Parametric Uncertainty Assessment in Hydrological Modeling Using the Generalized Polynomial Chaos Expansion

Junjun Hu\textsuperscript{a}, Sheng Chen\textsuperscript{b*}, Ali Behrang\textsuperscript{c}, Huiling Yuan\textsuperscript{d}

\textsuperscript{a}Cooperative Institute for Mesoscale Meteorological Studies, The University of Oklahoma, Norman, Oklahoma, USA

\textsuperscript{b}School of Atmospheric Sciences, Sun Yat-sen University, Guangzhou, China

\textsuperscript{c}Department of Hydrology and Atmospheric sciences, University of Arizona, Tucson, Arizona, USA

\textsuperscript{d}School of Atmospheric Sciences and Key Laboratory of Mesoscale Severe Weather/Ministry of Education, Nanjing University, Nanjing, China

Submitted to Journal of Hydrology for publication

Corresponding author: Sheng Chen
Email: chenshengbj@gmail.com
Address: 135 West Xingang Road Guangzhou, Guangdong 510275, China
Abstract

An integrated framework is proposed for parametric uncertainty analysis in hydrological modeling using a generalized polynomial chaos expansion (PCE) approach. PCE represents model output as a polynomial expression in terms of critical random variables that are determined by parameter uncertainties, thus offers an efficient way of sampling without running the original model, which is appealing to computationally expensive models. To demonstrate the applicability of generalized PCE approach, both second- and third-order PCEs (PCE-2 and PCE-3) are constructed for the Xinanjiang hydrological model using three selected uncertain parameters. Uncertainties in streamflow predictions are assessed by sampling the random inputs. Results show that: 1) both PCE-2 and PCE-3 are capable of capturing the uncertainty information in hydrological predictions, generating consistent mean, variance, skewness and kurtosis estimates with the standard Monte Carlo (MC) methodology; 2) Using more collocation points and more polynomial terms, PCE-3 approximation slightly improves the model simulation and provides more matched distribution with that of MC compared to PCE-2; 3) the computational cost using the PCE approach is greatly reduced by 71% (20%) with PCE-2 (PCE-3). In general, PCE-2 is recommended to serve as a good surrogate model in future Xinanjiang hydrological modelling with much higher computation speed, more efficient sampling, and compatible approximation results.

Keywords: Polynomial Chaos Expansion, Collocation Points, Hydrological Model, Uncertainty Quantification
1. Introduction

A hydrological model is a simplified representation of the real watershed hydrological processes. It uses simple mathematical equations to conceptualize and aggregate the complex, spatially distributed, and highly interrelated water, energy, and vegetation processes in a watershed (Vrugt et al., 2005). The conceptualization and aggregation produce extensive uncertainties in model parameters and structures, thus leading to great uncertainties in hydrological predictions (Singh and Bárdossy, 2012). Therefore, it is necessary to quantify the parameter uncertainties inherent to hydrological models in water resource applications, such as flood forecasting, drought prediction, and water resource management.

Up to date, different uncertainty analysis techniques have been introduced in hydrologic literature, including Monte Carlo simulation (MC, Ang and Tang, 1984), Generalized Likelihood Uncertainty Estimation (GLUE, Beven and Binley, 1992), Bayesian recursive estimation (BaRE, Thiemann et al., 2001), the Shuffled Complex Evolution Metropolis algorithm (SCEM, Vrugt et al., 2003a), the multi-objective extension of SCEM (Vrugt et al., 2003b), the dual state-parameter estimation methods (Moradkhani et al., 2005a,b), the Simultaneous Optimization and Data Assimilation (SODA, Vrugt et al., 2005), and the model averaging methods (e.g. least squares model averaging, Diks and Vrugt, 2010). Among these, MC provides the greatest flexibility for uncertainty propagation and is widely used. With sufficiently large number of samples, stable estimates of model output distribution can be easily derived (Liu and
Gupta, 2007; Mishra, 2009). The main drawback of MC technique (and sampling-based method, e.g. GLUE) is that huge number of model simulations are required to get satisfactory estimates of the output statistics, which is a challenge for models that have high computational demand. Possible ways to overcome the problem could be: (1) using more efficient sampling scheme, such as Latin hypercube method; (2) using a faster surrogate model to replace the computationally intensive model; (3) combining the above two techniques (Khu and Werner, 2003). The polynomial chaos expansion (PCE) approach could be an alternative to surrogate-modeling scheme.

In addition to the applications in engineering and uncertainty quantification (UQ) of dynamical systems (Hu and Youn, 2011; Sanzida and Nagy, 2014), the PCE approach has received attention in hydrologic studies recently (Fan et al., 2015; Huang and Qin, 2014; Wang et al., 2015; Zheng et al., 2011). Originating from Wiener’s homogeneous chaos theory (Wiener, 1938) and generalized by Xiu and Karniadakis (2002) using Wiener-Askey polynomial chaos, PCE provides a spectral expression to the random process in terms of orthogonal polynomials. PCE is beneficial because:

1) PCE can be used to expand any second-order random process. A second-order random process \( \{X_t: t \in \mathbb{T}\} \) is one for which \( E(X_t^2) < \infty \) for all \( t \), i.e. a process with finite (bounded) variance. It applies to most physical processes (Xiu and Karniadakis, 2002).

2) The generalized PCE can handle both Gaussian and non-Gaussian random processes (Xiu and Karniadakis, 2002; Xiu and Karniadakis, 2003). The selection of
the polynomial basis depends on the distribution of the random inputs. It is shown that exponential convergence rate can be ensured if an “optimal” polynomial basis is selected (Xiu and Karniadakis, 2002).

3) PCE offers an approximation to the dynamic model, allowing for the fully probabilistic distribution of the model output. The statistical information (for example, the mean, variance, covariance, skewness, etc.) can be estimated by sampling the random variable in PCE.

4) In virtue of a polynomial expression, PCE offers an efficient way of sampling without running the original model, which is appealing particularly to computationally expensive models.

It is of interest to examine the usefulness of generalized PCE approach in quantifying the uncertainty in hydrological predictions due to uncertain parameters. In this study, the PCE approach is applied to the widely used Xinanjiang hydrological model with the assumption that three of its 15 parameters are nondeterministic. A guidance for its potential usage in quantifying the uncertainties of Xinanjiang model is proposed. The effectiveness of PCE in quantifying the modeling uncertainty is investigated by comparison against the standard MC simulations. By offering a surrogate model, the accuracy and computational cost of PCE approximation are assessed further through comparison against those from running the Xinanjiang model.

The structure of this paper is as follows: Section 2 provides a brief review of PCE methodology and the framework of parametric uncertainty analysis using generalized
PCE for UQ in hydrological modeling. In Section 3, the second- and third-order PCEs are constructed after introducing the Xinanjiang hydrological model and its parameters. Experimental results with discussions are provided in Section 4. Finally, the concluding remarks along with possible future extensions to the current study are provided in Section 5.

2. Methodology

2.1. Polynomial Chaos Expansion

The concept of PCE method was originally proposed by Wiener (1938). In Wiener’s chaos theory, it is suggested to express a second-order random process (which applies to most physical process) in terms of orthogonal polynomials (Hermite polynomials in Wiener (1938)). Some earlier studies have employed Hermite PCE to solve stochastic differential equations and demonstrated its effectiveness for Gaussian or some non-Gaussian random inputs (Ghanem, 1999; Ghanem and Spanos, 1991; Spanos and Ghanem, 1989; Xiu and Karniadakis, 2003). However, for general non-Gaussian random inputs, the convergence rate may not be optimal. By extending Hermite polynomials to polynomials in Wiener-Askey family, Xiu and Karniadakis (2002) proposed the generalized PCE. They suggested that an “optimal choice” is made if the polynomial basis is chosen based on the distribution of the random inputs according to Table 1 (adapted from Table 4.1 in Xiu and Karniadakis, 2002). For example, the Hermite polynomials are used for Gaussian random variables and the Legendre
polynomials are for uniform distributions. They further demonstrated that the
exponential convergence rate can be achieved if an optimal choice is made.

Given a dynamic model

\[ y(t) = f(x, t) \]  \hspace{1cm} (1)

where \( x \in \mathbb{R}^n \) is the model input, \( t \) is time, \( y(t) \in \mathbb{R}^m \) is the model output and
\( f(x, t) \) maps the input \( x \) into output \( y \). The output \( y(t) \) would be random if some
of the input fields are indeterminate. Generally, PCE is seeking to approximate the
random output \( y(t) \) as an orthogonal polynomial expression of a predefined random
variable \( \xi \) by

\[ y(t) = \sum_{||i||=0}^{\infty} v_i(t) \Phi_i(\xi), \]  \hspace{1cm} (2)

where \( v_i(t) \in \mathbb{R}^m \) is defined as expansion coefficient vector at time \( t \), \( \Phi_i(\xi) \) is a
Wiener-Askey polynomial in terms of a \( N \)-dimensional random variable \( \xi =
(\xi_1, \xi_2, \ldots, \xi_N) \) and \( i = (p_1, p_2, \ldots, p_N) \) is a multi-index. The \( N \)-variate
polynomial \( \Phi_i(\xi) = \Phi_{p_1p_2\ldots p_N}(\xi) \) is constructed as the product of those univariate
polynomials \( \phi_{p_j}(\xi_j) \), i.e,

\[ \Phi_i(\xi) = \prod_{j=1}^{N} \phi_{p_j}(\xi_j), \]  \hspace{1cm} (3)

where \( \phi_{p_j}(\xi_j) \) is the orthogonal polynomial in \( \xi_j \) dimension with order \( p_j \). The order
of \( \Phi_i(\xi) \) is defined as \( p = ||i|| = p_1 + p_2 + \cdots + p_N \).

In practice, approximations are made by a finite summation in a finite dimensional
space. This is accomplished by truncating the expression in equation (2) to a low-order
PCE as expressed in equation (4),
\[ \hat{y}(t) = \sum_{i=0}^{P} v_i(t) \Phi_i(\xi). \] (4)

The selection of truncation order $P$ is made according to the accuracy requirement (Wan et al., 2004).

For a $P$-th order (the highest order of the polynomial chaos) $N$-dimensional (the number of random variables) PCE, the total number of polynomial basis terms will be

\[ \binom{P+N}{N} = \frac{(P+N)!}{P!N!}. \] (5)

The vital part in PCE is to determine the expansion coefficients $v_i(t)$. Generally, there are two classes of approaches to estimate the coefficients, namely intrusive and non-intrusive methods (Oladyshkin and Nowak, 2012). Intrusive method transforms the original dynamic model to a set of equations (usually coupled) with the expansion coefficients as unknowns. The coefficients are obtained by solving the resulting equations (Hu et al., 2017). The procedure is usually cumbersome and sometimes impractical, especially for high complex nonlinear problems. In contrast, no model transformation is needed in non-intrusive method. The superiority of non-intrusive method is remarkable when the dynamic takes complicated forms and it is difficult to derive the equations for the expansion coefficients. The projection method (Le Maître et al., 2002), stochastic collocation method (Xiu and Hesthaven, 2005), regression method (Berveiller et al., 2006), and gradient-based method (Perez, 2008) are some examples of non-intrusive method. The non-intrusive method, specifically the stochastic collocation (SC) method, is used in this study due to its merit.
2.2. The Stochastic Collocation Method

The SC method assumes the PCE approximations be equal to the model outputs at predefined points (called collocation points) in random input space. The expansion coefficients are then obtained by solving a set of linear equations. The widely used strategy to specify the collocation points is based on the roots of the orthogonal polynomials adopted in PCE.

A hydrological model usually has multiple parameters and the distribution of each varies with time, study region, initial conditions, etc. The focus of this study is to demonstrate the applicability of generalized PCE approach in quantifying the uncertainties in hydrological predictions, therefore less efforts were made to study the parameters. Following previous studies (Bárdossy, 2007; Feyen et al., 2007, 2008), parameters in hydrological models are assumed uniformly distributed if their actual distributions are unknown. It is worth noting that even though uniform distribution is adopted in this study, the generalized PCE approach works for other distributions (listed in Table 1) as well. The generalized PCE approach utilizes the optimal choice of Legendre polynomials to describe uniform distribution (Table 1), which is different from previous studies where Hermite polynomials are used and transformation between uniform and normal distributions is required (Fan et al., 2015; Wang et al., 2015; Zheng et al., 2011).

The first few Legendre polynomials \((0 \leq m \leq 3, \ m \) is the order of the polynomial) with one and two random variables are defined as follows:
\{1, \xi, 3\xi^2 - 1, 5\xi^3 - 3\xi\}, \quad (6) \\
\{1, \xi_1, \xi_2, 3\xi_1^2 - 1, \xi_1\xi_2, 3\xi_2^2 - 1, 5\xi_1^3 - 3\xi_1, 3\xi_1^2\xi_2 - \xi_2, 3\xi_1\xi_2^2 - \xi_1, 5\xi_2^3 - 3\xi_1\xi_2\}, \quad (7)

For one-dimensional case, the choice of collocation points follows Gaussian quadrature rule for estimating integrals, and they are the roots of the polynomials. Specifically, suppose the truncation order is \(P\), the roots of the \((P+1)\)th-order polynomial and value of zero are selected as collocation points. For example, if \(P = 2\), the collocation points are 0, \(\sqrt{15}/5\), and \(-\sqrt{15}/5\); if \(P = 3\), the points are 0, \(\sqrt{3/7 + 2\sqrt{30}/35}\), \(\sqrt{3/7 - 2\sqrt{30}/35}\), \(-\sqrt{3/7 + 2\sqrt{30}/35}\), \(\sqrt{3/7 - 2\sqrt{30}/35}\). For multidimensional case, the tensor product is the most intuitive strategy to construct the collocation points (Xiu and Hesthaven, 2005; Zio and Fernando, 2012). As an example, Table 2 lists the collocation points for two-variate second- and third-order PCE following tensor product rule. It is worth noting that the number of collocation points following tensor product rule increases exponentially as dimension \(N\) increases. Hence, the tensor product rule is only applicable to low dimension \(N\). Several studies have been conducted on effectively constructing collocation points to alleviate the computation burden while ensuring the degree of accuracy. For example, sparse grid strategy (Jia et al., 2012; Xiu, 2007), the efficient collocation method (ECM) (Isukapalli, 1999; Villadsen and Michelsen, 1978), the regression-based Stochastic Response Surface Method (SRSM) (Isukapalli, 1999), etc. For the sake of simplicity, the tensor product rule is used in this study, as we only consider three random variables.
Once achieving the expansion coefficients, a polynomial approximation is constructed and can be used as a surrogate model. Without running the original model, ensembles of the dynamic model simulation can be easily created by sampling the random variable $\xi$ in PCE. For time-consuming models, large amount of time will be saved. One can estimate the statistics based on the ensembles and further build the histogram to estimate the probability density function (PDF).

2.3. Parameter Uncertainty Quantification for the hydrological model

For brevity, the procedure of UQ using PCE is summarized in the following:

1. Determine random input (parameters of the hydrological model in this study) of concern, denoted as $x = \{x_1, x_2, ..., x_s\}$;

2. Specify the random variable $\xi = (\xi_1, \xi_2, ..., \xi_N)$ together with polynomials $\Phi_i(\xi)$ according to the probabilistic distribution of random input $x$;

3. Construct the P-th order PCE of model output $y(t)$ as $g(\xi, t) = \sum_{|i|=0}^P v_i(t) \Phi_i(\xi)$ and express the random input $x$ as a function of the random variable $\xi$;

4. Select $L$ collocation points $\xi_1, \xi_2, ..., \xi_L$ for random variable $\xi$;

5. Obtain input values $x_j$, $j = 1, 2, ..., L$ at collocation points $\xi_j$, run the original model $f(x_j, t)$ and get $L$ outputs $y_1(t), y_2(t), ..., y_L(t)$ at each time $t$;

6. By letting $g(x_j, t) = y_j(t), j = 1, 2, ..., L$, a set of linear equations with the expansion coefficients $v_i(t), |i| = 0,1, ..., P$ as unknowns are established;
(7) Solve linear equations to obtain the expansion coefficients $v_i(t)$ at each time.

Now a surrogate model $g(\xi, t)$ for the original model $f(x, t)$ has been established, and uncertainty assessment can then be performed on $g(\xi, t)$.

2.4. Metrics for quantitative assessment

To quantitatively evaluate the UQ ability of generalized PCE approximation, the mean uncertainty error (MUE) and mean uncertainty absolute error (MUAE) of estimates on the mean value, standard deviation (STD), skewness and kurtosis are calculated relative to the standard MC simulation. The formulas are given as follows,

$$
MUE = \frac{\sum_{i=1}^{T} (u_{pc} - u_{mc})}{T},
$$

$$
MUAE = \frac{\sum_{i=1}^{T} \text{abs}(u_{pc} - u_{mc})}{T},
$$

where $T$ is the number of simulation times, $U$ stands for uncertainty index including mean value, standard deviation, skewness and kurtosis, and subscript “pc” denotes PCE approximation and “mc” refers to MC simulation.

To further quantitatively evaluate the accuracy of the PCE approximation to output variable $Q$ compared to the model simulation, metrics include the Pearson linear correlation coefficient (CC), root mean square error (RMSE), mean absolute error (MAE), Bias, and relative bias (RB) are calculated with following formulas:

$$
CC = \frac{\text{Cov}(Q_{pc}, Q_{model})}{\sigma_{pc} \sigma_{model}},
$$

$$
RMSE = \frac{\sqrt{\sum_{i=1}^{T} (Q_{pc}^i - Q_{model}^i)^2}}{T},
$$

$$
MAE = \frac{\sum_{i=1}^{T} \text{abs}(Q_{pc}^i - Q_{model}^i)}{T},
$$
Bias = \frac{\sum_{i=1}^{T}(Q_{pc}^i - Q_{model}^i)}{T}, \quad (13)

RB = \frac{\sum_{i=1}^{T}(Q_{pc}^i - Q_{model}^i)}{\sum_{i=1}^{T}Q_{model}^i}, \quad (14)

where RMSE, MAE and Bias are in m³/s, CC and RB are dimensionless; “Cov” in equation (10) refers to covariance, and \( \sigma \) stands for STD. The RB denotes the degree of overestimation or underestimation, \( i \) is the simulation time index and \( T \) is the total number of simulation times. The abbreviation “pc” stands for the simulation from PCE approximation and “model” is from the hydrological model.

3. Experimental Configurations

3.1. Study Region

The proposed framework is applied to the Tar River basin that confluences to the United States Geological Survey (USGS) gauging station “USGS02083500”. This basin is located in the northern North Carolina state (NC) (Figure 1) with latitude ranging from 35.81 °N to 36.45 °N and longitude ranging from -78.87 °W to -77.35°N. The drainage area is about 2,183 square miles with the runoff outlet at the location of 77°31'59" E and 35°53'40" N. The basin is topographically complicated with mountains and hills distributed in the western basin except the relatively flat eastern basin where the alluvial plains govern the lower reach of Tar River. The east of the basin is the Pamlico Sound which drains into Atlantic Ocean. The basin has a humid subtropical climate type, and is impacted by monsoon and hurricane precipitation...
systems that usually bring significant amount of precipitation to this basin and trigger
floods.

3.2. Hydrological Model

The hydrological model used in this study is Xinanjiang model, a well-known
physically based conceptual model developed by Zhao (1992) in 1979s. This model has
been widely used in China and other countries since its development (Chen et al., 2015).
In this model, the runoff is separated into surface, interflow and ground water
components. The core of this model is a water storage capacity distribution curve that
describes the spatial heterogeneity of tension water and free water within a basin. Such
curve has been applied in the well-known three-layer Variable Infiltration Capacity
(VIC-3L) hydrologic model (Liang, 1994; Liang et al., 1996). Xinanjiang model has 15
major parameters defined in Table 3.

Usually, not all parameters have a large impact on the model output. Sensitivity
analysis can be performed (e.g. Christiaens and Feyen, 2002) to discriminate between
sensitive and non-sensitive parameters and usually, the non-sensitive parameters are
fixed and only those sensitive parameters are assessed (Blasone et al., 2008). Study by
(Ren et al., 2010) shows that the sensitivity of the parameters in Xinanjiang model
differs with the objective function, initial conditions, flood type, and other factors.
Among the 15 parameters, three of them namely SM, CKI, and CKG are found
commonly sensitive with different objective functions. This study investigates the
uncertainty in hydrological modeling prediction under these three sensitive parameters SM, CKI, and CKG. To achieve this, SM, CKI and CKG are assumed uniformly distributed (see discussion in Section 2.2) and the remaining 12 parameters are deterministic. The values of the parameters (Table 3) are calibrated using daily streamflow observations from Tar River basin between 1 January 2002 and 31 December 2004 with the Shuffled Complex Evolution – University of Arizona (SCE-UA) algorithm (Duan et al., 1992). Note that the aim is to demonstrate the ability of PCE in quantifying the uncertainty in hydrological prediction compared to MC sampling strategy. Therefore, the value should be fine as long as it falls within reasonable range and the methodology is applicable to different parameter settings. Daily simulations from 1 January 2008 to 31 December 2009 (731 days altogether) are conducted in this study.

3.3. PCE and MC Experimental Configurations

Second-order PCE (denoted as PCE-2 hereafter) in the form of 

\[ g(\xi, t) = \sum_{|\mathbf{i}|=0}^{2} v_i(t) \Phi_i(\xi), \] 

where \( \xi = (\xi_1, \xi_2, \xi_3) \) and \( \xi_j \in [-1, 1], \ j = 1,2,3 \) is first constructed. At each time \( t \) (each day here), expression \( g(\xi, t) \) has \( \binom{2+3}{3} = 10 \) polynomial terms \( \Phi_i(\xi), |\mathbf{i}| = 0, 1, 2, \) and by tensor product, a number of \( 3^3 = 27 \) collocation points (27 model runs) are used to obtain the expansion coefficients \( v_i(t) \).

As long as the coefficients are acquired, a surrogate model in the form of PCE-2 is created. By sampling the random variable \( \xi \) (a total number of 100,000 samples in this
study), samples for PCE-2 approximation are created afterwards and further statistical analysis is performed.

It is assumed that a higher-order PCE can give better estimates, for complete study, the third order PCE (PCE-3 hereafter) approximation is also examined. Compared to PCE-2, more polynomial terms (28) are adopted in PCE-3 and correspondingly more collocation points (125) are involved to obtain the expansion coefficients. The same set of samples for $\xi$ in PCE-2 are used to create ensembles of PCE-3 approximation.

The performance of PCE in quantifying the uncertainty of hydrological modeling prediction is assessed by comparison with the classic MC simulation. In MC simulation, the values of the parameters SM, CKI, and CKG are initiated based on the 100,000 samples of $\xi$ created and the Xinanjiang model is run with these parameter values. The design of the experiment ensures that both PCE and MC use the same parameter values which helps to examine the accuracy of the PCE approximation.

4. Experimental Results

In this section, the experimental results of PCE are compared with those using the MC method in terms of uncertainty analysis, simulation accuracy, and computational cost.

4.1. Uncertainty Quantification
Figure 2 displays the mean and STD of daily streamflow simulation over the period of two years calculated from the 100,000 samples by PCE-2 approximation and MC simulation. It is noticed that both mean and STD values obtained from PCE-2 agree well with those from the MC method (Figure 2a-b). The differences of mean and STD for PCE-2 and MC methods are marginally small, with maximum difference less than 10 m³/s (Figure 2c-d). The scatter plots in Figure 3 show that the mean (STD) simulations with PCE-2 and MC lie very close to the 1:1 line. This indicates that the PCE approach has very good ability to capture the mean and standard deviation in hydrological outputs.

Similarly, the comparison of the mean and STD values of the daily streamflow from PCE-3 and MC approaches are displayed in Figure 4 and Figure 5 respectively. The difference between the mean (STD) estimates from PCE-3 and MC is very small with maximum difference less than 5 (8) m³/s (Figure 4c-d). It is noticed that the estimates from PCE-3 match better with those from MC when compared to PCE-2, especially for the STD values. As shown in Table 4, the MUAE is reduced from 0.66 (PCE-3) to 0.46 (PCE-2) for mean value, 1.33 to 0.54 for STD, 0.44 to 0.23 for skewness, and 1.46 to 0.93 for kurtosis. The MUE values exhibit similar behavior.

In order to present an overall view of the simulation distribution, Figure 6 displays the histograms constructed using the samples from PCE-2, PCE-3, as well as MC at selected days (i.e., the 148th, 323rd, and 685th days). The left column plots are from MC, those from PCE-2 are in the middle column, and the right column is from PCE-3. Here,
the performance of PCE approximation is demonstrated at high (the 685th day), low (the 148th day), and medium (the 323th day) streamflow simulations. Overall, samples from PCE-2 and PCE-3 have similar distributions to those from MC especially at those days with high streamflow outputs. The detailed statistical information including mean, STD, skewness and kurtosis listed in Table 5 indicate that both PCE-3 and PCE-2 generally capture the behavior of the hydrological model, hence they are able to give reliable uncertainty assessment to the model outputs. Compared to PCE-2, PCE-3 shows overall slightly closer estimates to the statistics obtained from MC. For example, mean value for the 323th day from MC is 166.62., PCE-2 has a value of 164.38, and PCE-3 has an estimate of 165.59. One thing needs to be kept in mind is that the computational cost also increases with a higher order PCE since more terms and collocation points are introduced. In practice, the balance between accuracy requirement and computational cost needs to be taken into account.

4.2. Simulation Accuracy

Besides examining the ability of UQ using PCE, the accuracy of PCE approximation is also assessed. The streamflow ensembles from PCE and MC at selected times in Table 5 together with CC values are displayed in Figure 7. It can be seen that the ensembles from both PCE-2 and PCE-3 match well with those from MC approach, especially for higher streamflow simulations. Overall, PCE-3 generates higher CC values. Analysis of the 2008-2009 (2-year) simulations reveals that with
PCE-2, 713 of 731 simulation days have CC value greater than 0.8 and 623 days with CC greater than 0.9; while with PCE-3, all 731 days have CC value greater than 0.8 and 721 days with CC greater than 0.9. This indicates that the difference between PCE-3 approximation and model simulation is smaller, i.e. PCE-3 is more accurate than PCE-2.

Figure 8 displays the model simulation, PCE-2 approximation and PCE-3 approximation at three different parameter settings: (a) SM = 49.32, CKG = 0.95, CKI = 0.97, (b) SM = 24.02, CKG = 0.95, CKI = 0.97, and (c) SM = 13.01, CKG = 0.99, CKI = 0.91, which are randomly chosen. It is noticed that the simulations from Xinanjiang model, PCE-2 and PCE-3 approximations almost overlap with each other, especially PCE-3 is almost identical to the model. Quantitatively, it is found that the PCE approximations are relatively well compared with Xinanjiang model simulations: high CC (>0.9), small RMSE (<6 m³/s) and MAE (<4 m³/s), and a margin of Bias (<2 m³/s) and RB (<±1%). Particularly, PCE-3 shows smaller RMSE and MAE, and a smaller margin of Bias and RB, indicating higher-order PCE outperforms low-order PCE in surrogating the Xinanjiang Model. Compared to PCE-3, PCE-2 overestimates high streamflow simulations, but not remarkable.

4.3. Computational cost

The computational time is investigated in the study as well. Under the same computational environment, it takes MC 5823 seconds to make 100,000 runs for a
period of 731 days. The computation of sampling by PCE approach is divided into two parts: the expansion coefficient calculation and sampling process. It takes PCE-2 six seconds to obtain the expansion coefficients and 1654 seconds to finish the sampling process. Therefore, about 69 minutes (71%) are saved by adopting PCE-2. As anticipated, PCE-3 needs more time: 15 seconds for coefficient acquiring and 4662 seconds for sampling. Consequently, it takes 19 minutes (20%) less by PCE-3 than MC. It is worth mentioning that PCE spends most of its time in evaluating the polynomial basis at 100,000 random input values. In fact, this task can be accomplished in advance and a lookup table could be built for reference. By excluding this part in PCE, the sampling process can be done in less than one minute, and it is irrelevant to the hydrological model, which in turn saves large amount of time. In conclusion, PCE provides an efficient way of quantifying uncertainty, which is especially appealing to complex models.

5. Conclusions

This study proposes an integrated framework for parametric uncertainty analysis in hydrological modeling by using the generalized PCE approach. The essence of this PCE approach is to expand the model as a spectral expression in terms of orthogonal polynomials, which in turn serves as a surrogate model of the specific model, like Xinanjiang hydrological model. The selection of the polynomials is determined by the probability distribution of input parameters. The generalized PCE deals with not only
Gaussian, but also non-Gaussian random inputs. Being a polynomial expression, PCE offers an efficient way of sampling for uncertainty assessment without actually running the hydrological model, which is crucial if the model runs are computationally intensive and considerably timesaving in computational efforts may be achieved. Note that the samples of the random input in PCE and values of the polynomial basis can be calculated in advance to construct a lookup table for future use, which further saves large amount of computational time.

The popular Xinanjiang hydrological model is used to demonstrate the applicability of PCE in hydrological uncertainty analysis. We used three sensitive parameters (SM, CKG and CKI) out of the 15 parameters of Xinanjiang model. In this study, uniform distribution was assumed for the three sensitive parameters and Legendre polynomials are adopted for the construction of PCE. Non-intrusive approach, specifically the SC method, is used to estimate the expansion coefficients in PCE. The collocation points are selected based on the roots of the Legendre polynomials and tensor product rule is used for constructing multi-dimensional collocation points. Daily streamflow is simulated during a period of two years (from 1 January, 2008 to 31 December, 2009) in Tar River basin. Both PCE-2 and PCE-3 are utilized to assess the uncertainties in streamflow simulation. Mean, STD, skewness, kurtosis and histograms from PCE are compared with those from the standard MC approach. Results indicate that both PCE-2 and PCE-3 provide consistent uncertainty estimates with MC. Not surprisingly, with
more terms and more collocation points, PCE-3 exhibits better performance in
capturing the uncertainty information than PCE-2.

Besides the capability of uncertainty assessment, the accuracy of the PCE
approximation of the hydrological model is investigated in this study. Overall, both
PCE-2 and PCE-3 are able to provide reliable approximations to long-term simulations
as well as extreme streamflows. Like the UQ ability, the proximity of PCE-3 is slightly
improved over PCE-2 by reducing RMSE, MAE, Bias and RB but increasing CC (Fig.
8 (a-c)). Generally, the higher accuracy is achieved with more computational costs. In
this application, the computational time decreases rapidly from 71% using PCE-2 to
20% using PCE-3 when compared with the MC method. Considering the accuracy and
computational saving, PCE-2 is suggested to use as a surrogate model of the Xinanjiang
hydrological model in the future.

In summary, the generalized PCE offers a reliable surrogate model which further
provides an efficient alternative to analyze the uncertainties in hydrological modeling.
Though uniform distribution was used to demonstrate the applicability of generalized
PCE in this study, it can be extended to other types of distribution as well. Further
studies related to hydrological modeling using generalized PCE approach can be
performed on: 1) addressing the uncertainty of complex distributed hydrological
models, 2) improving hydrological forecasts by combining PCE with data assimilation
methods, etc. These studies are underway.
Acknowledgements

The authors acknowledge the help from Dr. S. Lakshmivarahan from the University of Oklahoma, and Dr. John M. Lewis from National Severe Storms Laboratory, Norman, Oklahoma and Desert Research Institute, Reno, Nevada. This research was partially sponsored by the “100 Top Talents Program” (74110-52601108) at Sun Yat-sen University, Guangzhou, Guangdong, China, and by the National Natural Science Foundation of China (41675109, 41875182). The authors thank the editor and the anonymous reviewer for their constructive review.
References


Oladyshkin, S., Nowak, W., 2012. Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion, Reliability Engineering & System Safety, 106,


Table Captions

Table 1 The correspondence between the random variables and the type of Wiener-Askey polynomial chaos (Xiu and Karniadakis, 2002).

Table 2 Collocation points for the second- and third- order PCEs with two-dimensional random inputs.

Table 3 Parameters in Xinanjiang Model and their values used in this study.

Table 4 Comparison of statistics from PCE-2 and PCE-3 against those from MC.

Table 5 Statistics from PCE-2, PCE-3, and MC simulations at specific times.
Figure Captions

Figure 1 The study region (Tar river basin) in the eastern U.S.A.

Figure 2 The (a) mean and (b) standard deviation values of daily streamflow from PCE-2 and MC simulations and their differences for (c) mean and (d) standard deviation.

Figure 3 The scatter plots of the (a) mean and (b) standard deviation values of daily streamflow from PCE-2 and MC simulations.

Figure 4 The (a) mean and (b) standard deviation values of daily streamflow from PCE-3 and MC simulations and their differences for (c) mean and (d) standard deviation.

Figure 5 The scatter plots of the (a) mean and (b) standard deviation values of daily streamflow from PCE-3 and MC simulation.

Figure 6 Histograms generated through ensembles of MC (the left column), PCE-2 (the middle column), and PCE-3 (the right column) at the 148th day for low stream flow (a-c) indicated by red letter “A” in Fig.2 (b), the 323rd day for medium stream flow (d-f) indicated by red letter “B” in Fig.2 (b), and the 685th day for high stream flow (g-i) indicated by red letter “C” in Fig.2 (b).
Figure 4 Comparison of the ensembles generated from PCE-2 and PCE-3 against MC at specific times, (a) and (d) the 148th day, (b) and (e) the 323rd day, (c) and (f) the 685th day.

Figure 5 Daily streamflow simulation from Xinanjiang hydrological model and approximations from PCE-2 and PCE-3 using specified parameter values: (a) SM = 49.32, CKG = 0.95, CKI = 0.97, (b) SM = 24.02, CKG = 0.95, CKI = 0.97, and (c) SM = 13.01, CKG = 0.99, CKI = 0.91.
Table 1 The correspondence between the random variables and the type of Wiener-Askey polynomial chaos (Xiu and Karniadakis, 2002).

<table>
<thead>
<tr>
<th>Random variable</th>
<th>Wiener-Askey chaos</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gaussian</td>
<td>Hermite-Chaos</td>
<td>$(-\infty, \infty)$</td>
</tr>
<tr>
<td>Gamma</td>
<td>Laguerre-Chaos</td>
<td>$[0, \infty)$</td>
</tr>
<tr>
<td>Beta</td>
<td>Jacobi-Chaos</td>
<td>$[a,b]$</td>
</tr>
<tr>
<td>Uniform</td>
<td>Legendre-Chaos</td>
<td>$[a,b]$</td>
</tr>
<tr>
<td>Discrete</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poisson</td>
<td>Charlier-Chaos</td>
<td>${0, 1, 2, ...}$</td>
</tr>
<tr>
<td>Binomial</td>
<td>Krawtchouk-Chaos</td>
<td>${0, 1, ..., N}$</td>
</tr>
<tr>
<td>Negative Binomial</td>
<td>Meixner-Chaos</td>
<td>${0, 1, 2, ...}$</td>
</tr>
<tr>
<td>hypergeometric</td>
<td>Hahn-Chaos</td>
<td>${0, 1, ..., N}$</td>
</tr>
</tbody>
</table>
Table 2 Collocation points for the second- and third-order PCEs with two-dimensional random inputs.

<table>
<thead>
<tr>
<th>Collocation points</th>
<th>Second-order</th>
<th>Third-order</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ξ₁</td>
<td>ξ₂</td>
</tr>
<tr>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>0.00</td>
<td>0.77</td>
</tr>
<tr>
<td>3</td>
<td>0.00</td>
<td>-0.77</td>
</tr>
<tr>
<td>4</td>
<td>0.77</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>0.77</td>
<td>0.77</td>
</tr>
<tr>
<td>6</td>
<td>0.77</td>
<td>-0.77</td>
</tr>
<tr>
<td>7</td>
<td>-0.77</td>
<td>0.00</td>
</tr>
<tr>
<td>8</td>
<td>-0.77</td>
<td>0.77</td>
</tr>
<tr>
<td>9</td>
<td>-0.77</td>
<td>-0.77</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>-0.86</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>-0.86</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>-0.86</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>-0.86</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>-0.86</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>0.34</td>
</tr>
<tr>
<td>17</td>
<td></td>
<td>0.34</td>
</tr>
<tr>
<td>18</td>
<td></td>
<td>0.34</td>
</tr>
<tr>
<td>19</td>
<td></td>
<td>0.34</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>0.34</td>
</tr>
<tr>
<td>21</td>
<td></td>
<td>-0.34</td>
</tr>
<tr>
<td>22</td>
<td></td>
<td>-0.34</td>
</tr>
<tr>
<td>23</td>
<td></td>
<td>-0.34</td>
</tr>
<tr>
<td>24</td>
<td></td>
<td>-0.34</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>-0.34</td>
</tr>
</tbody>
</table>
Table 3 Parameters in Xinanjiang Model and their values used in this study.

<table>
<thead>
<tr>
<th>No.</th>
<th>Parameters</th>
<th>Definition</th>
<th>Value or Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WM</td>
<td>Areal mean tension water capacity (mm)</td>
<td>256.2825</td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>Ratio of WUM to WM (0 to 1)</td>
<td>0.8968</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WUM: Average basin storage capacity of the upper layer</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Y</td>
<td>Ratio of WLM to (1-X)*WM (0 to 1)</td>
<td>0.9999</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WLM: Average basin storage capacity of the lower layer</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(WM = WUM + WLM + WDM)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>WDM (Average basin storage capacity of the deeper layer)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>KE</td>
<td>Ratio of potential evaporation to pan evaporation (0 to 1)</td>
<td>0.8468</td>
</tr>
<tr>
<td>5</td>
<td>B</td>
<td>Exponential number of storage capacity distribution curve</td>
<td>0.4892</td>
</tr>
<tr>
<td>6</td>
<td>SM</td>
<td>Areal mean free water storage capacity (mm)</td>
<td>5 – 50</td>
</tr>
<tr>
<td>7</td>
<td>EX</td>
<td>A parameter in the distribution of free water storage capacity</td>
<td>0.500</td>
</tr>
<tr>
<td>8</td>
<td>CI</td>
<td>A coefficient relating RI, a contribution to interflow, to free water storage</td>
<td>0.1565</td>
</tr>
<tr>
<td>9</td>
<td>CG</td>
<td>A coefficient relating RG, a contribution to groundwater, to free water storage</td>
<td>0.0001</td>
</tr>
<tr>
<td>10</td>
<td>CIMP</td>
<td>Proportion of impermeable area to the total area</td>
<td>0.0012</td>
</tr>
<tr>
<td>11</td>
<td>C</td>
<td>Evapotranspiration coefficient from deep layer</td>
<td>0.3005</td>
</tr>
<tr>
<td>12</td>
<td>CKI</td>
<td>The interflow recession coefficient (0 to 1)</td>
<td>0.900-0.999</td>
</tr>
<tr>
<td>13</td>
<td>CKG</td>
<td>The groundwater recession coefficient (0 to 1)</td>
<td>0.950-0.999</td>
</tr>
<tr>
<td>14</td>
<td>CN</td>
<td>Number of cascade linear reservoir for runoff routing</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>CNK</td>
<td>Scale parameter of cascade linear (delta) reservoir</td>
<td>21.0271</td>
</tr>
</tbody>
</table>

Table 4 Comparison of statistics from PCE-2 and PCE-3 against those from MC.

<table>
<thead>
<tr>
<th>Index</th>
<th>Mean</th>
<th>STD</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCE-2</td>
<td>PCE-3</td>
<td>PCE-2</td>
<td>PCE-3</td>
</tr>
<tr>
<td>MUE</td>
<td>-0.28</td>
<td>-0.05</td>
<td>-1.03</td>
<td>-0.34</td>
</tr>
<tr>
<td>MUAE</td>
<td>0.66</td>
<td>0.46</td>
<td>1.33</td>
<td>0.54</td>
</tr>
</tbody>
</table>

Table 5 Statistics from PCE-2, PCE-3, and MC simulations at specific times.

<table>
<thead>
<tr>
<th>Day</th>
<th>Mean</th>
<th>STD</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MC</td>
<td>PCE-2</td>
<td>PCE-3</td>
<td>PCE-2</td>
</tr>
<tr>
<td>148</td>
<td>24.92</td>
<td>24.56</td>
<td>24.57</td>
<td>5.55</td>
</tr>
<tr>
<td>323</td>
<td>166.62</td>
<td>164.38</td>
<td>165.59</td>
<td>87.00</td>
</tr>
<tr>
<td>685</td>
<td>589.08</td>
<td>587.99</td>
<td>590.49</td>
<td>258.38</td>
</tr>
</tbody>
</table>
Figure 1 The study region (Tar river basin) in the eastern U.S.A.
Figure 2 The (a) mean and (b) standard deviation values of daily streamflow from PCE-2 and MC simulations and their differences for (c) mean and (d) standard deviation.
Figure 3 The scatter plots of the (a) mean and (b) standard deviation values of daily streamflow from PCE-2 and MC simulations.
Figure 4 The (a) mean and (b) standard deviation values of daily streamflow from PCE-3 and MC simulations and their differences for (c) mean and (d) standard deviation.
Figure 5 The scatter plots of the (a) mean and (b) standard deviation values of daily streamflow from PCE-3 and MC simulation.

Figure 6 Histograms generated through ensembles of MC (the left column), PCE-2 (the middle column), and PCE-3 (the right column) at the 148th day for low stream flow (a-c) indicated by red letter “A” in Fig.2 (b), the 323rd day for medium stream flow (d-f) indicated by red letter “B” in Fig.2 (b), and the 685th day for high stream flow (g-i) indicated by red letter “C” in Fig.2 (b).
Figure 7 Comparison of the ensembles generated from PCE-2 and PCE-3 against MC at specific times, (a) and (d) the 148th day, (b) and (e) the 323rd day, (c) and (f) the 685th day.
Figure 8 Daily streamflow simulation from Xinanjiang hydrological model and approximations from PCE-2 and PCE-3 using specified parameter values: (a) SM = 49.32, CKG = 0.95, CKI = 0.97, (b) SM = 24.02, CKG = 0.95, CKI = 0.97, and (c) SM = 13.01, CKG = 0.99, CKI = 0.91.