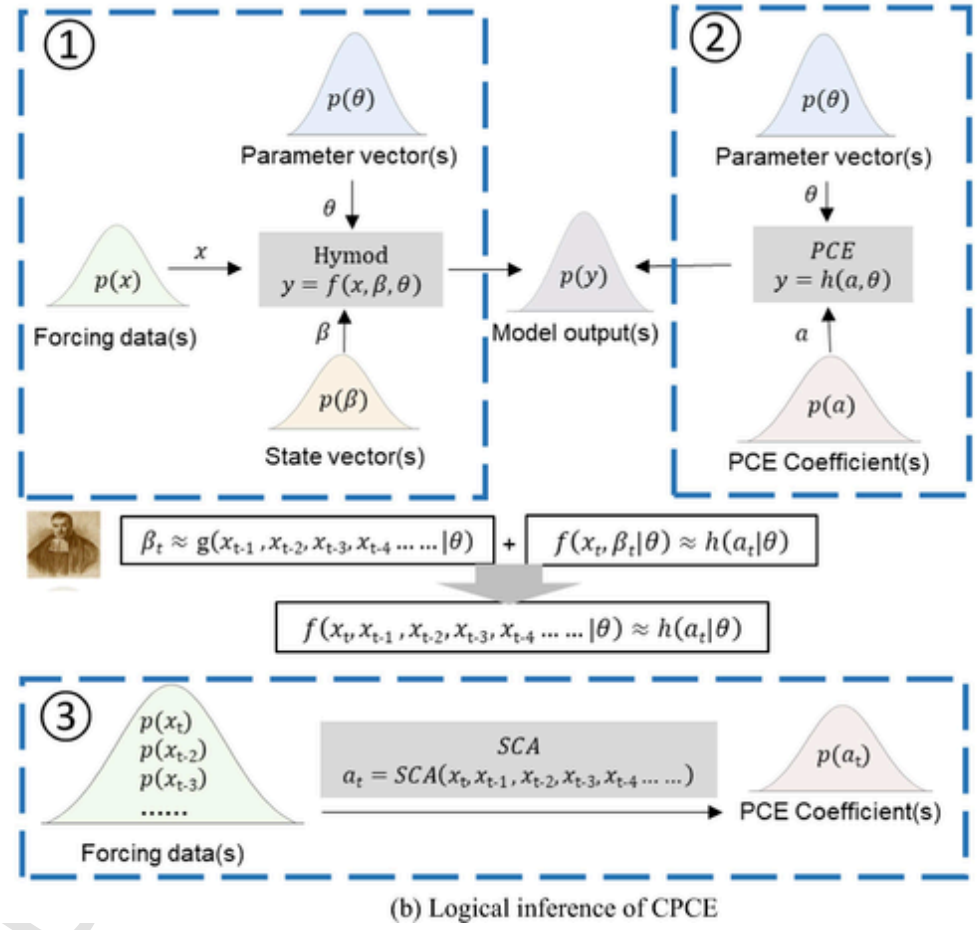
Generally, the output of a hydrological model is expressed by a nonlinear function of stochastic inputs (i.e., presents as a set of random variables). To express the evolution of uncertainty in dynamical system with random inputs, Wiener (1938) introduced the theory and application of polynomial chaos (PC) comprehensively. For Gaussian random variables, the modeling stochastic process can be decomposed by Hermite polynomials (Fan et al., 2015a), and the general polynomial chaos expansion (PCE) can be expressed as follows:

$$\begin{aligned} y = a_0 &+ \sum_{i_1=1}^{n} a_{i_1}\Gamma_1\left(x_{i_1}\right) + \sum_{i_1=1}^{n}\sum_{i_2}^{i_1} a_{i_1 i_2}\Gamma_2\left(x_{i_1}, x_{i_2}\right) \\ &+ \sum_{i_1=1}^{n}\sum_{i_2}^{i_1}\sum_{i_3=1}^{i_2} a_{i_1 i_2 i_3}\Gamma_3\left(x_{i_1}, x_{i_2}, x_{i_3}\right) + \ldots \end{aligned} \tag{5}$$

where y is the output of a model, $\Gamma_p(x_{i_1}, x_{i_2}, \ldots, x_{i_p})$ is the p^{th} order polynomial in terms of the multi-dimensional random variables, and $a_0, a_{i_1}, a_{i_1 i_2}, a_{i_1 i_2 i_3}, \cdots$ are the unknown coefficients to be determined. Take the standard normal variable as an example, the



(a) Framework of the CPCE



(b) Logical inference of CPCE

Fig. 1. Framework (a) and Logical inference (b) of the Clustered Polynomial Chaos Expansion (CPCE) model.

Hermite polynomial can be expressed as:

$$\Gamma_p\left(x_{i_1}, x_{i_2}, \ldots, x_{i_M}\right) = (-1)^p e^{1/2\mathrm{x}^T\mathrm{x}} \frac{\partial^M}{\partial x_{i_1} \partial x_{i_2} \ldots \partial x_{i_M}} e^{-1/2\mathrm{x}^T\mathrm{x}} \quad (6)$$

where $(x_{i_1}, x_{i_2}, \ldots, x_{i_p})$. ($\boldsymbol{x}$ is the vector form) present the standard normal random variables. Consequently, Eq. (6) is often written in a simple formulation as:

$$y = a_0 + \sum_{i=1}^{\infty} a_i \Gamma_i(x) \quad (7)$$

where a_i (i = 0, 1, …) are the unknown coefficients to be determined. The truncated PCE model is generally applied based on the dimension of the random variables and the highest order of the Hermite polynomials. If the p^{th} order Hermite polynomials is involved, the truncated PCE for M dimensional random variables can be expressed as (Fan et al., 2016a):

$$y \approx a_0 + \sum_{i=1}^{n-1} a_i \Gamma_i(x) \quad (8)$$

where $n = \frac{(M+p)!}{M!p!}$. Table S2 gives some explicit values of n for given dimensions (i.e. M) and degrees (i.e. p) of PCE. For example, the second order Hermite polynomials with two random variables are $\{1, \zeta_1, \zeta_2, \zeta_1^2 - 1, \zeta_2^2 - 1, \zeta_1\zeta_2\}$. Then, according to Eq (8), the expansion of truncated PCE for a second-order two-random-variables model is presented as follows:

$$y \approx a_0 + a_1\zeta_1 + a_2\zeta_2 + a_3\left(\zeta_1^2 - 1\right) + a_4\left(\zeta_2^2 - 1\right) + a_5\zeta_1\zeta_2 \quad (9)$$

To determine the unknown coefficients contained in the PCE model, the probabilistic collocation method (PCM) was used in previous studies (Fan et al., 2015a). The PCM is executed through approaching a model output with a PCE in terms of random inputs. In other words, the basic idea of PCM is to let the PCE responses equal to the model simulations at selected collocation points. Collocation points can be specified by the algorithm proposed by Webster et al. (1996) in this research. In this algorithm, the collocation points are selected from combinations of the roots for the higher-order Hermite polynomial (Huang et al., 2007). Moreover, the regression based collocation method is employed to obtain the unknown coefficients in PCE (Huang et al., 2007). More detail of PCM are supplied in (Fan et al., 2015a).

2.2. Estimation of PCE coefficients through stepwise cluster analysis

The PCE model can be adopted to replace the original hydrological model for uncertainty quantification and propagation. Nevertheless, it can hardly do the prediction work since no streamflow observations are available to estimate the coefficients of PCE in the forecasting period. To address this challenge, the stepwise cluster analysis (SCA) will be introduced into PCE, leading to a clustered polynomial chaos expansion (CPCE) model. In detail, nonlinear relationships between hydrological model inputs (e.g., precipitation, potential evapotranspiration) and PCE coefficients are established through SCA in the historical period where streamflow observations are available. Then, the PCE coefficients in the forecasting period can be estimated with meteorological data and the trained SCA model. Finally, the PCE surrogate model can be adopted to generate the probabilistic streamflow predictions. More details for this procedure is provided in the supporting materials.

SCA is a multivariate analysis tool which can simulate variations of single or multiple dependent variables (y_s) with a group of independent variables (x_s) (Huang, 1992). According to the theory of multivariate analysis of variance (MANOVA), the sample sets of independent-dependent variables (i.e., x_s-y_s) pairs are divided into groups through a series of cutting and merging operations (Li et al., 2015; Wang et al., 2013). According to (Huang, 1992), the cutting and merging procedures are conducted through the F-test based Wilks' likelihood ratio criterion (Wilks, 1963). For example, assume a cluster $A_{m\times n}$ contains m samples of n dimensional predictors. The cluster $A_{m\times n}$ can be cut into two sub-clusters $A^1_{\alpha\times n}$. and $A^2_{\beta\times n}$, where $\alpha + \beta = m$ based on Wilks' likelihood ratio criterion. The value of Wilks' statistic Λ can be calculated as follows:

$$\Lambda = \frac{|W|}{|W + B|} \quad (10)$$

where W is the within-groups sums of squares and cross products matrices; B is the between-group sums of squares and cross products.

$$W = \sum_{i=1}^{p}\left(A^1_{\alpha\times n} - \bar{A}^1_{\alpha\times n}\right)^{\mathrm{T}}\left(A^1_{\alpha\times n} - \bar{A}^1_{\alpha\times n}\right) + \sum_{i=1}^{q}\left(A^2_{\beta\times n} - \bar{A}^2_{\beta\times n}\right)^{\mathrm{T}}\left(A^2_{\beta\times n} - \bar{A}^2_{\beta\times n}\right) \quad (11)$$

$$B = \frac{\alpha\beta}{\alpha + \beta}\left(\bar{A}^1_{\alpha\times n} - \bar{A}^2_{\beta\times n}\right)^{\mathrm{T}}\left(\bar{A}^1_{\alpha\times n} - \bar{A}^2_{\beta\times n}\right) \quad (12)$$

$\bar{A}^1_{\alpha\times n}$ and $\bar{A}^1_{\beta\times n}$ are the sample means of sub-clusters $A^1_{\alpha\times n}$ and $A^2_{\beta\times n}$, respectively:

$$\bar{A}^1_{\alpha\times n} = \frac{1}{\alpha}\sum_{i=1}^{p} A^1_{\alpha\times n} \quad (13)$$

$$\bar{A}^2_{\beta\times n} = \frac{1}{\beta}\sum_{i=1}^{p} A^2_{\beta\times n} \quad (14)$$

$|W|$ and $|W + B|$ indicate the determinants of matrices. The smaller the Λ value is, the larger the difference between the sub-clusters of $A^1_{\alpha\times n}$ and $A^2_{\beta\times n}$. The cutting point is optimal, if and only if the value of Λ is minimum (Huang, 1992; Wilks, 1963). On the contrary, sub-clusters $A^1_{\alpha\times n}$ and $A^2_{\beta\times n}$ cannot be cut if the Λ value is very large, which may be merged into a new cluster. By Rao's F approximation (Rao et al., 1973), we have R-statistic as following:

$$R = \frac{1 - \Lambda^{1/S}}{\Lambda^{1/S}} \frac{ZS - P(K-1)/2 + 1}{P(K-1)} \quad (15)$$

where K is the number of groups and P is the number of predictors. Z and S can be calculated as following:

$$Z = m - 1 - (P + K)/2 \quad (16)$$

$$S = \frac{P^2 \times (K-1)^2 - 4}{P^2 + (K-1)^2 - 5} \quad (17)$$

Here, $K = 2$ (two sub-clusters $A^1_{\alpha\times n}$ and $A^2_{\beta\times n}$) and the R-statistic will be an exact F-variate:

$$F(P, m - P - 1) = \frac{1 - \Lambda}{\Lambda} \times \frac{m - P - 1}{P} \quad (18)$$

Therefore, the process of cutting and merging clusters is conducted through a number of F tests (Martin and Maes, 1979;

Rao et al., 1973; Var, 1998). In other words, the F test is used to identify whether sub-clusters $A^1_{\alpha\times n}$ and $A^2_{\beta\times n}$ are significantly different. Cluster $A_{p\times n}$ can be cut into two sub-clusters $A^1_{\alpha\times n}$ and $A^2_{\beta\times n}$ if $F(P, m-P-1)$ is larger than $F_{p-cutting}$. The p-cutting is the significance level of cutting process. On the other hand, the F test could also be used to identify whether any two sub-clusters are similar between each other. For two clusters $A^i_{\alpha'\times n}$ and $A^j_{\beta'\times n}$ with samples of α' and β', if $F\left(P, \alpha'+\beta'-P-1\right)$ is smaller than $F_{p-merging}$, the two clusters can be merged into a new cluster. The p-merging is the significance level of merging process.

As shown in Fig. 2, the main procedures to estimate the PCE coefficients through SCA are as follows: (1) Select forcing data of the hydrological model (i.e., x_s) and PCE coefficients (y_s), prepare a training matrix, and set the significant level p (default 0.05); (2) Rank sample sets of x_s-y_s pairs according to the value of $x_{.j}$ (the j^{th} variable); (3) Cut the matrix as two groups from row t randomly; (4) Estimate the Wilks' statistic Λ value of the two groups; (5) Repeat steps 2–4 for all j and t to find the minimum Λ; (6) Cut operation is performed if $F\left(P, \alpha'+\beta'-P-1\right)$ is larger than $F_{p-merging}$; (7) Cut loops until all hypotheses of further cut are rejected; (8) Estimate Wilks' statistic Λ value of ys' difference of any two child-clusters; (9) Merge the two clusters into one if $F\left(P, \alpha'+\beta'-P-1\right)$ is smaller than $F_{p-merging}$; (10) Repeat steps 8–9 for all clusters; (11) Merging loops until all hypotheses of further merge are rejected; (12) Repeat cut-merge to the end when hypotheses of further cut or merge are rejected; (13) Generate the cluster tree of the training samples; (14) The PCE coefficients are predicted according to the generated cluster tree when new forcing data are available.

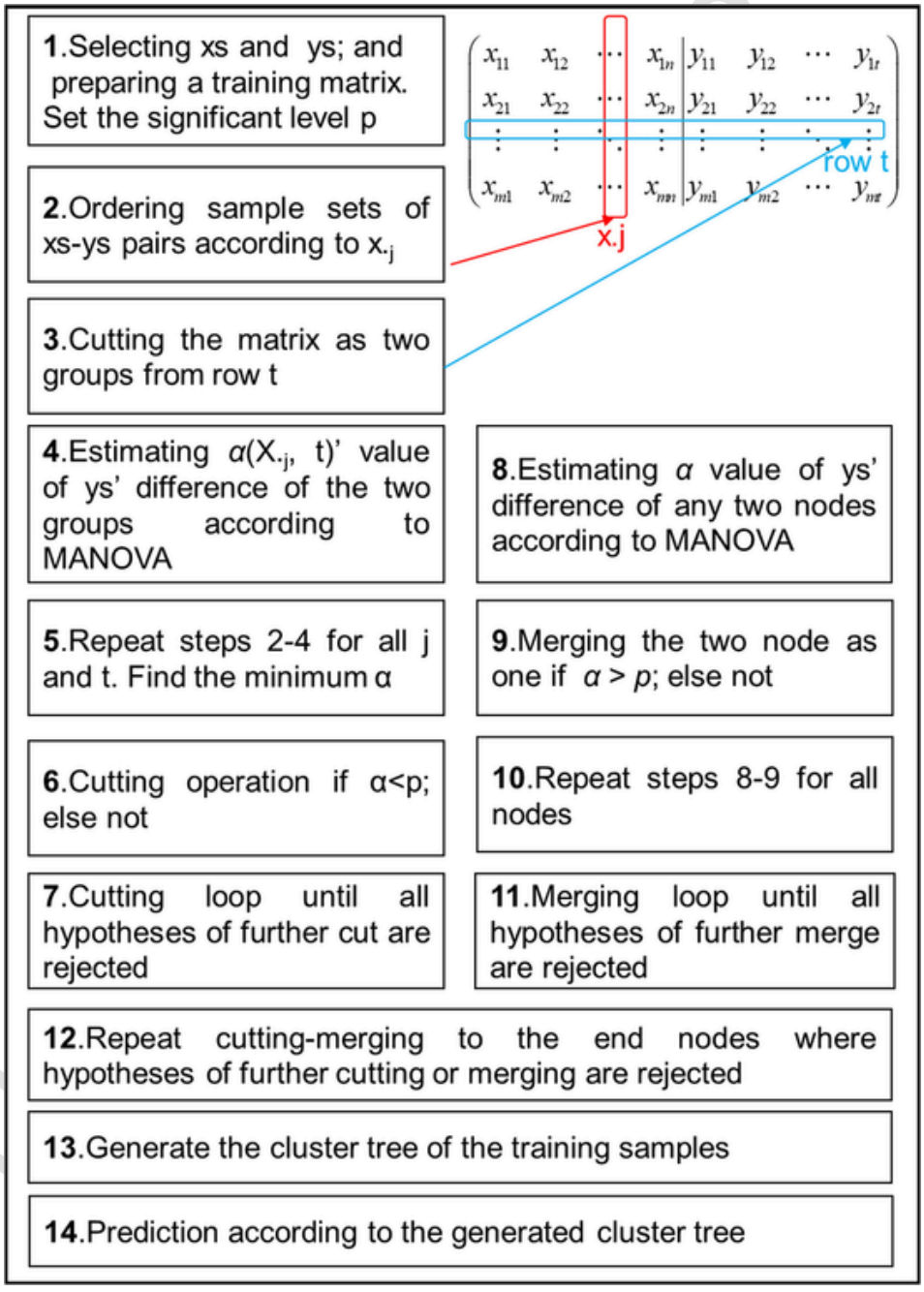


Fig. 2. Procedures of stepwise cluster forecasting method (SCA).

Here, a simple example has been provided to show the structural of SCA model in the supporting material. More detailed description of the SCA method can be accessible from relevant studies by (Cheng et al., 2016; Fan et al., 2016b; Huang, 1992; Huang et al., 2006).

2.3. Multilevel factorial analysis

Multilevel factorial analysis (MFA) is one of powerful tools to quantify the effects of several independent variables as well as their interactions on a dependent variable (Duan et al., 2020). In order to use the same terminology to present MFA approach, a generalized three-variable model is defined as $\psi = \mathrm{F}(A, B, C)$. A, B, C represent the independent variables (e.g., inputs, parameters, or model structure) and Ψ represents the response (e.g., the model performances). Assume there are a, b, c elements (or levels) for variables A, B and C, respectively. According to the analysis of variance (ANOVA) theory, the total sum of the squares (SST) can be divided into the sum of squares due to individual variable and their interactions as follows (Saltelli et al., 2010, 2008):

$$SST = SS_A + SS_B + SS_C + SS_{AB} + SS_{AC} + SS_{BC} + SS_{ABC} \tag{19}$$

where SS_A, SS_B, SS_C respectively represents the square due to the individual effect of factor A, B, and C; $SS_{AB}, SS_{BC}, SS_{AC}, SS_{ABC}$ represent the squares due to interactions. In this model, all interaction terms are summarized into the term SS_{int}.

$$\begin{aligned} SS_{\mathrm{int}} &= SS_{AB} + SS_{AC} + SS_{BC} + SS_{ABC} \\ &= SST - SS_A - SS_B - SS_C \end{aligned} \tag{20}$$

Then, for each effect, the variance fractions η^2 are derived as follows:

$$\begin{aligned} \eta^2_{\mathrm{int}} &= \frac{SS_{\mathrm{int}}}{SST} \\ \eta^2_A &= \frac{SS_A}{SST} \\ \eta^2_B &= \frac{SS_B}{SST} \\ \eta^2_C &= \frac{SS_C}{SST} \end{aligned} \tag{21}$$

$$\eta^2_A + \eta^2_B + \eta^2_C + \eta^2_{\mathrm{int}} = 1 \tag{22}$$

The value of η^2 indicates the contribution of independent variable and their interaction to the total variance.

$$SST = \sum_{A=1}^{a}\sum_{B=1}^{b}\sum_{C=1}^{c}(\psi^{A,B,C} - \psi^{o,o,o})^2 \tag{23}$$

$$\begin{aligned} SS_A &= \sum_{A=1}^{T_1}\sum_{B=1}^{T_2}\sum_{C=1}^{T_n}(\psi^{A,o,o} - \psi^{o,o,o})^2 \\ SS_B &= \sum_{A=1}^{T_1}\sum_{B=1}^{T_2}\sum_{C=1}^{T_n}(\psi^{o,B,o} - \psi^{o,o,o})^2 \\ SS_C &= \sum_{A=1}^{T_1}\sum_{B=1}^{T_2}\sum_{C=1}^{T_n}(\psi^{o,o,C} - \psi^{o,o,o})^2 \end{aligned} \tag{24}$$

The symbol "o" indicates averaging over the particular index. For a general N-variable model, the calculation process of factorial analysis is shown in the supporting material.

3. Overview of research area

3.1. Complexity of Ruihe watershed

The developed CPCE model is applied to address the stochastic hydrological predictions for the Ruihe River, which is one of the major tributaries of Jinghe River. Ruihe River is located in the middle of the Loess Plateau in China, as shown in Fig. 3. The catchment area of the Ruihe River is 1670 km^2 with a mainstream length of 119 km and the elevation varying from 1200 to 2645 m. The annual mean temperature is 8.8 °C, and the annual mean precipitation is 562 mm with approximate 60–70% of precipitation occurs between June and September. The Ruihe River plays a vital role in reducing soil erosion, alleviating water loss, and protecting the ecosystem of the middle reach of the Jinghe River. Hydrological modeling in this watershed would benefit the description of hydrological processes and water resources management.

In this study, the climatic and hydrological data of the Ruihe River watershed were collected from different sources. The areal daily precipitation (P) data were interpolated from site precipitation measurements distributed over the catchment, and the areal potential daily evapotranspiration (E) were interpolated from the national meteorological stations. The streamflow observations of the Ruihe River (from 1981 to 1987) were obtained from Yuanjiaan hydrometric station as shown in Fig. 3(b). The time series of streamflow, precipitation, and potential evapotranspiration of Ruihe River basin are presented in Fig. 3(c). It can be found that the climatic and hydrological data remained stable during the study period. Here, it is assumed that the relationship between climate conditions and streamflow has not changed significantly in this short term (seven years). Thus, the historically normal years from 1981 to 1984 will be used as the calibration period in this study. The input topography map was collected to derive the flow direction and hydrographic network. DEM (1:250,000) is obtained from Data Center for Resources and Environmental Sciences, Chinese Academy of Sciences (RESDC) (http://www.resdc.cn).

3.2. Hymod

In this study, rainfall–runoff simulation is undertaken by using a simple conceptual model-Hymod, which is proposed by Moore (1985). Based on the probability distribution of soil moisture, water storage dynamics are built to represent the mass balance principles among inflow from rainfall, losses to evaporation, drainage to groundwater, and production of direct runoff. In brief, assume that a catchment consists of infinite points and each of them can be defined by a soil moisture capacity. Soil moisture capacities vary within the catchment as a result of variability in soil texture and depth. Therefore, to describe the soil moisture variability of the whole catchment, a cumulative distribution function can be presented as follows:

$$F(c) = 1 - \left[1 - \frac{c}{C\max}\right]^{B\exp} \quad 0 \leqslant c \leqslant C\max \tag{25}$$

where c is soil moisture capacity, Cmax is the maximum storage capacity, and the exponent Bexp is the degree of spatial variability of storage capacity over the basin. A schematic representation of Hymod is shown in Fig. 4, where Alpha is introduced to represent how much of the subsurface runoff is routed over the fast (Rq) and slow (Rs) pathways. The descriptions of the five parameters and their prior fluctuating ranges are shown in Table1. The five parameters of Hymod, including Cmax, Bexp, Alpha, Rs, and Rq can be obtained through a model calibration process. In this study, the prior range of the five parameters are obtained from previous literatures (Fan et al., 2017a, 2017b, 2015b). These prior distributions are assumed spread over a wide range to represent vague prior knowledge. Input data including daily precipitation, P (mm/d), and daily potential evapotranspiration, E (mm/d), are used to drive Hymod. For the sake of brevity, we recommend to access the studies (Moore, 1985, 2007) for a comprehensive description of this model. For Hymod, a second-order two-random-variables PCE will be built as:

$$\begin{aligned} Q \approx\ & a_0 + a_1 * C\max' + a_2 * B\exp' + a_3 * \left(C\max'^2 - 1\right) \\ & + a_4 * \left(B\exp'^2 - 1\right) + a_5 * C\max' * B\exp' \end{aligned} \tag{26}$$

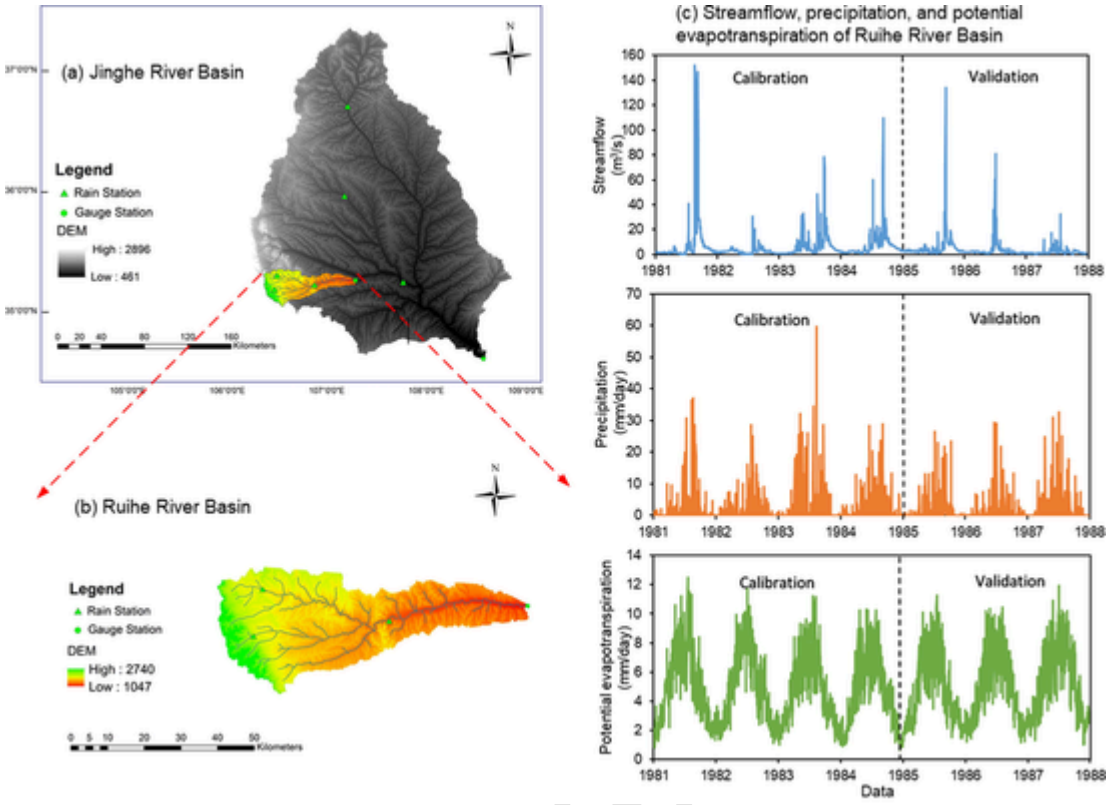


Fig. 3. The location of the studied catchments: (a) the Jing River basin, (b) the Ruihe River basin, and (c) presents the time series of streamflow, precipitation, and potential evapotranspiration of Ruihe River basin. The green triangles show the location of national rain stations used to generate potential evapotranspiration and areal precipitation, and the green circle shows the location of the streamflow station. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

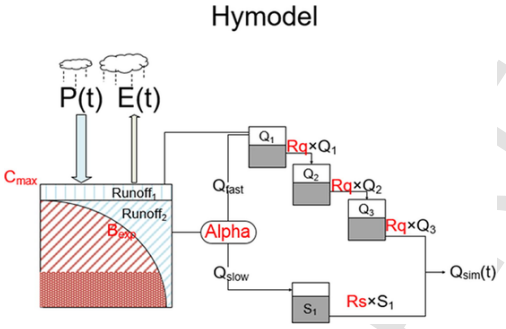


Fig. 4. Schematic representation of Hymod.

here, Q presents the streamflow, $Cmax'$ and $Bexp'$ are standard Gaussian variables which are transformed from Cmax and Bexp.

In this study, Nash–Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970) is used to evaluate the deterministic prediction of the hydrological modeling.

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}\left(Q_{obs,i} - Q_{sim,i}\right)^2}{\sum_{i=1}^{N}\left(Q_{obs,i} - \bar{Q}_{obs}\right)^2} \tag{27}$$

where N is the total number of observations (or predictions); $Q_{obs,i}$ is the observed stream;.$Q_{sim,i}$. is the estimated streamflow; $\bar{Q}_{obs}$ is the mean of all observations.

The probabilistic prediction is evaluated by continuous ranked probability score (CRPS), which is defined as the integrated squared difference between the CDF of forecasts and observations (Fan et al., 2017a; Hersbach, 2000; Murphy and Winkler, 1987):

$$CRPS = \int_{-\infty}^{+\infty}\left[\mathrm{F}^{\mathrm{sim}}(x) - \mathrm{F}^{\mathrm{obs}}(x)\right]^2 dx \tag{28}$$

where $\mathrm{F}^{\mathrm{sim}}$ and $\mathrm{F}^{\mathrm{obs}}$ are CDFs for predicted streamflow and observed streamflow, respectively. A small CRPS value indicates a better model performance, with the value of zero suggesting a perfect accuracy for model prediction.

Table 1
The descriptions of Hymod parameters.

Parameter	Description	Prior range	Posterior range
Cmax	Maximum soil moisture capacity within catchment (mm)	[0, 700]	[100,123]
Bexp	Shape factor that is dependent on the degree of spatial variability in soil moisture capacities	[0.01,15]	[0.26, 0.32]
Alpha	Fraction coefficient for distribution of water between slow and quick reservoirs	[0.5,0.9]	0.58
Rq	Inverse of residence time in quick reservoirs (1/day)	[0.01,0.2]	0.022
Rs	Inverse of residence time in a slow reservoir (1/day)	[0.3,0.7]	0.58

4. Influence of the parameter distribution on surrogate system

To evaluate the uncertainty of Hymod predictions, the second-order PCE ($Q \approx a_0 + a_1 * Cmax' + a_2 * Bexp' + a_3 * \left(Cmax'^2 - 1\right) + a_4 * \left(Bexp'^2 - 1\right) + a_5 * Cmax' * Bexp'$) is built

with the historical data, and then applied to quantify the uncertainty of the Hymod predictions. As presented in Table S3, nine collocation points are obtained to establish the linear regression equations and generate the values of the six unknown coefficients in the second-order PCE (Fan et al., 2015a). In the prediction period, 50 values of Cmax' and Bexp' are randomly generated through the standard Gaussian distribution which are transformed from Cmax and Bexp. Thus, total 2,500 PCE simulation values would be generated for each time point with the generated Cmax' and Bexp', as well as PCE coefficients. The mean and variance values of CPCE model would be obtained with the 2500 realizations. To evaluate PCE's reliability for uncertainty quantification, the Monte Carlo (MC) simulation, is applied as the benchmark results. Similarly, 50 values of Cmax and Bexp are randomly generated from their distribution randomly, and total 2500 combinations of parameter values are applied to run Hymod simulations for MC analysis. And again, the mean and varianance values of MC model would be obtained. In this study, the prior distributions for the parameters (i.e. Cmax and Bexp) of Hymod are assumed to uniformly spread over predefined ranges to represent vague prior knowledge as presented in Table 1. Posterior distributions of Cmax and Bexp are obtained by Metropolis-Hastings (MH) algorithm through the calibration of Hymod over the period of 1981–1984. After the calibration process, the parameters posterior distributions are obtained as shown in Fig. 5. The detail of MH and the calibration process are provided in the supporting material.

To analyze the influence of the parameter distributions on the surrogate system, PCE surrogate model is established based on the prior and posterior distributions of the parameters in Hymod, respectively. The comparison for the mean and variance values of daily streamflow obtained through the 2-order PCE and MC simulation methods are presented in Fig. 6. As shown in Fig. 6(a) and (b), with the parameter prior distribution, the mean streamflow obtained by PCE surrogate model are highly identical with the MC simulation results. By linearly fitting the mean values of MC simulation and PCE simulation, the coefficient of determination $R^2 = 0.999$. In Fig. 6(c) and (d), the variances of the streamflow obtained through MC and PCE simulations are compared with the corresponding coefficient of determination $R^2 = 0.995$. These indicate that the PCE results would fit well with the MC simulation results in both means and variances. However, as can be seen from Fig. S2, neither the MC simulation nor the PCE model would fit well with the observed streamflows as they are established based on the parameter prior distributions. Due to the wide prior distributions of parameters, MC simulations cannot reflect the true relationship between rainfall and runoff, and the simulation results cannot match the actual runoff observations. Not surprisingly, the PCE simulation where the collocation points are selected in the parameter prior spaces also cannot match the actual runoff observations.

In comparison, as shown in Fig. 7, the PCE surrogate model based on the posterior distributions of parameters cannot only better match the MC simulations in both means and variances, but also fit well with the runoff observations (i.e., NSE = 0.84). In other words, the PCE surrogate model can generally replace the hydrologic model (i.e. Hymod) with posterior parameter distributions to reflect the temporal variations for the streamflow. Combining the prior distribution and actual observations, the posterior distributions of parameters are obtained based on MH algorithm. The PCE surrogate model with the collocation points selected from parameter posterior distribution spaces can well match the observed runoff. These indicate that the mean values obtained through PCE surrogate model are highly identical with the MC simulation results no matter whether they are established based on the parameters' prior or posterior distributions. In other words, the PCE surrogate model can generally replace the hydrologic model (i.e. Hymod). However, only the PCE surrogate model and MC simulation based on parameters' posterior distributions can reflect the temporal variations for streamflow accurately. Therefore, the MH method would be employed to approximate the posterior probabilities of model parameters and improve the forecasting accuracy based on the observed measurements.

5. Projection of CPCE model

For the historical period, the unknown coefficients in PCE can be assessed through regression-based PCM method with model input (e.g., precipitation and evapotranspiration) and output (e.g., streamflow) (Fan et al., 2015a). However, for the forecasting period, these coefficients are unavailable due to the lack of observed streamflow (Ghaith and Li, 2020). Therefore, before exploring PCE's capability to forecast streamflow stochastically, the potential predictors of PCE coefficients need to be identified. As shown in Fig. 1(b), since the state variable (i.e., x-loss) is continuously updated, its specific value depends on the previous climatic and hydrological relationship (Xie et al., 2017). Hence, it can be assumed that the state variableis a function of the previous forcing

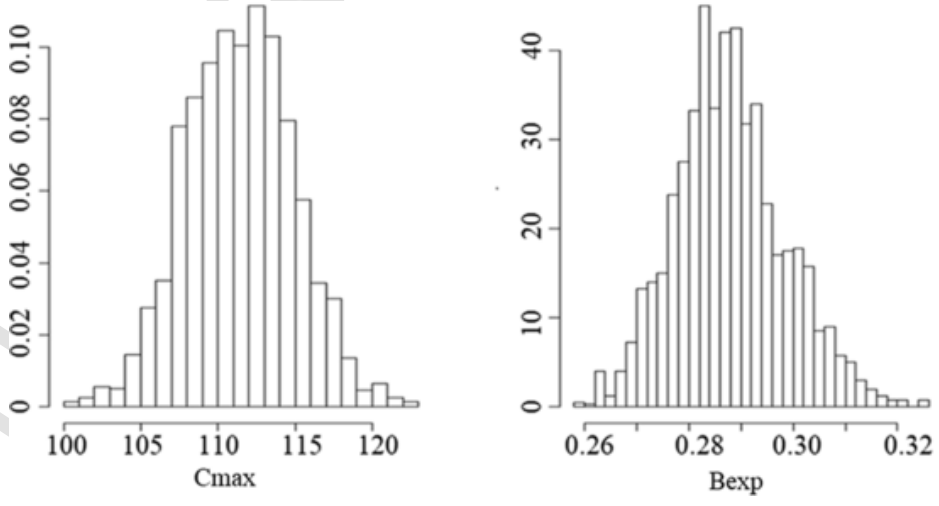


Fig. 5. Posterior distributions of Cmax and Bexp obtained by metropolis–hastings (MH) algorithm for Ruihe River.

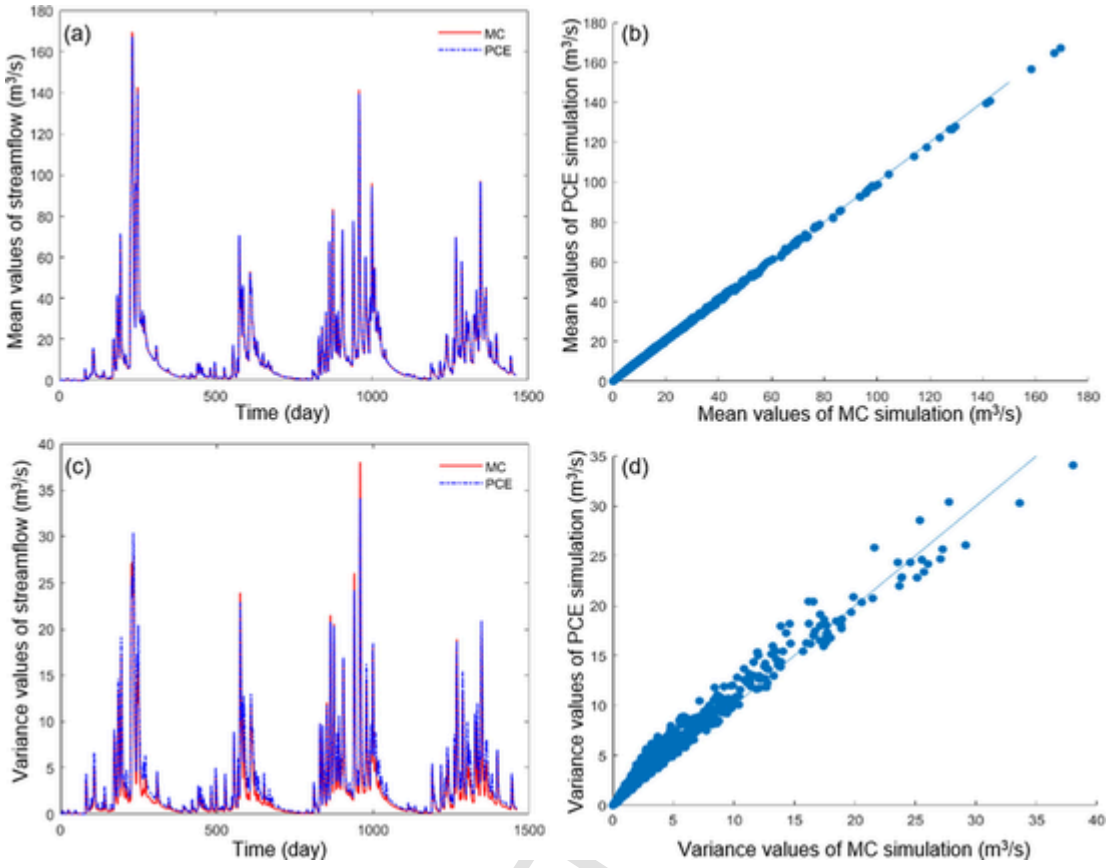


Fig. 6. The comparison of MC simulation, PCE simulation and observation based on parameters prior distribution. (a) The mean values of the streamflow obtained through MC and 2-order PCE simulation; (b) Scatter plot of the mean values of MC simulation and 2-order PCE results; (c) The variance values of the streamflow obtained through MC and 2-order PCE simulation; (d) Scatter plot of the variance values of MC simulation and 2-order PCE results.

data. For convenience, the PCE coefficients are firstly correlated with various climatic variables, including multi-day cumulative precipitation (denoted as P, sumP3, sumP5, sumP7, sumP10, sumP15, sumP20, sumP30, sumP60, and sumP90) and multi-day cumulative potential evapotranspiration (denoted as E, sumE3, sumE5, sumE7, sumE10, sumE15, sumE20, sumE30, sumE60, and sumE90). As presented in Fig. 8, there are significant correlations among the PCE coefficients and climate variables. For example, except for a_5, all PCE coefficients have strong correlations with climate variables (especially the cumulative precipitation). The biggest correlation (0.853) exists between a_2 and sumP5 (i.e., the cumulative 5-day rainfall). Therefore, a reasonable guess is that the coefficents of the PCE prediction model can be obtained based on climate factors.

The PCE coefficients and climate data during 1981–1984 are used to traing SCA model, and the obtained SCA tree is shown in Fig. S3. Then, the PCE coefficients and climate data in 1984 are used to test the obtained SCA model. As shown in Fig. 9, most of the PCE coefficients (e.g., a_0, a_2, a_4) predicted by the SCA model can fit their theoretical values. As mentioned in Section 3.2, a second-order two-random-variables PCE is built for Hymod as:

$$Q \approx a_0 + a_1 * C\max' + a_2 * B\exp' + a_3 * \left(C\max'^2 - 1\right) + a_4 * \left(B\exp'^2 - 1\right) + a_5 * C\max' * B\exp'$$

. Here, $C\max'$ and $B\exp'$ are standard Gaussian variables which are transformed from the distributions of Cmax and Bexp. The variability in PCE coefficients implies the different main (and/or interactive) effects of these parameters on model outputs (i.e., streamflow). For instance, the high predictability of a_2 (NSE = 0.80) indicates that the model performance is more sensitive to Bexp (as shown in Fig. 12). At the same time, it is difficult to predict a_5 accurately (NSE = 0.14), which means that the interaction between Cmax and Bexp has little effect on the model results. This also corresponds to the result of Fig. 12, which shows that the interactive effects of Cmax and Bexp in CPCE model are lower than 0.004. However, the relationship between PCE coefficients and climate variables is unknown, and has been rarely studied and not quantified in previous studies (Ghaith and Li, 2020). This study attempts to construct and quantify this relationship through the SCA approach for the first time. The advantage of SCA is that it can quantify nonlinear non-functional relationships (Huang, 1992). As shown in Fig. S3, the unknown relationships between PCE coefficients and climatic conditions can be revealed through the loop of cutting and merging, which is described through a clustering tree. With this trained tree, SCA model is expected to predict the unknown coefficients in PCE for the forecasting period with obtained climatic conditions. The accuracy of PCE coefficient predictions by SCA reflects the strength of these relationships. Using SCA to analyze the relationship between PCE coefficients and climate variables is one of the advantages of CPCE method.

The mean and variance values of streamflow obtained by MC and CPCE during the validation period are presented in Fig. 10. It is found that the predicted streamflows obtained by CPCE can fit very well with the results from MC. When compared with observed streamflows, MC and CPCE performed well for low streamflow while produced underestimations for the peak streamflows. This may be due to the peak clipping phenomenon of SCA method. For the validation period, NSE values of MC and CPCE are 0.79 and 0.57. Fig. 11 presents the ensembles of the forecasted streamflow

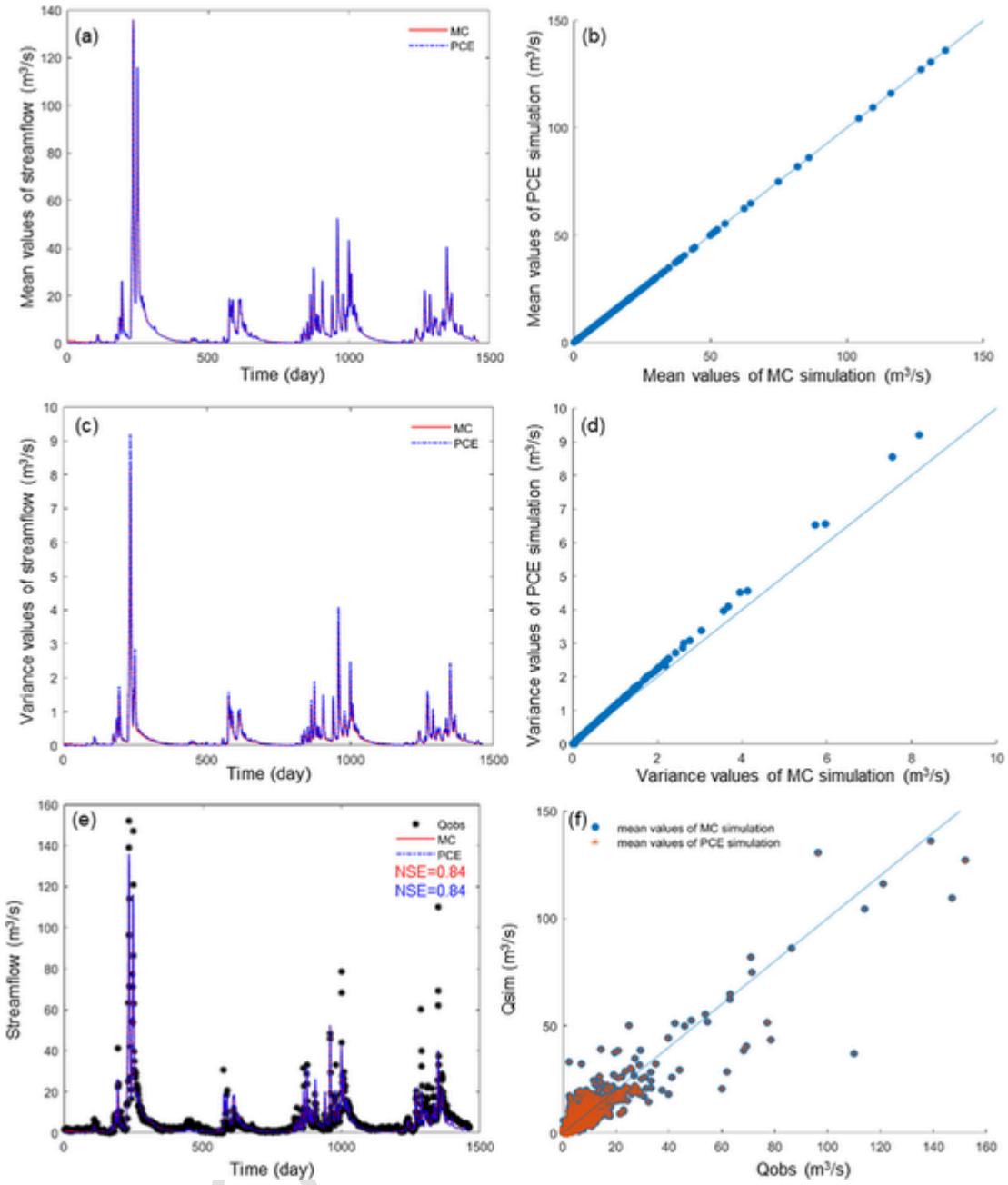


Fig. 7. The comparison of MC simulation, PCE simulation and observation under the posterior distribution of parameters obtained by AM. (a) The mean values of the streamflow obtained through MC and 2-order PCE simulation; (b) Scatter plot of the mean values of MC simulation and 2-order PCE results; (c) The variance values of the streamflow obtained through MC and 2-order PCE simulation; (d) Scatter plot of the variance values of MC simulation and 2-order PCE results; (e) The mean values of simulation and observation; (f) Scatter plot of the mean values of simulation and observation.

obtained by MC and CPCE during the validation period. The distributions show that the probability distributions generated by CPCE are very similar with those obtained by MC. For the validation period, CRPS values of MC and CPCE are 1.9 and 2.9, respectively. Similar to deterministic predictions, CPCE is likely to provide probabilistic forecasts reliably for low streamflow while generate underestimations for the peak streamflows. To make it clear, the histograms of streamflow obtained by MC and CPCE on several selected days are presented in Fig. S4. The different performances of the CPCE reflect the inconsistency in the process of uncertainty transmission. In CPCE model, the PCE coefficients are obtained through the SCA tree, where the predicted value is the average of the samples in each leaf node. For low runoff or normal runoff with a large number of samples, the samples of each leaf node are concentrated, and the average value can reflect the predicted value well. However, for flood peak with small sample size, the average value of leaf node may underestimate the predicted value and the uncertainty of prediction is amplified. This resaerch is the first attempt to establish the relationship between PCE coefficients and climatic conditions through SCA model and the traditional SCA method is chosen here. In the following research, the authors will consider using modified SCA (Fan et al., 2016b) to improve the prediction ability of CPCE further.

One major issue of the proposed CPCE model is its computation demand, since training a SCA tree would require a considerable computer time for repeating cut-merge. As presented in

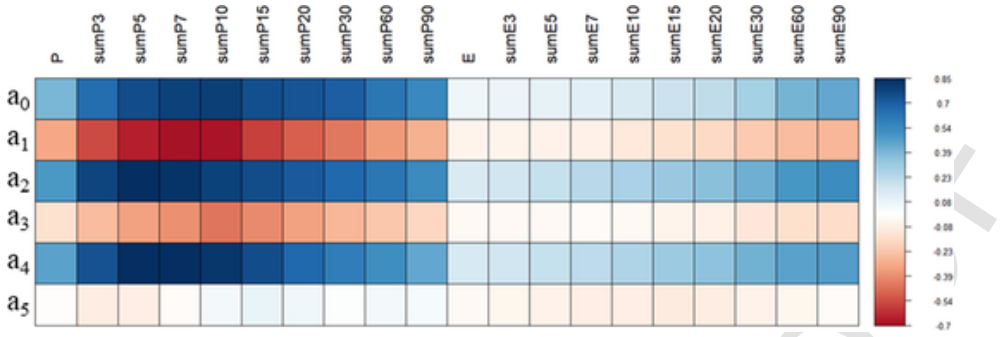


Fig. 8. Correlations between PCE coefficients and climatic conditions.

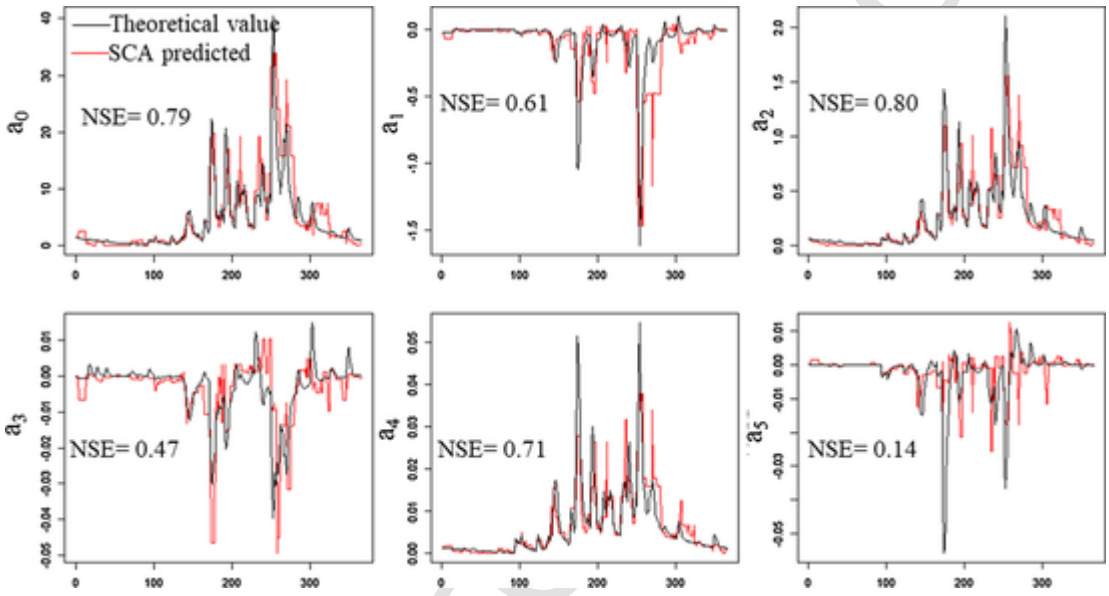


Fig. 9. The PCE coefficient predicted by SCA model.

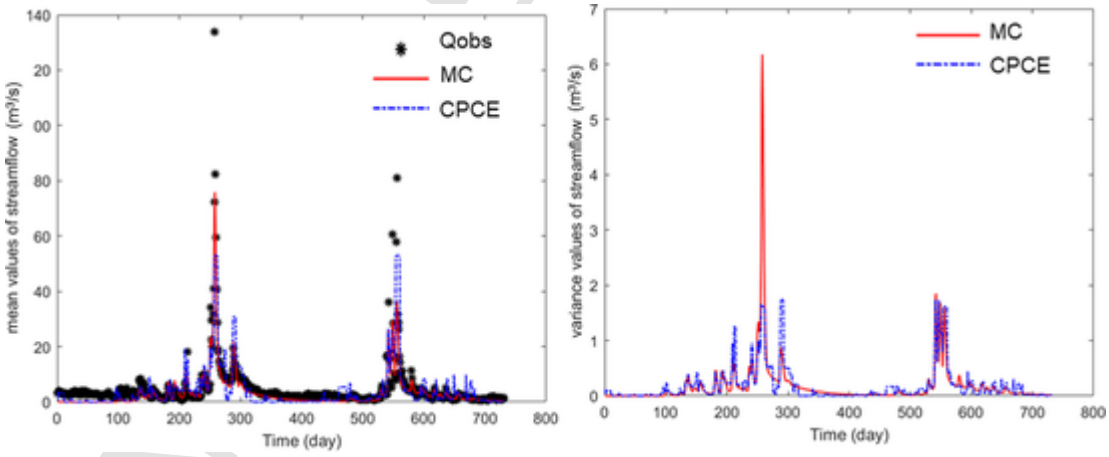


Fig. 10. Time series of streamflow obtained by MC and CPCE during the validation period. (a) mean values; (b) variance values.

Table 2, if the above CPCE procedures (i.e., include training and prediction) are applied for Hymod the computation demand would be about 481.22 and 1.54 (s) for training and prediction, and these would be 65.88 and 65.35 (s) for MC. The results show that CPCE needs more computation time than MC for this simple five-parameter Hymod. In CPCE model, SCA training does not need to be repeated, and random predictions can be quickly obtained once the SCA model is built. The main calculation requirements of CPCE for prediction are not affected by the complexity of the hydrological model. While, MC needs to run the hydrological model repeatedly, and its calculation time increases rapidly as the complexity of the model increases. Thus, the acceptable computational requirements and reliable performances make CPCE be an appealing hydrological tool in revealing interactive rainfall-runoff relationships (Ghaith and Li, 2020). Compared with traditional PCE method, the developed CPCE model cannot only reflect un-

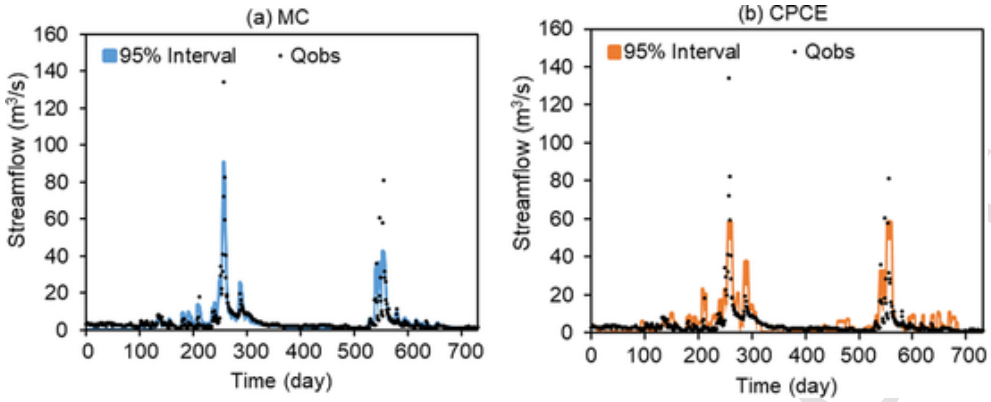


Fig. 11. Comparison between the ensembles of the forecasted streamflow obtatined by (a) MC and (b) CPCE during the validation period.

certainty propagation in stochastic hydrological simulation, but also have the capability of random forecasting. It has reported that the performance of SCA is sensitive to the clustering significance (i.e., *p* value), which determines the threshold value for cutting and merging operations (Qin et al., 2008; Sun et al., 2018). Therefore, to explore the sensitivity of CPCE on clustering significance (i.e., *p* value), three clustering significance values (i.e., $p = 0.01$, 0.05, and 0.10) are examined. It is found that the generated PCE coefficients are insensitive with *p* values (Fig. S5). The CPCE has the best prediction skill when *p* value equals to 0.05 (Fig. S5).

6. Dynamic sensitivity

Since the CPCE variables (i.e., Cmax and Bexp) have different contributions to the variability of the model outputs, a multilevel factorial analysis is introduced to quantify their main and interactive effects on uncertainty propagation (Fan et al., 2020). As an illustration, the functional approximation of a stochastic system with two-dimension second-order Hermite polynomials can be presented as: $Q \approx a_0 + a_1 * C\max' + a_2 * B\exp' + a_3 * \left(C\max'^2 - 1\right) + a_4 * \left(B\exp'^2 - 1\right) + a_5 * C\max' * B\exp'$. Here, $C\max'$ and $B\exp'$ are standard Gaussian variables which are transformed from *Cmax* and *Bexp*. Thus a 5^2 factorial experiment can be conducted to explore the sensitivities of ζ_1(Cmax) and ζ_2(Bexp) in CPCE model for each day. Fig. 12 shows the dynamic sensitivities of the parameters in the CPCE model on the streamflow. Among the 2553 daily streamflows, 70.2% of them have a main effect of Cmax higher than 0.8. In comparison, only 7.6% of daily streamflows have a main effect of Bexp higher than 0.5. As shown in Fig. 12, Cmax mainly has high main effects for low-flows, and Bexp would have high main effects for high-flows. In Hymod, Cmax indicates the maximum soil moisture capacity within the catchment, and Bexp presents the degree of spatial variability in soil moisture capacities. The conceptual hydrological model-Hymod adopts the saturation excess runoff generation mechanism (Moore, 1985). Saturation-excess overland flow is generated when the soil becomes saturated to the extent that additional precipitation cannot infiltrate. Runoff production at a point in the catchment is controlled by the maximum soil moisture capacity and precipitation (Li and Ishidaira, 2012), which is more likely to occur during small precipitation period. Thus, it can be concluded that the maximum soil moisture capacity within the catchment has the strongest main effect, which plays a key role in the forecasting streamflow especially for low-flows. Any change in the value of Cmax would cause a considerable variation in the accuracy of the low-flow forecasting. While for the high streamflow, the spatial variability in soil moisture capacities has a remarkable impact on the accuracy of hydrological forecasting. This can be explained by that saturation-excess flow generation mechanism may has the limited contribution to the total runoff during intensive storm events. It has been reported that the extreme rainfall event is the dominant trigger of annual maximum runoff events in the Yellow River basin (Ran et al., 2020). Extreme rainfall generally lasts for a short time. Under this condition, the spatial variability of river basins may have certain influence on the time and magnitude of runoff generation.

7. Conclusions

Based on the uncertainty evolution and sensitivity analysis of the hydrological process, this study introduced a clustered polynomial chaos expansion (CPCE) model for stochastic hydrological prediction. With effectively reflecting the stochastic characteristics of hydrological processes, CPCE model can directly predict future runoff sequences and generate probability distributions of all runoff values. The CPCE model characterizes the random characteristics of state variables in the hydrological model through a polynomial chaotic expansion (PCE) model based on the probabilistic collocation method. At the same time, the multivariate discrete non-functional relationship between PCE coefficients and hydrological model inputs (i.e. precipitation and potential evapotranspiration) is established based on stepwise cluster analysis. The CPCE model integrates the PCE and SCA approaches to generate the probabilistic predictions for streamflow. Finally, the dynamic parameters' sensitivities for streamflow forecasting are assessed with the help of multilevel factorial analyses.

The main conclusions are as follows: (1) The PCE surrogate model based on the posterior distributions of parameters can generally replace the hydrologic model (i.e. Hymod) to reflect the temporal variations for the streamflow. (2) The PCE coefficients can be estimated through SCA model through revealing the complicated multivariate nonlinear relation between the forcing data of hydrological model and the coefficients of the PCE model. (3) The proposed CPCE model is likely to provide reliable probabilistic predictions, which is an appealing modelling tool in revealing interactive rainfall-runoff relationships. (4) Dynamic sensitivity analysis indicates that the maximum soil moisture capacity within the catchment plays a key role in the accuracy of the low-flow forecasting while the degree of spatial variability in soil moisture capacities has a remarkable impact on the accuracy of the high-flow forecasting. The different performances of the CPCE model also reflect the inconsistency in the process of uncertainty transmission.

The main innovation of this research is the development of a clustered polynomial chaos expansion (CPCE) model which over-

Table 2
Computational time of both MC and CPCE.

	Period	Main steps	ComputationalTime (s)
MC	Training	i) Sampling 2500 from parameter prior ranges	0.01
		ii) Running hydrological model (i.e., Hymod) and calculating likelihood function for 2500 times to obtain parameter posterior distributions	65.87
	Prediction	i) Sampling 2500 from the obtained parameter posterior distributions	0.01
		ii) Runinig hydrological model (i.e., Hymod) for 2500 times	64.34
CPCE	Training	i) Solving regression equations based on probabilistic collocation method to determine PCE coefficients in the history	0.01
		ii) Training SCA model with obtained PCE coefficients and climate data	481.21
	Prediction	i) Sampling 2500 from parameter posterior distributions	0.01
		ii) Predicting PCE coefficients with trained SCA model and new climate data	1.26
		iii) Running PCE for 2500 times with predicted PCE coefficient	0.27

comes the shortcoming of the traditional PCE models in doing prediction work. The main finding is that the maximum soil moisture capacity within catchment plays a key role in the accuracy of the low-flow forecasting while the degree of spatial variability in soil moisture capacities has a remarkable impact on the accuracy of the high-flow forecasting. The results conclude that the clustered polynomial chaos expansion (CPCE) model can provide accurate predictions with less computational burden.

Moreover, some extensions, improvements, or applications are still needed to be conducted in future studies. For instance, the relationship between precipitation and runoff are assumed to be consistent during a short term (seven years in this study). Thus, the trained SCA model is not constantly updated here. However, the hydrological model also needs to consider the non-stationarity of streamflow for the long-term prediction. Thus, in the following research, the authors will consider updating the SCA model constantly during the forecasting period. The combination of dynamic SCA with PCE model is expected to update the relationship between PCE coefficient and climate conditions in real time, so as to maintain good prediction accuracy of CPCE model for long-term streamflow. Secondly, a high-dimensional high-order expansion model is not involved in this study. The total number of PCE terms increases rapidly with the dimensionality and the degree of the PCE. For instance, there are 56 PCE terms and coefficients for one five-dimension three-order PCE surrogate model. It is difficult to accurately predict 56 expansion coefficients at the same time, which may lead to inaccurate predictions. One potential solution is to remove the unimportant PCE items based on sensitivity analysis, as reported by Wang et al. (2015). Therefore, in the following work, we will also consider the combination of stepwise cluster analyses and reduced PCE models to solve the projection problem of a high-dimensional high-order expansion model.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the National Key Research and Development Plan (2016YFC0502800), the Natural Sciences Foundation (51520105013, 51679087), the 111 Program (B14008) and the Natural Science and Engineering Research Council of Canada. The data that support the findings of this study are available from https://www.researchgate.net/publication/342065388_Yuanjiaan1981-1987. The code used in this paper are available from the corresponding author upon reasonable request.

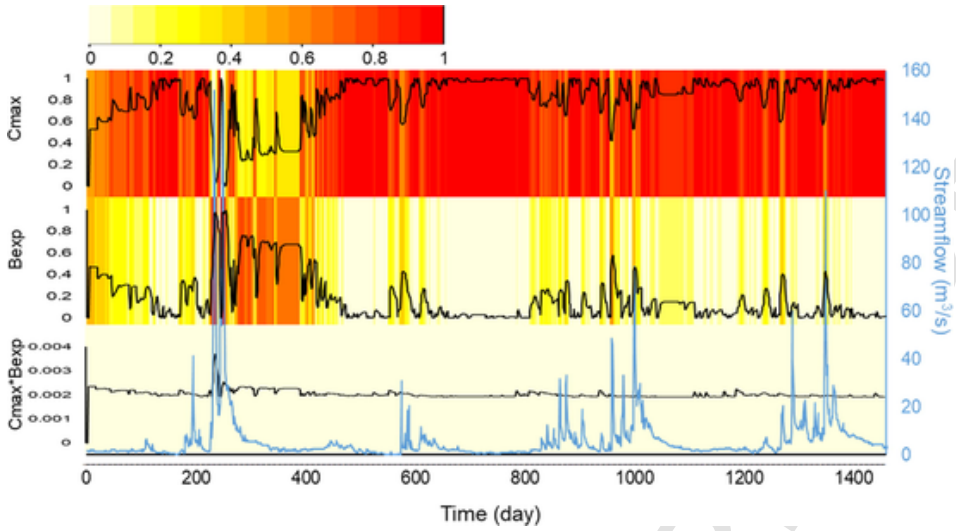


Fig. 12. Time series of the main and interactive effects of Cmax and Bexp in CPCE model. (Note: the color bars represent the strength of the main and interactive effects).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2021.126022.

References

Chen, J., Brissette, F.O.P., Leconte, R., 2011. Uncertainty of downscaling method in quantifying the impact of climate change on hydrology. J. Hydrol. 401 (3–4), 190–202.

Cheng, G., Dong, C., Huang, G., Baetz, B.W., Han, J., 2016. Discrete principal-monotonicity inference for hydro-system analysis under irregular nonlinearities, data uncertainties, and multivariate dependencies Part I: methodology development. Hydrol. Process.. doi:10.1002/hyp.10909.

Chow, V.T., Kareliotis, S., 1970. Analysis of stochastic hydrologic systems. Water Resour. Res. 6 (6), 1569–1582.

Chowdhury, K., 2019. Supervised machine learning and Heuristic algorithms for outlier detection in irregular spatiotemporal datasets. J. Environ. Inf. 33.

Duan, R.X., Huang, G.H., Li, Y.P., Zhou, X., Ren, J.Y., Tian, C.Y., 2020. Stepwise clustering future meteorological drought projection and multi-level factorial analysis under climate change: a case study of the Pearl River Basin, China. Environ. Res. 110368.

Fan, Y.R., Huang, G.H., Baetz, B.W., Li, Y.P., Huang, K., 2017a. Development of a copula-based particle filter (CopPF) approach for hydrologic data assimilation under consideration of parameter interdependence. Water Resour. Res. 53 (6), 4850–4875. doi:10.1002/2016wr020144.

Fan, Y.R., Huang, G.H., Baetz, B.W., Li, Y.P., Huang, K., Chen, X., Gao, M., 2017b. Development of integrated approaches for hydrological data assimilation through combination of ensemble Kalman filter and particle filter methods. J. Hydrol. 550, 412–426. doi:10.1016/j.jhydrol.2017.05.010.

Fan, Y.R., Huang, G.H., Baetz, B.W., Li, Y.P., Huang, K., Li, Z., Chen, X., Xiong, L.H., 2016a. Parameter uncertainty and temporal dynamics of sensitivity for hydrologic models: A hybrid sequential data assimilation and probabilistic collocation method. Environ. Modell. Softw. 86, 30–49. doi:10.1016/j.envsoft.2016.09.012.

Fan, Y.R., Huang, G.H., Li, Y.P., Wang, X.Q., Li, Z., 2016b. Probabilistic prediction for monthly streamflow through coupling stepwise cluster analysis and quantile regression methods. Water Resour. Manage. 30 (14), 5313–5331. doi:10.1007/s11269-016-1489-1.

Fan, Y.R., Huang, W., Huang, G.H., Huang, K., Zhou, X., 2015a. A PCM-based stochastic hydrological model for uncertainty quantification in watershed systems. Stoch. Env. Res. Risk Assess. 29 (3), 915–927. doi:10.1007/s00477-014-0954-8.

Fan, Y., Huang, K., Huang, G., Li, Y., Wang, F., 2020. An Uncertainty partition approach for inferring interactive hydrologic risks. Hydrol. Earth Syst. Sci. 24 (9), 4601–4624. doi:doi.org/10.5194/hess-24-4601-2020.

Fan, Y.R., Huang, W.W., Li, Y.P., Huang, G.H., Huang, K., 2015b. A coupled ensemble filtering and probabilistic collocation approach for uncertainty quantification of hydrological models. J. Hydrol. 530, 255–272. doi:10.1016/j.jhydrol.2015.09.035.

Ghaith, M., Li, Z., 2020. Propagation of parameter uncertainty in SWAT: A probabilistic forecasting method based on polynomial chaos expansion and machine learning. J. Hydrol. 586, 124854. doi:10.1016/j.jhydrol.2020.124854.

Ghaith, M., Siam, A., Li, Z., El-Dakhakhni, W., 2020. Hybrid hydrological data-driven approach for daily streamflow forecasting. J. Hydrol. Eng. 25 (2) 04019063.1-04019063.9.

Gong, W., et al., 2016. Multiobjective adaptive surrogate modeling-based optimization for parameter estimation of large, complex geophysical models. Water Resour. Res. 52 (3), 1984–2008. doi:10.1002/2015wr018230.

Gou, J., Miao, C., Duan, Q., Tang, Q., Di, Z., Liao, W., Wu, J., Zhou, R., 2020. Sensitivity analysis-based automatic parameter calibration of the VIC model for streamflow simulations over China. Water Resour. Res. 56 (1) e2019WR025968.

Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecasting 15 (5), 559–570.

Hu, J., Chen, S., Behrangi, A., Yuan, H., 2019. Parametric uncertainty assessment in hydrological modeling using the generalized polynomial chaos expansion. J. Hydrol. 579, 124158. doi:10.1016/j.jhydrol.2019.124158.

Huang, G., 1992. A stepwise cluster analysis method for predicting air quality in an urban environment. Atmos. Environ. Part B 26 (3), 349–357.

Huang, G., Huang, Y., Wang, G., Xiao, H., 2006. Development of a forecasting system for supporting remediation design and process control based on NAPL-biodegradation simulation and stepwise-cluster analysis. Water Resour. Res. 42 (6).

Huang, S., Mahadevan, S., Rebba, R., 2007. Collocation-based stochastic finite element analysis for random field problems. Probab. Eng. Mech. 22 (2), 194–205.

Kavetski, D., Kuczera, G., Franks, S.W., 2006. Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. Water resources research 42 (3).

Khaiter, P., Erechtchoukova, M., 2019. Conceptualizing an environmental software modeling framework for sustainable management using UML. J. Environ. Inf. 34 (2).

Li, Q., Ishidaira, H., 2012. Development of a biosphere hydrological model considering vegetation dynamics and its evaluation at basin scale under climate change. J. Hydrol. 412, 3–13.

Li, Z., Huang, G., Han, J., Wang, X., 2015. Development of a stepwise-clustered hydrological inference model. J. Hydrol. Eng. 20 (10), 04015008.

Lindenschmidt, K., Rokaya, P., 2019. A stochastic hydraulic modelling approach to determining the probable maximum staging of ice-jam floods. J. Environ. Inf. 34 (1).

Lu, W., Qin, X.S., Jun, C., 2017. A parsimonious framework of evaluating WSUD features in urban flood mitigation. J. Environ. Inf..

Marrel, A., Iooss, B., Van Dorpe, F., Volkova, E., 2008. An efficient methodology for modeling complex computer codes with Gaussian processes. Comput. Stat. Data Anal. 52 (10), 4731–4744.

Martin, N., Maes, H., 1979. Multivariate analysis. Academic press London.

Meng, J., Li, H., 2018. Uncertainty quantification for subsurface flow and transport: coping with nonlinearity/irregularity via polynomial chaos surrogate and machine learning. Water Resour. Res. 54 (10), 7733–7751.

Moore, R., 1985. The probability-distributed principle and runoff production at point and basin scales. Hydrol. Sci. J. 30 (2), 273–297.

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

Murphy, A.H., Winkler, R.L., 1987. A general framework for forecast verification. Monthly Weather Rev. 115 (7), 1330–1338.

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.

Papalexiou, S.-M., Koutsoyiannis, D., Montanari, A., 2011. Can a simple stochastic model generate rich patterns of rainfall events? J. Hydrol. 411 (3–4), 279–289.

Qin, X.S., Huang, G.H., Zeng, G.M., Chakma, A., 2008. Simulation-based optimization of dual-phase vacuum extraction to remove nonaqueous phase liquids in subsurface. Water Resour. Res. 44 (4). doi:10.1029/2006wr005496.
Ran, Q., Zong, X., Ye, S., Gao, J., Hong, Y., 2020. Dominant mechanism for annual maximum flood and sediment events generation in the Yellow River basin. CATENA 187, 104376.
Rao, C.R., Rao, C.R., Statistiker, M., Rao, C.R., Rao, C.R., 1973. Linear Statistical Inference and Its Applications, 2. Wiley, New York.
Razavi, S., Tolson, B.A., Burn, D.H., 2012. Review of surrogate modeling in water resources. Water Resour. Res. 48 (7).
Rui, Q., Ouyang, H., Wang, H., 2013. An efficient statistically equivalent reduced method on stochastic model updating. Appl. Math. Model. 37 (8), 6079–6096.
Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., Tarantola, S., 2010. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. Comput. Phys. Commun. 181 (2), 259–270. doi:10.1016/j.cpc.2009.09.018.
Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S., 2008. Global Sensitivity Analysis: the Primer. In: Wiley, John, Ltd, Sons (Eds.), The Atrium. Southern Gate, Chichester.
Singh, V.P., Woolhiser, D.A., 2002. Mathematical modeling of watershed hydrology. J. Hydrol. Eng. 7 (4), 270–292.
Sun, J., Li, Y.P., Gao, P.P., Suo, C., Xia, B.C., 2018. Analyzing urban ecosystem variation in the City of Dongguan: A stepwise cluster modeling approach. Environ. Res. 166, 276–289. doi:10.1016/j.envres.2018.06.009.
Var, I., 1998. Multivariate data analysis. Vectors 8 (2), 125–136.
Vinogradov, Y.B., Semenova, O., Vinogradova, T., 2011. An approach to the scaling problem in hydrological modelling: the deterministic modelling hydrological system. Hydrol. Process. 25 (7), 1055–1073.
Wang, H.M., Chen, J., Cannon, A.J., Xu, C.Y., Chen, H., 2018. Transferability of climate simulation uncertainty to hydrological impacts. Hydrol. Earth Syst. Sci. 22 (7), 3739–3759.
Wang, C., Duan, Q., Gong, W., Ye, A., Di, Z., Miao, C., 2014. An evaluation of adaptive surrogate modeling based optimization with two benchmark problems. Environ. Modell. Software 60, 167–179. doi:10.1016/j.envsoft.2014.05.026.
Wang, F., Huang, G., Fan, Y., Li, Y., 2020. Robust subsampling ANOVA methods for sensitivity analysis of water resource and environmental models. Water Resour. Manage. 34, 3199-3217.
Wang, H., Gong, W., Duan, Q., Di, Z., 2020. Evaluation of parameter interaction effect of hydrological models using the sparse polynomial chaos (SPC) method. Environ. Modell. Software 125, 104612. doi:10.1016/j.envsoft.2019.104612.
Wang, S., Huang, G.H., Baetz, B.W., Huang, W., 2015. A polynomial chaos ensemble hydrologic prediction system for efficient parameter inference and robust uncertainty assessment. J. Hydrol. 530, 716–733. doi:10.1016/j.jhydrol.2015.10.021.
Wang, X., Huang, G., Lin, Q., Nie, X., Cheng, G., Fan, Y., Li, Z., Yao, Y., Suo, M., 2013. A stepwise cluster analysis approach for downscaled climate projection–A Canadian case study. Environ. Modell. Software 49, 141–151. doi:10.1016/j.envsoft.2013.08.006.
Webster, M.D., Tatang, M.A., McRae, G.J., 1996. Application of the probabilistic collocation method for an uncertainty analysis of a simple ocean model.
Wiener, N., 1938. The homogeneous chaos. American Journal of Mathematics 60 (4), 897–936.
Wilks, S.S., 1963. Multivariate statistical outliers. Sankhyā: Indian J. Stat. Series A 407–426.
Xiang, Y., Li, L., Chen, J., Xu, C.Y., Xia, J., Chen, H., Liu, J., 2019. Parameter uncertainty of a snowmelt runoff model and its impact on future projections of snowmelt runoff in a Data-Scarce Deglaciating River Basin. Water 11 (11), 2417.
Xie, Y., Crary, D., Bai, Y., Cui, X., Zhang, A., 2017. Modeling grassland ecosystem responses to coupled climate and socioeconomic influences in multi-spatial-and-temporal scales. J. Environ. Inf..
Yan, S., Minsker, B., 2006. Optimal groundwater remediation design using an adaptive neural network genetic algorithm. Water Resour. Res. 42 (5).
Zhang, X., Srinivasan, R., Van Liew, M., 2009. Approximating SWAT model using artificial neural network and support vector machine 1. JAWRA J. Am. Water Res. Assoc. 45 (2), 460–474.
Zhong, L., Huang, G., Wang, X., Han, J., Fan, Y., 2016. Impacts of future climate change on river discharge based on hydrological inference: A case study of the Grand River Watershed in Ontario, Canada. Sci. Total Environ. 548–549, 198–210.