HYBRID APPROACHES FOR UNCERTAINTY QUANTIFICATION

DEVELOPMENT OF HYBRID APPROACHES FOR UNCERTAINTY QUANTIFICATION IN HYDROLOGICAL MODELING

By

Maysara Mostafa Ahmed Ghaith

BSc., MSc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy

Doctor of Philosophy (2020) (Civil Engineering)	McMaster University Hamilton, Ontario
TITLE:	Development of Hybrid Approaches for Uncertainty Quantification in Hydrological Modeling
AUTHOR:	Maysara Mostafa Ahmed Ghaith BSc., MSc. (Cairo University)
SUPERVISORS:	Dr. Zhong (Zoe) Li
NUMBER OF PAGES:	xviii, 132

Dedications

To Mostafa Ghaith & Mona Mannoon,

Mohra, Manar and Mayar,

Ahmed Ghaith and Mamdoha Mostafa

Lay Abstract

There is a water scarcity problem in the world, so it is vital to have reliable decision support tools for effective water resources management. Researchers and decision-makers rely on hydrological modelling to predict water availability. Hydrological model results are then used for water resources allocation and risk mitigation. Hydrological modelling is not a simple process, as there are different sources of uncertainty associated with it, such as model structure, model parameters, and data. In this study, data-driven techniques are used with process-driven models to develop hybrid uncertainty quantification approaches for hydrological modelling. The overall objectives are: *i*) to generate more robust probabilistic forecasts; *ii*) to improve the computational efficiency for uncertainty quantification without compromising accuracy; and, iii) to overcome the limitations of current uncertainty quantification methods, such as parameter interdependency. The developed hybrid approaches can be used by decision-makers in water resources management, as well as risk assessment and mitigation.

Abstract

Water is a scarce resource especially as the water demand is significantly increasing due to the rapid growth of population. Hydrological modelling has gained a lot of attention, as it is the key to predict water availability, optimize the use of water resources and develop risk mitigation schemes. There are still many challenges in hydrological modelling that researchers and designers are trying to solve. These challenges include, but not limited to: i) there is no single robust model that can perform well in all watersheds; ii) model parameters are often associated with uncertainty, which makes the results inconclusive; iii) the required computational power for uncertainty quantification increases with the increase in model complexity; iv) some modelling assumptions to simplify computational complexity, such as parameter independence are, are often not realistic. These challenges make it difficult to provide robust hydrological predictions and/or to quantify the uncertainties within hydrological models in an efficient and accurate way. This study aims to provide more robust hydrological predictions by developing a set of hybrid approaches. Firstly, a hybrid hydrological data-driven (HHDD) model based on the integration of a physically-based hydrological model (HYMOD) and a data-driven model (artificial neural network, ANN) is developed. The HHDD model is capable of improving prediction accuracy and generating interval flow prediction results. Secondly, a hybrid probabilistic forecasting approach is developed by linking the polynomial chaos expansion (PCE) method with ANN. The results indicate that PCE-ANN can be as reliable as but much more

efficient than the traditional Monte-Carlo (MC) method for probabilistic flow forecasting. Finally, a hybrid uncertainty quantification approach that can address parameter dependence is developed through the integration of principal component analysis (PCA) with PCE. The results from this dissertation research can provide valuable technical and decision support for hydrological modeling and water resources management under uncertainty.

Acknowledgment

I would like to express my sincere appreciation to my supervisor Dr. Zoe Li for her continuous help and guidance through my study since the first day. It was a great opportunity to work with such an amazing deliberate supervisor. I have learned a lot from her not only during supervision but also during the course I was teaching with her for three years. I would also like to thank my supervisory committee members Dr. Brian Baetz and Dr. Yipping Guo for their constructive feedback that helped me expand my work. Dr. Baetz was a great mentor for my health as well since the first year.

I am also very grateful to all my professors at Cairo University who helped me to become the person I am today. I have learned a lot from my home university as well as McMaster University. I want to give a special thanks to my friend, roommate and brother Ahmed Yosri and my supporter friend Ahmed El-Sayed as well as all the Egyptian friends that made me feel as if I am home and on top of them Dr. Shady Salem. I owe special thanks to Dr. Moataz Mohamed and Dr. Wael El-Dakhakhni.

I am also grateful to part of the civil engineering department with all the great fellows that made the working environment fruitful. I want to thank Joanne Gadawski for her continuous help and support in managing any academic problem. I can’t forget Sarah Sullivan, who made all her effort to make everything happen smoothly without any complications since day one.

Last but not least, I want to express my sincere gratitude to my family who supported me all the way during my study. Both my grandparents (Ahmed Ghaith and Mamdoha Mostafa) helped me to embrace myself since I was a kid to love science by helping in my studies. My parents (Mostafa Ghaith and Mona Mannoon) gave me all the support they have through all my life to be successful. No such words can be enough to thank them for what they have done. Also, this work could not be completed without the support from my sisters (Mohra, Manar, and Mayar).

Table of contents

List of Figures

List of Tables

Declaration of Academic Achievement

This dissertation was prepared per the guidelines set by the School of Graduate Studies at McMaster University. This is a sandwich thesis that contains a compilation of research papers published or prepared for publishing as journal articles. Chapters 2 and 3 have been published in peer-reviewed journals, and the research paper presented in Chapter 4 is under review. This dissertation presents the work carried out solely by Maysara Ghaith, where technical advice and guidance were provided for the whole thesis by the academic supervisor Dr. Zoe Li. Dr. Wendy Huang provided editorial comments for the paper presented in Chapters 3. Information from outside sources, which has been used towards analysis or discussion, has been cited where appropriate. The original contributions of the authors to each paper (Chapter) and the reasons for including them in this dissertation are outlined below:

Chapter 2: Ghaith, M, Siam, A., Li, Z., and El-Dakhakhni, W., 2020. "**Hybrid Hydrological Data-Driven Approach for Daily Streamflow Forecasting**". Journal of Hydrologic Engineering. 25, 1–9. With the permission from ASCE. This material may be downloaded for personal use only. Any other use requires prior permission of the American Society of Civil Engineers. The material may be found at https://doi.org/10.1061/(ASCE)HE.1943-5584.0001866.

The idea and concept for this paper came from Dr. Zoe Li with some modifications from Maysara Ghaith. Both Dr. Ahmad Siam and Dr. Wael El-

Dakhakhni were involved in the discussion of the research idea. Maysara Ghaith carried out the analysis as well as the preparation of the first draft. Dr. Ahmad Siam provided editorial comments to the earlier versions of this paper. The manuscript was reviewed and edited by Dr. Zoe Li and Dr. Wael El-Dakhakhni. This work should be included in this dissertation as it shows how to integrate a data-driven model with a physically-based model to obtain more robust hydrological forecasting results and to tackle model structure uncertainty.

Chapter 3: Ghaith, M., and Li, Z., 2020. "**Propagation of parameter uncertainty in SWAT: A probabilistic forecasting method based on polynomial chaos expansion and machine learning**". Journal of Hydrology. 586, 124854. https://doi.org/10.1016/j.jhydrol.2020.124854

Dr. Li came out with conceptualization of the idea for this paper, while Maysara Ghaith performed programming, analysis, and the interpretation of the results. Maysara Ghaith prepared the manuscript draft, and Dr. Zoe Li reviewed and revised it. This work should be included in this dissertation as it shows how polynomial chaos expansion (PCE) can be used as an efficient and accurate technique for the analysis of parameter uncertainty compared to the traditional Monte-Carlo simulation method. A data-driven model was introduced to the PCE framework to overcome a major limitation of PCE and provide probabilistic hydrological forecasts.

Chapter 4: Ghaith, M., Li, Z., and Baetz, B. "**Uncertainty analysis for hydrological models with interdependent parameters: an improved Polynomial Chaos Expansion approach.**" *Water Resources Research*. Submitted for publication in April 2020.

The idea for this paper was ignited from the second paper. Maysara Ghaith carried out the programming and analysis. Maysara Ghaith prepared the manuscript, and Dr. Zoe Li reviewed and edited it. Dr. Brian Baetz participated in the discussion of the research idea and provided editorial comments to the final version. This work should be included in this dissertation as it enables the use of PCE for the quantification of the uncertainty from correlated model parameters. The principle component analysis technique is integrated with PCE to address parameter dependence in hydrological models and thus to improve the accuracy and reliability of uncertainty analysis results.

Chapter 1

INTRODUCTION

1.1. BACKGROUND

Lately due to climate change, extreme events are becoming more frequent, which requires more preparedness to minimize risk and avoid catastrophic circumstances. Researchers, designers, and decision-makers in the hydrology area have been seeking to develop more robust daily (or sub-daily) probabilistic forecasting techniques. Hydrological modelling is essential for water resources allocation, flood risk management, water infrastructure operation and planning.

There are various types of hydrological models for flow forecasting. Hydrological models can be classified as physically-based or data-driven models. Physically-based models can be divided into three categories based on spatial variability: lumped, semi-distributed, and fully distributed models. The lumped model is also called a conceptual model as the catchment is considered as a single unit, where a parameter takes only one value for the whole watershed with no spatial variability within the watershed. The semi-distributed model divides the watershed into sub-basins, where each sub-basin has its own parameter values. The fully distributed model is the most complex as it divides the catchment into grids, and each grid has its properties and processes. Although it seems that fully distributed models are the best, as the model becomes complex, it requires more input data. With more parameters, it also brings more uncertainties to the model. On the other

hand, data-driven models are based on building a relationship between input and output variables without taking into consideration the hydrological processes. In data-driven models, sometimes the input data are tailored to have more or less some resemblance of the hydrological processes, but the model built is purely based on the relationship between input and output data.

1.2. PHYSICALLY-BASED MODELS

A physically-based model is a system of hydrological and/or hydraulic processes to simulate the catchment response to precipitation events. In physically-based models, there are some simplifications of the real-world hydrological processes. The degree of simplification depends on the type of the physically-based model. Lumped models are the most simplified, and fully-distributed models have the most details of the actual catchment response. In order to provide valid and reliable predictions, physically-based models need to be calibrated first.

Depending on the model type, the number of parameters to be calibrated can range from less than ten to hundreds. Sensitivity analysis can be carried out first to reduce the computational time and requirements. During the calibration process, non-sensitive parameters identified from the sensitivity analysis can be excluded. There are two different types of sensitivity analysis: local and global sensitivity analysis. Local sensitivity is done by fixing all parameters and changing only one parameter at a time around the value of interest to check how the output change accordingly. The analysis of local sensitivity is straightforward and fast, and it is

helpful when the computational resources are limited (Karkee and Steward, 2010; López-Cruz et al., 2012). Global sensitivity analysis takes account of output change with respect to the changes in all parameters within the entire parameter space and identifies the parameters that have the most significant influence on the model output (Dos Santos and Lu, 2015; López-Cruz et al., 2012; Scire et al., 2001). With the rapid development of high-performance computing technology, global sensitivity is becoming more common as it provides more realistic results than local sensitivity.

Previous studies have investigated how to improve physically-based models through different avenues. The first avenue is to improve the simulation of hydrological and hydraulic processes within the model or to develop new hydrological models (Ehteram et al., 2018; Farzin et al., 2018). The second avenue is to enhance the calibration algorithms to obtain more accurate parameter values that can reproduce the observed outflow (Singh et al., 2013; Yang et al., 2008). The third avenue is to investigate the trade-off between simplicity and accuracy (Herman et al., 2013; Vos et al., 2010). The goal is to find hydrological models with a structure as simple as lumped models but could provide results as accurate as fully-distributed models when the data is available.

Currently, lumped, and semi-distributed models are more widely used than fully-distributed models due to the limitation of data and the fact that the fully-distributed models require a lot of time to set up and calibrate. Both lumped and semi-distributed models could provide acceptable prediction with sufficient

watershed input data and the selection of an appropriate model. There is no universal model that performs well in all watersheds. The search for more robust hydrological models is still a very active research topic.

1.3. DATA-DRIVEN MODELS

Data-driven models are based on building relationships between input and output without explicit knowledge of the physical processes. In the past two decades, the development of technology has made it possible for researchers and decision-makers to collect more frequent and accurate data (Montáns et al., 2019). With the increases in data availability and computational capability, data-driven models have been widely investigated and used in hydrological modelling (Jothiprakash and Kote, 2011; Solomatine and Ostfeld, 2008; Taormina and Chau, 2015). Data-driven models are very dependent on the quality and quantity of the data. To use a data-driven model for forecasting, the model has to be trained on a certain percentage of the data then the rest is used for validation and testing. The most common percentage for training is 70%, while the rest 30% are divided equally for validation and testing (Wu and Chau, 2010; Zeroual et al., 2016). The percentage can vary depending on the length of available data. Input data is a crucial part of data-driven models. Depending on the data-driven model and the output required, the input variables to be included in the model have to be selected carefully.

In data-driven hydrological modelling, the selection of input variables is a challenging process. Including more input variables does not always lead to better model performance. Although there has been a lot of research on the input variable selection techniques for data-driven models, there is no universally accepted procedure for identifying input variables. Another common issue with data-driven models is the overfitting of models, as data-driven models capture the noises in the output, which will produce errors in forecasting results. The overfitting problem can happen because of adding more layers and elements in the model or adding correlated input variables. There is still no concrete solution to address this problem. By applying cross-validation or testing the model for unseen data, the model can be detected if it is overfitting or not. Moreover, if there are extreme events in the forecasting period that were not included during the training period, then the forecasting output might not be accurate (Amasyali and El-Gohary, 2018; Sudheer et al., 2002). Data-driven models rely on the data fed to it during the training model unlike the physically-driven model which rely on the hydrological and hydraulics process. Data-driven models are not extensively investigated in the hydrological flow forecast yet.

1.4. HYBRID MODELS

To overcome the drawbacks of both data-driven and physically-based models, the hybrid modelling approach has been recently investigated. The main idea of hybrid modelling is to have two or more layers of modelling techniques to

enhance the model performance. Hybrid modelling can be consisting of a physically-based model as the first layer, then a data-driven model as a second layer that uses the output from the physically-based model as its input (Humphrey et al., 2016; Mekonnen et al., 2015). Another way of hybrid modelling is to have two or more layers of the data-driven model in sequence (He et al., 2015; Tiwari and Chatterjee, 2011). Hybrid modelling has shown a lot of potential for further applications.

Even though the hybrid models can produce better results, there are still large uncertainties associated with the modeling process as a result of the different choices of member models. Also, by increasing the number of member models, the model complexity increases significantly, as each model has to be calibrated. Moreover, there might be errors propagating from one model to another. Hybrid modelling is promising; however, it requires more investigation and justification.

1.5. UNCERTAINTY ANALYSIS

In all hydrological models, physically-based, data-driven or hybrid, there are different types of uncertainties. First, model parameters are a major source of uncertainty. Even after the model calibration is performed, the parameter values might not be representing the actual values. Another source of uncertainty is associated with input data or observations. Moreover, the model structure is another source of uncertainty. Each hydrological model uses different mathematical representations of hydrological relationships; there could be different modeling

methods even for the same process in the same model (Talebizadeh et al., 2010; Tolessa et al., 2015). There have been many studies on the analysis and quantification of parameter uncertainty. To reduce the computational requirement, the analysis of parameter uncertainty is usually performed only on the most sensitive parameters identified from sensitivity analysis.

Common techniques for analyzing parameter uncertainty and generating probabilistic predictions include Monte Carlo (MC), generalized likelihood uncertainty estimation (GLUE), and bootstrap sampling (Li et al., 2009; Wu and Liu, 2012; Zhang et al., 2016). A major effort is being made to find uncertainty quantification algorithms that can reduce the required computational time and resources while maintaining the accuracy of the probabilistic prediction. More recently, Polynomial chaos expansion (PCE) showed some potential for quantifying parameter uncertainties in an effective and efficient manner (Fan et al., 2016, 2014; Wang et al., 2015). However, it has not been widely applied in hydrological modeling due to a few limitations. The existing PCE method relies on observations to quantify the propagation of parameter uncertainties with a model, which prevents it from providing hydrological forecasts under uncertainty. Moreover, model parameters have to be independent to be able to use PCE to quantify parameter uncertainties where most hydrological models have interdependency relationship between its parameters.

1.6. OBJECTIVES

Based on the gaps in previous studies, this research aims to develop a set of hybrid approaches to support the analysis and quantification of uncertainties in hydrological modelling. The main objectives of this research include the following:

1) To develop a hybrid hydrological model to generate more reliable predictions and address model structure uncertainty: the hybrid model will leverage the advantages of both physically-based and data-driven models by integrating a physically-based modelling layer with a second data-driven modeling layer.

2) To improve existing methods for addressing parameter uncertainty: the improved methods will be able to provide reliable probabilistic forecasts and support effective and robust water resources planning and management.

The developed approaches will provide results that can be used in scenario-based optimization models. The results of the optimization models can be used for planning to accommodate for any future risk or generate operation rules based on future scenarios.

1.7. DISSERTATION ORGANIZATION

The five chapters of the dissertation can be summarized as follows:

In Chapter 1 provides the background required for this research, brief literature review, an overview of the objectives, and a description of the dissertation organization

In Chapter 2, a hybrid model that consists of two layers is developed to generate a more robust prediction and address the model structure uncertainty. The first layer is a lumped model named HYMOD, and the second layer is a data-driven model, i.e., artificial neural network (ANN). Results from HYMOD and the hybrid model are compared and discussed to demonstrate the advantages of the developed hybrid model.

In Chapter 3, An innovative uncertainty analysis method based on the integration of PCE and ANN is developed. The introduction of ANN enables the developed PCE-ANN method to generate probabilistic forecasts, which cannot be done by the traditional PCE approach. The parameter uncertainty in Soil & Water Assessment Tool (SWAT) is analyzed using MC simulation, the traditional PCE method and the PCE-ANN method. The results are compared to show the advantages of the developed hybrid method.

In Chapter 4, the traditional PCE is further improved to address parameter dependency during the analysis of parameter uncertainty. Principle component analysis (PCA) is integrated with PCE to enable PCE to quantify the uncertainties of interdependent model parameters. The new PCA-PCE method is applied to SWAT to demonstrate its applicability.

Finally, Chapter 5 presents the conclusions of this dissertation research and recommendations for future work.

1.8. REFERENCES

Amasyali, K., El-Gohary, N.M., 2018. A review of data-driven building energy consumption prediction studies. Renew. Sustain. Energy Rev. 81, 1192–1205. https://doi.org/10.1016/j.rser.2017.04.095

Dos Santos, R.F., Lu, C.-T., 2015. Geography Markup Language (GML). Springer International Publishing, pp. 1–6. https://doi.org/10.1007/978-3-319-23519-6_480-2

Ehteram, M., Othman, F.B., Yaseen, Z.M., Afan, H.A., Allawi, M.F., Malek, M.B.A., Ahmed, A.N., Shahid, S., Singh, V.P., El-Shafie, A., 2018. Improving the Muskingum flood routing method using a hybrid of particle swarm optimization and bat algorithm. Water (Switzerland) 10, 1–21. https://doi.org/10.3390/w10060807

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

Fan, Y.R., Huang, W., Huang, G.H., Huang, K., Zhou, X., 2014. A PCM-based stochastic hydrological model for uncertainty quantification in watershed systems. Stoch. Environ. Res. Risk Assess. 29, 915–927. https://doi.org/10.1007/s00477-014-0954-8

Farzin, S., Singh, V.P., Karami, H., Farahani, N., Ehteram, M., Kisi, O., Allawi, M.F., Mohd, N.S., El-Shafie, A., 2018. Flood routing in river reaches using a three-parameter Muskingum model coupled with an improved Bat algorithm. Water (Switzerland) 10. https://doi.org/10.3390/w10091130

He, X., Guan, H., Qin, J., 2015. A hybrid wavelet neural network model with mutual information and particle swarm optimization for forecasting monthly rainfall. J. Hydrol. 527, 88–100. https://doi.org/10.1016/j.jhydrol.2015.04.047

Herman, J.D., Reed, P.M., Wagener, T., 2013. Time-varying sensitivity analysis clarifies the effects of watershed model formulation on model behavior. Water Resour. Res. 49, 1400–1414. https://doi.org/10.1002/wrcr.20124

Humphrey, G.B., Gibbs, M.S., Dandy, G.C., Maier, H.R., 2016. A hybrid approach to monthly streamflow forecasting : Integrating hydrological model outputs into a Bayesian artificial neural network. J. Hydrol. 540, 623–640. https://doi.org/10.1016/j.jhydrol.2016.06.026

Jothiprakash, V., Kote, A.S., 2011. Improving the performance of data-driven techniques through data pre-processing for modelling daily reservoir inflow. Hydrol. Sci. J. 56, 168–186. https://doi.org/10.1080/02626667.2010.546358

Karkee, M., Steward, B.L., 2010. Local and global sensitivity analysis of a tractor and single axle grain cart dynamic system model. Biosyst. Eng. 106, 352–366. https://doi.org/10.1016/j.biosystemseng.2010.04.006

Li, Z., Xu, Z., Shao, Q., Yang, J., 2009. Parameter estimation and uncertainty analysis of SWAT model in upper reaches of the Heihe river basin. Hydrol. Process. 23, 2744–2753. https://doi.org/10.1002/hyp.7371

López-Cruz, I.L., Rojano-Aguilar, A., Salazar-Moreno, R., Ruiz-García, A., Goddard, J., 2012. A comparison of local and global sensitivity analyses for greenhouse crop models. Acta Hortic. 957, 267–273. https://doi.org/10.17660/ActaHortic.2012.957.30

Mekonnen, B.A., Nazemi, A., Mazurek, K.A., Elshorbagy, A., Putz, G., 2015. Approche par modélisation hybride de l'hydrologie des Prairies : fusion de modèles hydrologiques dirigés par les données et basés sur les processus. Hydrol. Sci. J. 60, 1473–1489. https://doi.org/10.1080/02626667.2014.935778

Montáns, F.J., Chinesta, F., Gómez-Bombarelli, R., Kutz, J.N., 2019. Data-driven modeling and learning in science and engineering. Comptes Rendus - Mec. 347, 845–855. https://doi.org/10.1016/j.crme.2019.11.009

Scire, J.J., Dryer, F.L., Yetter, R.A., 2001. Comparison of global and local sensitivity techniques for rate constants determined using complex reaction mechanisms. Int. J. Chem. Kinet. 33, 784–802. https://doi.org/10.1002/kin.10001

Singh, V., Bankar, N., Salunkhe, S.S., Bera, A.K., Sharma, J.R., 2013. Hydrological stream flow modelling on Tungabhadra catchment :

parameterization and uncertainty analysis using SWAT CUP. Curr. Sci. 104, 1187–1199.

Solomatine, D.P., Ostfeld, A., 2008. Data-driven modelling : some past experiences and new approaches. J. Hydroinformatics 10, 3–22. https://doi.org/10.2166/hydro.2008.015

Sudheer, K.P., Gosain, A.K., Ramasastri, K.S., 2002. A data-driven algorithm for constructing artificial neural network rainfall-runoff models. Hydrol. Process. 16, 1325–1330. https://doi.org/10.1002/hyp.554

Talebizadeh, M., Morid, S., Ayyoubzadeh, S.A., Ghasemzadeh, M., 2010. Uncertainty analysis in sediment load modeling using ANN and SWAT model. Water Resour. Manag. 24, 1747–1761. https://doi.org/10.1007/s11269-009-9522-2

Taormina, R., Chau, K.W., 2015. Data-driven input variable selection for rainfall-runoff modeling using binary-coded particle swarm optimization and Extreme Learning Machines. J. Hydrol. 529, 1617–1632. https://doi.org/10.1016/j.jhydrol.2015.08.022

Tiwari, M.K., Chatterjee, C., 2011. A new wavelet–bootstrap–ANN hybrid model for daily discharge forecasting. J. Hydroinformatics 13, 500–519. https://doi.org/10.2166/hydro.2010.142

Tolessa, O., Nossent, J., Velez, C., Kumar, N., Griensven, A. Van, Bauwens, W.,

2015. Environmental Modelling & Software Assessment of the different sources of uncertainty in a SWAT model of the River Senne (Belgium). Environ. Model. Softw. 68, 129–146. https://doi.org/10.1016/j.envsoft.2015.02.010

Vos, N.J. De, Rientjes, T.H.M., Gupta, H. V, 2010. Diagnostic evaluation of conceptual rainfall – runoff models. Hydrol. Process. 2850, 2840–2850. https://doi.org/10.1002/hyp.7698

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. https://doi.org/10.1016/j.jhydrol.2015.10.021

Wu, C.L., Chau, K.W., 2010. Data-driven models for monthly streamflow time series prediction. Eng. Appl. Artif. Intell. 23, 1350–1367. https://doi.org/10.1016/j.engappai.2010.04.003

Wu, Y., Liu, S., 2012. Environmental Modelling & Software Automating calibration , sensitivity and uncertainty analysis of complex models using the R package Flexible Modeling Environment (FME): SWAT as an example. Environ. Model. Softw. 31, 99–109. https://doi.org/10.1016/j.envsoft.2011.11.013

Yang, J., Reichert, P., Abbaspour, K.C., Xia, J., Yang, H., 2008. Comparing uncertainty analysis techniques for a SWAT application to the Chaohe Basin

in China. J. Hydrol. 358, 1–23. https://doi.org/10.1016/j.jhydrol.2008.05.012

Zeroual, A., Meddi, M., Assani, A.A., 2016. Artificial Neural Network Rainfall-Discharge Model Assessment Under Rating Curve Uncertainty and Monthly Discharge Volume Predictions. Water Resour. Manag. 3191–3205. https://doi.org/10.1007/s11269-016-1340-8

Zhang, D., Chen, X., Yao, H., James, A., 2016. Moving SWAT model calibration and uncertainty analysis to an enterprise Hadoop-based cloud. Environ. Model. Softw. 84, 140–148. https://doi.org/10.1016/j.envsoft.2016.06.024

Chapter 2

HYBRID HYDROLOGICAL DATA-DRIVEN APPROACH FOR DAILY STREAMFLOW FORECASTING

ABSTRACT

Hydrological forecasting is key for water resources allocation and flood risk management. Although a number of advanced hydrological forecasting methods have been developed in the past, daily (or sub-daily) forecasting remains a major challenge in engineering hydrology. The uncertainties associated with input data, model parameters, and model structure necessitate developing more robust modeling techniques. In this study, a hybrid machine-learning approach based on hydrological and data-driven modeling is developed for daily stream-flow forecasting. The proposed hybrid hydrological data-driven model (HHDD) approach succeeds in improving daily prediction compared to that predicted by the standard conceptual hydrological model (HYMOD). In addition, the developed HHDD model is more robust in terms of providing direct uncertainty analysis results. The results indicate that a better resemblance of streamflow pattern is achieved by integrating physically based and data-driven approaches into the developed HHDD model. DOI: 10.1061/(ASCE)HE.1943-5584.0001866. © 2019 American Society of Civil Engineers.

Keywords: Daily streamflow forecasting; Data-driven modeling; Hybrid modeling; Hydrological modeling; HYMOD; Uncertainty analysis.

2.1. INTRODUCTION

The magnitude and frequency of natural disasters attributed to extreme metrological and hydrological hazards have been increasing in North America and in many regions all over the world due to climate change. Therefore, water resources planning, and flood risk forecasting and mitigation are priority research areas (Barati et al. 2012). In order to improve the resilience of water resources systems, accurate estimation of daily streamflow is key (Adams et al. 2018; Bagatur and Onen 2018). In this regard, several studies have been conducted to improve streamflow prediction models.

Previously, various methods have been proposed to improve the performance of simulation models through adjusting model parameters (Chen et al. 2018; Zhang et al. 2017; Barati 2011), addressing the propagation of parameter uncertainties (Fan et al. 2015, 2016; Zhang et al. 2016; Zheng and Han 2016), and improving the representation of hydrological and hydraulics processes (Ehteram et al. 2018; Farzin et al. 2018; Fu et al. 2014; Wi et al. 2015). In parallel, other researchers attempted to augment model spatial resolution by shifting from lumped to semidistributed rainfall-runoff models or from semidistributed to fully distributed models (Mendonça et al. 2018; Singh and Marcy 2017; Wi et al. 2015), which enriches the representation of hydrological processes. Despite the previous research, there are still major challenges in hydrological simulations, including model calibration and uncertainty quantification. Particularly, for fully distributed models,

more data-collection efforts are necessary for the model setup. These drawbacks highlight the need for developing robust data-driven hydrological modeling and streamflow forecasting techniques.

In the past decade, the increase of computational power and data availability have made the development of data-driven models more appealing (Bertone et al. 2017; Dariane et al. 2018; Kothari and Gharde 2015; Nanda et al. 2016; Zeroual et al. 2016). Recent research has found promising results using data-driven techniques such as artificial neural network (ANN) (Khan et al. 2016; Wang et al. 2015), fuzzy logic (FL) (Chen et al. 2013; Özger 2009; Özger et al. 2012; Wang and Altunkaynak 2012), and support vector ma- chine (SVM) (Ch et al. 2013; Sudheer et al. 2014; Wu et al. 2014) for streamflow forecasting. However, such data-driven models are heavily influenced by data availability, data pretreatment, and selection of input variables (Feng et al. 2017; Galelli et al. 2014). Another common drawback of data-driven models is overfitting, which essentially means that noise within the data could negatively impact the models' predictive performance when handling new data due to the lack of understanding of the physical hydrological processes.

In order to overcome the drawbacks of the aforementioned modeling techniques, hybrid modeling has been recently introduced and implemented (Nourani et al. 2014). Hybrid models integrate both process-driven and data-driven models in order to enhance the overall model performance (Nourani et al. 2014). For example, Humphrey et al. (2016) developed a hybrid model based on the conceptual hydrological model Génie Rural à 4 paramètres Journalier (Agricultural

Engineering Model 4 Parameters Daily) (GR4J) and ANN to improve monthly streamflow prediction. Song et al. (2012) developed a hybrid model based on the semidistributed model Xinanjiang (XAJ). They took the output of each subcatchment as an input to the ANN to train it. The results were promising for event-based simulation. The results of previous studies demonstrated the potential of the hybrid modeling approach. However, such studies mainly focused on either monthly streamflow forecasting (Nourani et al. 2014; Humphrey et al. 2016) or hourly event-based simulation (Song et al. 2012).

Daily streamflow forecasting, which could significantly improve flood and drought management, is still very challenging due to the prediction complexity originating from the fluctuations in daily measurements. As such, the current study focuses on developing a hybrid model for daily streamflow forecasting through integrating a lumped physically based rainfall-runoff simulation model with a data-driven model. The hybrid hydrological data-driven (HHDD) model developed in this study integrates hydrological model (HYMOD) as the model's first layer and ANN as a second layer. In this respect, HYMOD is first used to obtain simulated flow based on watershed characteristics. Subsequently, the HYMOD's output is used as input for ANN for final daily flow forecasting. The objective of this study is to develop a robust and reliable daily streamflow forecasting method based on HHDD modeling. In order to demonstrate its applicability, the developed HHDD model is applied to a case study in the Guadalupe River Watershed in Texas.

2.2. METHODOLOGY

2.2.1. HYMOD

HYMOD is a lumped rainfall-runoff model based on conceptually simplified physical processes (Moore 1985; Quan et al. 2015). The modeling process can be divided into three steps. First, the excess infiltration method is used to calculate the amount of infiltration and runoff produced. Runoff is predicted as the excess water after evapotranspiration and infiltration are subtracted. In this step, evapotranspiration is an input variable, whereas the infiltration is calculated based on the soil infiltration capacity, which is determined by two parameters, C_{max} and B_{exp}. C_{max} is the maximum storage capacity, and B_{exp} is used to address the spatial distribution of water storage. Second, the runoff (i.e., excess water) is divided into surface runoff and base flow using an α coefficient. Three consecutive, identical quick reservoirs with a travel time of R_q are used to calculate surface runoff, and the base flow is calculated using a slow reservoir with a travel time of R_s. Finally, the discharge is calculated as the summation of both discharges from the quick and the slow reservoirs. The schematic in Figure 2-1 summarizes the modeling process.

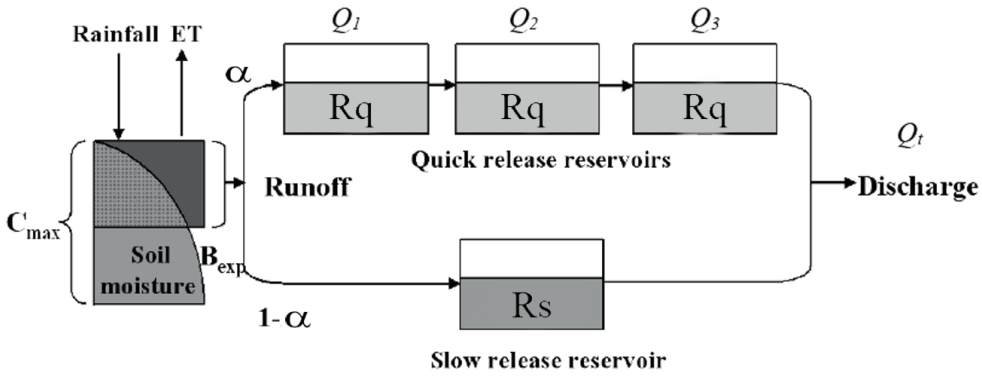

Figure 2-1: Schematic of the HYMOD modeling process.

The HYMOD model has five parameters as mentioned previously. The first three parameters (C_{max}, B_{exp}, and α) are used to calculate the generated runoff, whereas the other two parameters (R_q and R_s) are used in the routing process to estimate the discharge at the catchment outlet. Because this is a lumped rainfall-runoff model, the five parameters are not explicitly measured in the field per se; instead, their ranges have been defined in previous studies (Quan et al. 2015; Vrugt et al. 2008) as given in Table 2-1. In this study, the parameters' distribution is assumed to be uniform (Quan et al. 2015; Vrugt et al. 2008).

Table 2-1: List of HYMOD Parameters.

	Min	Max	Units	Description
C_{max}	1	500	mm	Maximum storage capacity
B_{exp}	0	15	-	Degree of spatial variability in the soil capacity
A	0.01	0.99	-	Factor of distribution of water to surface and base flow
R_s	0.01	0.99	day	Travel time of slow tank
R_q	0.01	1.2	day	Travel time of quick reservoirs (all three are identical)

.

2.2.2. ARTIFICIAL NEURAL NETWORK

ANNs are data-driven models that mimic the structure of the human brain to facilitate data mining for prediction and/or classification (Khan et al. 2016). An ANN can be divided into several layers:

• Input data are the first layer, which is connected to the hidden layer(s) by a number of neurons (Kothari and Gharde 2015).

• There can be one or more hidden layers depending on the depth of data mining. The number of hidden layers and the neuron weights can be determined by training the model with input and output data.

• The final layer is the sought results, which can be a single variable or multiple variables.

The performance of an ANN model depends on the quantity and quality of data, as well as the training processes. There is no universally accepted rule for determining the optimal number of input variables, neurons, or hidden layers; however, it has been shown that data pretreatment can effectively increase the performance of ANNs (Feng et al. 2017; Humphrey et al. 2016; Nanda et al. 2016).

2.2.3. HYBRID MODELING

As explained previously, hybrid modeling integrates two or more modeling techniques in order to improve the model performance. In this study, the HYMOD model (a process-driven model) is integrated with the ANN model (a data-driven

model) to develop the HHDD model. HYMOD can reflect the underlying physical processes for precipitation-runoff simulation, and ANN can better capture the nonlinear relationship between hydrological parameters and streamflow output. The developed HHDD model has only a few parameters to calibrate and requires less data collection effort, and thus is able to overcome the aforementioned disadvantages of both models.

In this study, four HHDD models are built and analyzed. All models take the data from the best models obtained from the calibration and testing periods to train the ANN model (1984–1993). The first model uses HYMOD inputs (rainfall, temperature, and potential evapotranspiration) and output time series (runoff, surface water from each quick reservoir, flow from slow reservoir, and total outflow) to train the ANN. The total outflow predicted by HYMOD, rather than observed flow, is used for the HHDD model because the HHDD model is intended to be used for future prediction when observation flow is unavailable. In the second model, cumulative precipitation is added as additional input variables of the first model, aiming to improve prediction accuracy. The third model is developed by introducing lagged time series of the HYMOD out- put to the first model in order to take into consideration the effects of mismatching error. The fourth model includes all the variables from the second and third model. The selected variables of the four HHDD models are summarized in Table 2-2.

Table 2-2: HHDD model input variables.

Model Name	HyMod In-Out variables	Cumulative Precipitation*	Previous flow lags**	Number of parameters
Model 1	X			7
Model 2	X	X		16
Model 3	X		x	14
Model 4	X	X	x	23

*Cumulative Precipitation of 2, 3, 7, 10, 15, 20, 30, 45, and 60 days

**Previous flow lags of 1, 2, 3, 4, 5, 6, and 7 days

The HYMOD modeling process is presented in Figure 2-2. To build and calibrate the HYMOD model, HYMOD is run several times using the same rainfall and weather data but with different parameter values each time during the calibration period. The model performance is evaluated using the Nash-Sutcliffe model efficiency coefficient (NSE) and the coefficient of determination (R^2). During each iteration, the parameters' values are selected from the range given in Table 1, and the output is compared with the observation. The NSE is evaluated and compared with the threshold of 0.6. If the NSE value satisfies the threshold, then the model is evaluated for the testing period. The NSE of the model for the testing period has to exceed 0.5 to accept the set of parameters and be used for forecasting during the validation period.

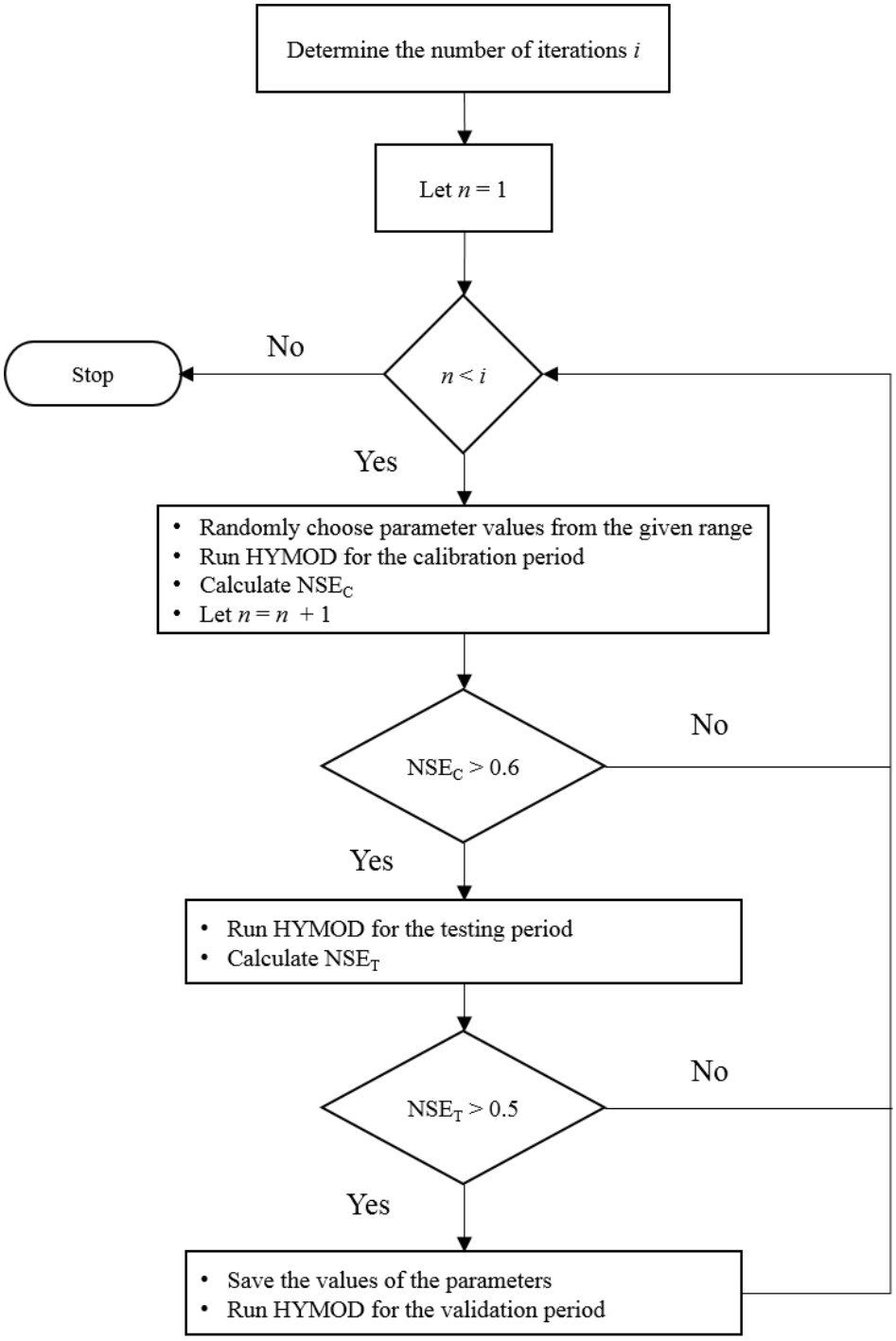

Figure 2-2: Flowchart of the HYMOD calibration, testing, and validation processes.

The flowchart of the HHDD modeling process is shown in Figure 2-3. In the proposed hybrid approach, HYMOD is run first using randomly selected parameter values for the calibration period. The input and output time series from HYMOD are used as the input of ANN. Only one hidden layer is used for the ANN to keep the hybrid model in its simplest form. For some HHDD models, other input variables are included as given in Table 2-2. The number of ANN neurons is calibrated through an iterative process by searching for the highest R^2 and lowest mean square error (MSE). In this study, the HHDD is trained for 250 realizations using the best number of neurons and the observation data during the calibration period. Each of the trained models is run for predicting the streamflow for the testing and validation periods. The time series of the streamflow is averaged from all the 250 iterations to avoid overfitting.

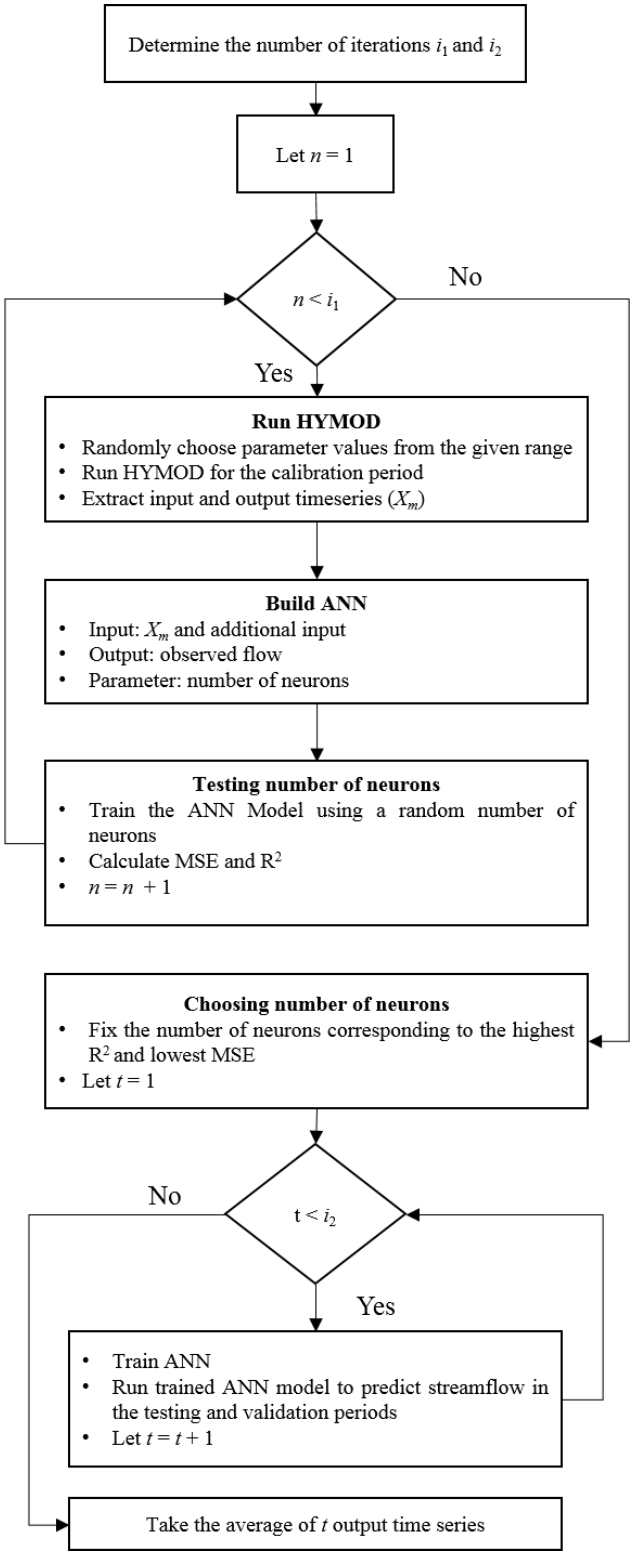

Figure 2-3: Flowchart of the HHDD modeling process.

2.3. STUDY AREA AND DATA COLLECTION

A case study of the Guadalupe Basin in Texas is used to demonstrate the applicability of the proposed HHDD approach. The Guadalupe Basin is located in the southeast part of Texas and discharges to the Gulf of Mexico, as shown in Figure 2-4. The basin has several subcatchments with a few flow gauges and weather stations. Due to limited data availability, only the upper subcatchment of the Spring Branch Basin was studied. The Spring Branch has only one mainstream, which verifies that using a lumped rainfall-runoff model, which does not consider the spatial distribution of subrivers or subcatchment, is reasonable. The total area of the Spring Branch catchment is approximately 3,500 km^2, which is considered a medium to large basin. The difference in elevation is 345 m over a stream length of 290 km with a mild average slope of approximate 0.12%.

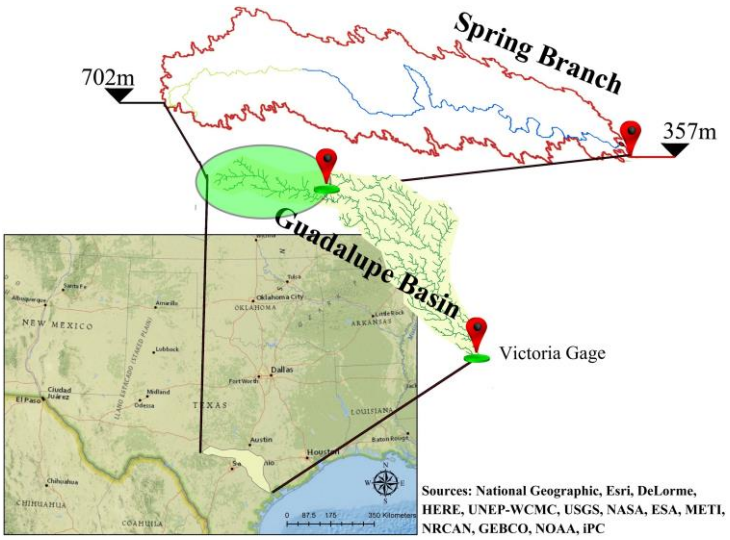

Figure 2-4: Study area in the Guadalupe Basin, Texas. (Base map from National Geographic, Esri, DeLorme, HERE, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA, iPC.).

Fourteen years of rainfall and temperature data for the Spring Branch Basin were collected. The weather data were obtained from the National Oceanic and Atmospheric Administration (NOAA) at Victoria rain gauge (USW00012912) and the flow data were obtained from the USGS at Spring Branch flow gauge. The data cover a period from 1984 to 1997 with a daily resolution and are divided into three parts: (1) the first 7 years (1984–1990) for calibration to choose best model parameters from HYMOD; (2) 3 years (1991– 1993) for testing to assess if the chosen HYMOD models are good enough; and (3) the last 4 years (1994–1997), which are used for validation to compare the performance of the HYMOD and the HHDD models.

The basin lies in a moderate temperature zone with an average maximum temperature of 27°C and an average minimum temperature of 16°C. The basin has a low annual rainfall with an average of 2.3 mm/day, which leads to a relatively low mean annual streamflow of 16.7 m^3/s; however, floods occur occasionally during spring and summer. There is no obvious annual or seasonal pattern of peak flow (Figure 2-5), and the magnitude of peak flow seems to be increasing, which makes it challenging to forecast streamflow.

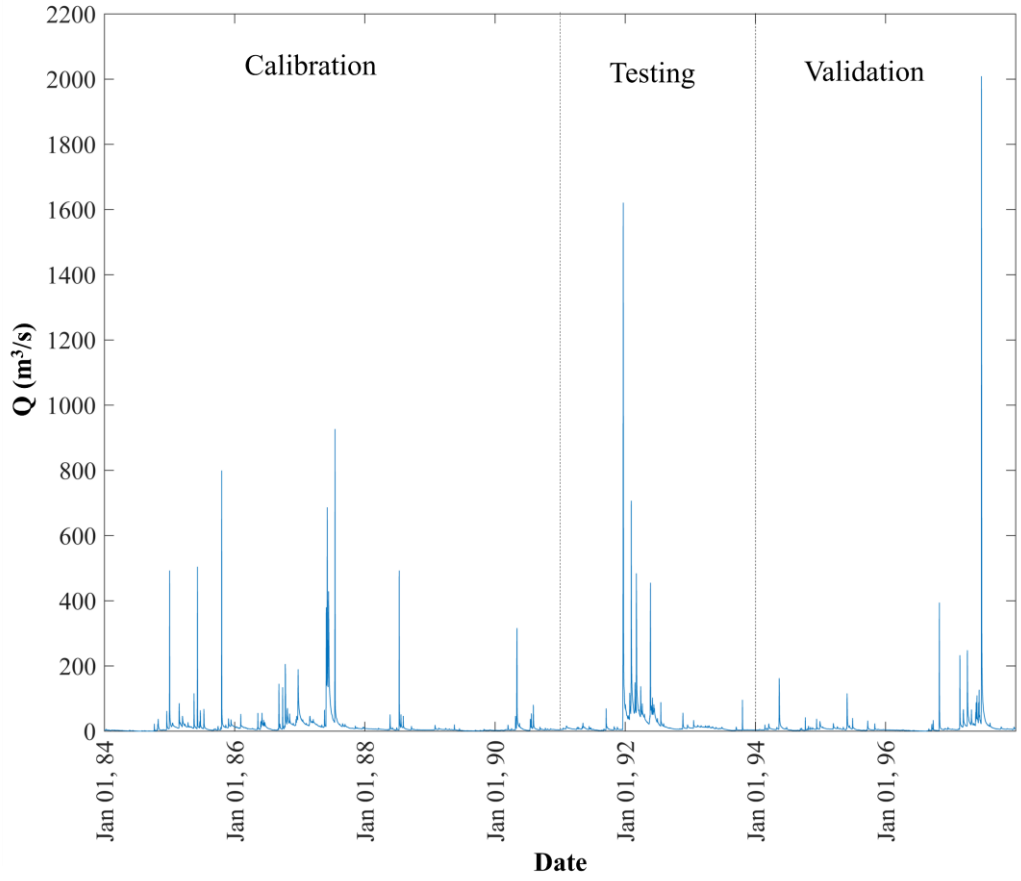

Figure 2-5: Time series of daily flow at the Spring Branch gauge.

2.4. RESULTS AND DISCUSSION

2.4.1. PERFORMANCE OF THE HHDD MODELS

According to Table 2-3, HHDD Models 1 and 4 both show better performance than the calibrated HYMOD model during the testing period, and Model 2 has the best performance during the validation period. Although there are some uncertainties about which model can provide the best performance, all four HHDD models perform better than the HYMOD during the validation period. Compared

with previous studies, the results of the Models 2 and 4 are considered very good, whereas HYMOD and Models 1 and 3 are considered only acceptable or good based on NSE (Jimeno-Sáez et al. 2018). For example, Jimeno-Sáez et al. (2018) compared the daily forecasting results using both the Soil and Water Assessment Tool (SWAT) and ANN for two different basins (Min˜ o-Sil and Segura River) in peninsular Spain. The NSE values for the two basins using SWAT were 0.57 and 0.48, respectively, and the R^2 values were 0.58 and 0.61, respectively. When ANN was used, the NSE values were 0.59 and 0.49, and the R^2 values were 0.61 and 0.52, respectively. The results indicate that the HHDD models developed in this study have better performance. Meanwhile, other daily forecasting–focused studies used an NSE value of 0.7 as an acceptable threshold for good models (Li et al. 2010; Yang et al. 2007, 2008; Zhang et al. 2016). Based on the mentioned studies, the HHDD models have acceptable or good results and have made a significant improvement to the traditional modeling approach.

Table 2-3: Performance of HYMOD and HHDD models.

	Testing		Validation	
	NSE	R^2	NSE	R^2
HYMOD	0.69	0.7	0.54	0.67
HHDD 1	0.73	0.74	0.61	0.68
HHDD 2	0.67	0.68	0.74	0.79
HHDD 3	0.65	0.74	0.64	0.71
HHDD 4	0.74	0.74	0.7	0.75

Note: NSE = Nash–Sutcliffe model efficiency coefficient, and R^2 = coefficient of determination.

2.4.2. SELECTION OF THE BEST HHDD MODEL

In order to choose the best of the four HHDD models, further investigation has been carried out. To mitigate the impacts of uncertain HYMOD parameters, the set of HYMOD parameters that lead to the highest NSE values in both calibration and testing periods was selected. Instead of choosing only one set of the output of HYMOD, an uncertainty analysis was carried out to check the robustness of these models against the change of the HYMOD parameters.

To conduct the uncertainty analysis, a Monte Carlo simulation was carried out by assuming a uniform distribution for each of the five HYMOD parameters. The ranges of the uniform distributions are given in Table 2-1. A total of 10,000 simulations were conducted, and the ones with an NSE value higher than 0.6 during the calibration period and NSE value higher than 0.5 during the testing period were selected for further analysis.

According to the aforementioned criteria, a total of 58 parameter sets were chosen for the development of the HHDD models. The results of HYMOD and the four HHDD models at Spring Branch are compared with the observations, and NSE values are plotted as box plots in Figure 2-6. Both Models 2 and 4 have high NSE values, which are absolutely higher than those of the original HYMOD model. The variation of Model 4's NSE values are smaller than that of Model 2. However, Model 2's performance, in terms of NSE, is better, and it also requires fewer input variables, which makes Model 2 more preferable. Another advantage of Model 2

compared with Model 4 is that Model 2 does not use previous flow calculated in HYMOD as Model 4 does; thus, the cascade error propagating from HYMOD to ANN can be avoid.

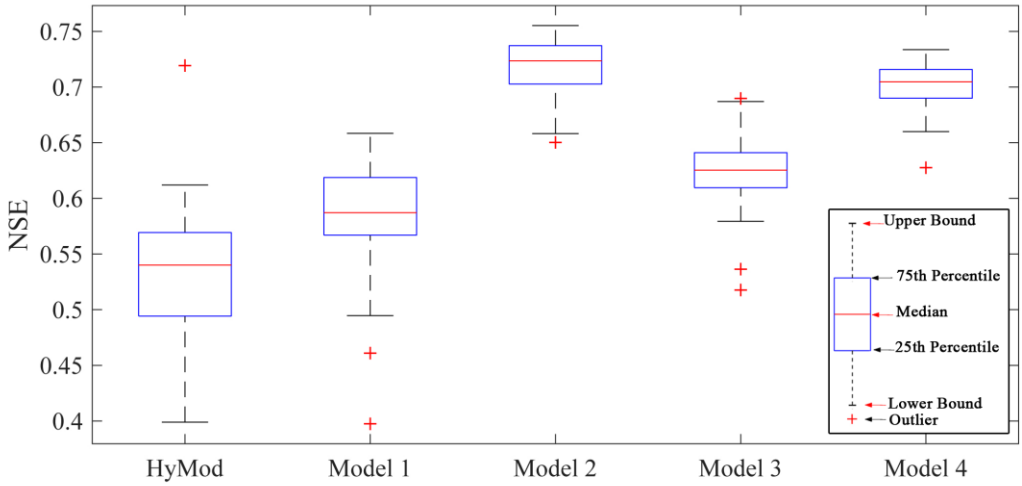

Figure 2-6: Nash-Sutcliffe efficiency results for Monte-Carlo analysis during the validation period.

2.4.3. ADVANTAGES OF THE HHDD MODELING APPROACH

Based on the previous analysis, the HHDD model selected (i.e., Model 2) has an overall better performance than the HYMOD model. Model 2 can also generate satisfactory forecasts even without prior knowledge of the best combination of HYMOD parameters. Figure 2-7 shows the results of HYMOD and HHDD, as well as the observed flow, during the validation period. The HYMOD results are generated from the best HYMOD model with the highest NSE during both the calibration and testing periods, and the averaged HHDD results are from Model 2. As shown in Figure 2-7(a), the HHDD peaks concur with the observations, whereas HYMOD does not accurately predict the time of occurrence of streamflow peaks.

This could be attributed to the fact that the values of the HYMOD parameters are kept constant, which is not realistic (Herman et al. 2013; De Vos et al. 2010). For example, because flow velocity is a function of the quantity, R_s and R_q should be variables dependent on rainfall (De Vos et al. 2010). Unlike HYMOD, the HHDD model implicitly creates new variables that change respectively as flow change because the ANNs generate relationships between the HYMOD variables and cumulative precipitation.

The performance of HHDD and HYMOD was further analyzed using two statistic criteria: the root-mean square error (RMSE) and the mean absolute error (MAE). The results show that the overall performance of HHDD is better than HYMOD in terms of modeling errors during the validation period. The RMSE values of HHDD and HYMOD are 34 and 44 m^3/s, respectively, and the MAE values are 7.5 and 10 m^3/s, respectively. As shown in Figure 2-7(b), although the HHDD model does not accurately predict the streamflow peak on June 22, 1997, it provides a better estimate than the HYMOD. The HHDD model was incapable of predicting the exact value because this value is higher than any of the training data.

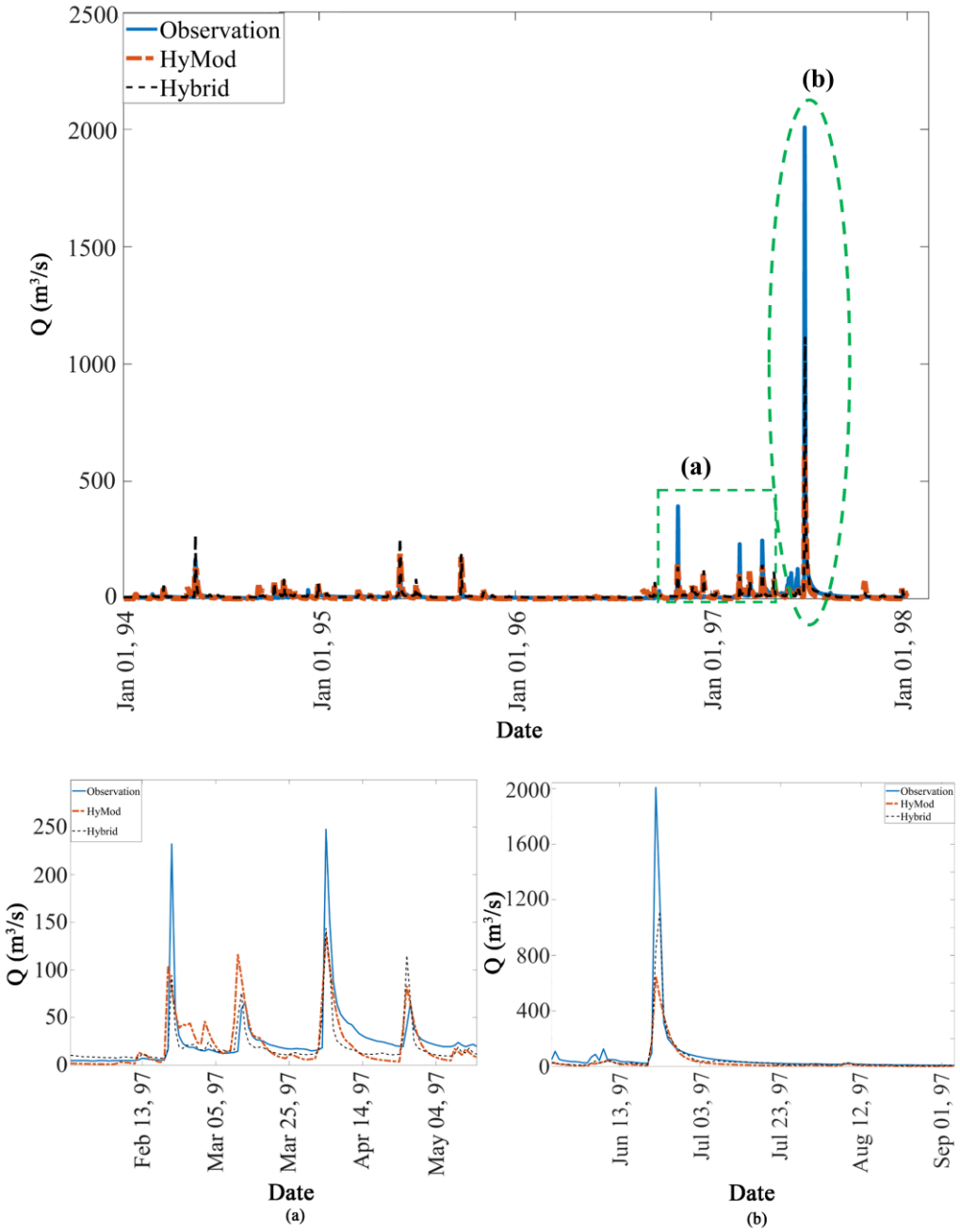

Figure 2-7: Time series results of the best HYMOD model and the average of HHDD Model 2.

The HHDD model performed better in simulating peak flow events. For example, the total observed volume during June 1997 is 419×10^6 m^3. The HHDD model estimated the volume to be 319×10^6 m^3, which is more accurate than the HYMOD value of 230×10^6 m^3. It is worth mentioning that peak flow estimation is one of the most challenging tasks in hydrological modeling. Although HHDD could provide more accurate peak flow prediction in comparison with HYMOD, the HHDD model can be further improved. For example, improving data quality and including more peak flow events when training the model could be helpful. Developing separate models for dry and wet seasons might also help better capture the peak flow patterns (Jothiprakash and Kote 2011). Moreover, wavelet transformation can be used to separate high flow and low flow before modeling to enhance the model performance in capturing the peak flows (Pramanik et al. 2011; Tiwari and Chatterjee 2011). Overall, the results demonstrated that the HHDD model has the ability to generate more accurate flow than that obtained from HYMOD. This is because the HHDD model takes into consideration the cumulative precipitation when training the model. Thus, the HYMOD parameters are translated into new time-dependent variables that is a function of the flow as well. Based on these analyses, it can be inferred that the HHDD model performance is affected by both the physical model structure and the new data fed to the data-driven model.

Additionally, the HYMOD is a lumped rainfall-runoff model that does not consider spatial variability within the catchment. As such, only one rainfall time

series (either rainfall measured at one rain gauge or the average rainfall from multiple rain gauges) can be used as HYMOD model input. Although not demonstrated in this study, spatial variability can be accounted for using the improved HHDD approach. When training the ANN model, the input and output data of the HYMOD are used as the feeding layer. Therefore, if there are multiple rain gauges, all gauge data can be used instead of the averaged data.

2.5. CONCLUSIONS

In this study, a hybrid modeling approach for daily streamflow forecasting was developed by integrating a conceptual physical process-based model (HYMOD) and a data-driven model (ANN). The developed HHDD model was applied to a subcatchment in the Guadalupe Basin in Texas. The HYMOD model was calibrated then tested based on 7- and 3-year data, respectively. The analysis was carried out in two steps. The first step focused on comparing the calibrated HYMOD model with the four developed HHDD models based on a 4-year validation period. To investigate which HHDD model is better, uncertainty analysis was then carried out as the second step.

The results demonstrated that integrating data-driven modeling with physical process–based modeling is not only promising, but it could also improve the performance of the model, especially when there are several parameters to be calibrated. In addition, including more information such as cumulative rainfall as

an input variable could improve the data-driven models' capability to address hydrological routing without characterizing the associated physical processes.

Overall, the developed HHDD model provides an effective way to improve process-based modeling while avoiding the complexity of parameter calibration. However, HHDD models still have some drawbacks. For example, data-driven models are very dependent on the quantity and quality of data, so it may not be suitable for an area with limited data. Introducing a new data-driven modeling layer(s) could also bring additional complexity and uncertainty; therefore, the development and integration of such layers to the data-driven model should be carefully considered. The developed hybrid approach has only been tested for a lumped hydrological model (i.e., HYMOD model) in this study. Hybrid modeling based on more complex hydrological models, including semidistributed and distributed models, could be investigated in future studies.

2.6. DATA AVAILABILITY STATEMENT

The rainfall data were obtained from the National Oceanic and Atmospheric Administration website at https://www.ncdc.noaa.gov/ (accessed August 1, 2017). The flow gauge data were obtained from the United States Geological Survey website at https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00012912/detail (accessed August 1, 2017). HYMOD is an open-source model and can be obtained from GitHub based

on the model preferences. The developed HHDD model is available from the corresponding author by request.

2.7. ACKNOWLEDGEMENT

This research was supported by the Natural Science and Engineering Research Council of Canada. The authors would like to thank the editor and the anonymous reviewers for their constructive comments that greatly contributed to improving the paper.

2.8. REFERENCES

Adams, T. E., S. Chen, and R. Dymond. 2018. "Results from operational hydrologic forecasts using the NOAA/NWS OHRFC Ohio river com- munity HEC-RAS model." J. Hydrol. Eng. 23 (7): 04018028. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001663.

Bagatur, T., and F. Onen. 2018. "Development of predictive model for flood routing using genetic expression programming." J. Flood Risk Manage. 11 (S1): S444–S454. https://doi.org/10.1111/jfr3.12232. Barati, R. 2011. "Parameter estimation of nonlinear Muskingum models using Nelder-Mead simplex algorithm." J. Hydrol. Eng. 16 (11):

946–954. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000379.

Barati, R., S. Rahimi, G. Akbari, and Y. Yu. 2012. "Analysis of dynamic wave model for flood routing in natural rivers." Water Sci. Eng. 5 (3): 243–258. https://doi.org/10.3882/j.issn.1674-2370.2012.03.001.

Bertone, E., K. O'Halloran, R. A. Stewart, and G. F. de Oliveira. 2017. "Medium-term storage volume prediction for optimum reservoir management: A hybrid data-driven approach." J. Cleaner Prod. 154 (Jun): 353–365. https://doi.org/10.1016/j.jclepro.2017.04.003.

Ch, S., N. Anand, B. K. Panigrahi, and S. Mathur. 2013. "Streamflow fore- casting by SVM with quantum behaved particle swarm optimization." Neurocomputing 101 (Feb): 18–23. https://doi.org/10.1016/j.neucom.2012.07.017.

Chen, C. S., Y. D. Jhong, T. Y. Wu, and S. T. Chen. 2013. "Typhoon event- based evolutionary fuzzy inference model for flood stage forecasting." J. Hydrol. 490 (May): 134–143. https://doi.org/10.1016/j.jhydrol.2013.03.033.

Chen, Y., X. Chen, C. Xu, M. Zhang, M. Liu, and L. Gao. 2018. "Toward improved calibration of SWAT using season-based multi-objective optimization: A case study in the Jinjiang Basin in southeastern China." Water Resour. Manage. 32 (4): 1193–1207. https://doi.org/10.1007/s11269-017-1862-8.

Dariane, A. B., M. Farhani, and S. Azimi. 2018. "Long term streamflow forecasting using a hybrid entropy model." Water Resour. Manage. 32 (4): 1439–1451. https://doi.org/10.1007/s11269-017-1878-0.

De Vos, N. J., T. H. M. Rientjes, and H. V. Gupta. 2010. "Diagnostic evaluation of conceptual rainfall: Runoff models." Hydrol. Processes 24 (20): 2840–2850. https://doi.org/10.1002/hyp.7698.

Ehteram, M., F. B. Othman, Z. M. Yaseen, H. A. Afan, M. F. Allawi,

M. B. A. Malek, A. N. Ahmed, S. Shahid, V. P. Singh, and A. El-Shafie. 2018. "Improving the Muskingum flood routing method using a hybrid of particle swarm optimization and bat algorithm." Water 10 (6): 1–21. https://doi.org/10.3390/w10060807.

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

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

Farzin, S., V. P. Singh, H. Karami, N. Farahani, M. Ehteram, O. Kisi,

M. F. Allawi, N. S. Mohd, and A. El-Shafie. 2018. "Flood routing in river reaches using a three-parameter Muskingum model coupled with an improved bat algorithm." Water 10 (9): 1130. https://doi.org/10.3390/w10091130.

Feng, C., M. Cui, B. M. Hodge, and J. Zhang. 2017. "A data-driven multi- model methodology with deep feature selection for short-term wind forecasting." Appl. Energy 190 (Mar): 1245–1257. https://doi.org/10.1016/j.apenergy.2017.01.043.

Fu, C., A. L. James, and H. Yao. 2014. "SWAT-CS: Revision and testing of SWAT for Canadian shield catchments." J. Hydrol. 511 (Apr): 719–735. https://doi.org/10.1016/j.jhydrol.2014.02.023.

Galelli, S., G. B. Humphrey, H. R. Maier, A. Castelletti, G. C. Dandy, and M. S. Gibbs. 2014. "An evaluation framework for input variable selection algorithms for environmental data-driven models." Environ. Modell. Software 62 (Dec): 33–51. https://doi.org/10.1016/j.envsoft.2014.08.015.

Herman, J. D., P. M. Reed, and T. Wagener. 2013. "Time-varying sensitivity analysis clarifies the effects of watershed model formulation on model behavior." Water Resour. Res. 49 (Feb): 1400–1414. https://doi.org/10.1002/wrcr.20124.

Humphrey, G. B., M. S. Gibbs, G. C. Dandy, and H. R. Maier. 2016. "A hybrid approach to monthly streamflow forecasting: Integrating hydrological model outputs into a Bayesian artificial neural network." J. Hydrol. 540 (Sep): 623–640. https://doi.org/10.1016/j.jhydrol.2016.06.026.

Jimeno-Sáez, P., J. Senent-Aparicio, J. Pérez-Sánchez, and D. Pulido- Velazquez. 2018. "A comparison of SWAT and ANN models for daily runoff simulation in different climatic zones of peninsular Spain." Water 10 (2): 192. https://doi.org/10.3390/w10020192.

Jothiprakash, V., and A. S. Kote. 2011. "Improving the performance of data-driven techniques through data pre-processing for modelling daily reservoir inflow." Hydrol. Sci. J. 56 (1): 168–186. https://doi.org/10.1080/02626667.2010.546358.

Khan, M. Y. A., F. Hasan, S. Panwar, and G. J. Chakrapani. 2016. "Neural network model for discharge and water-level prediction for Ramganga River catchment of Ganga Basin, India." Hydrol. Sci. J. 61 (11): 2084– 2095. https://doi.org/10.1080/02626667.2015.1083650.

Kothari, M., and K. D. Gharde. 2015. "Application of ANN and fuzzy logic algorithms for streamflow modelling of Savitri catchment." J. Earth Syst. Sci. 124 (5): 933–943. https://doi.org/10.1007/s12040-015-0592-7.

Li, Z., Q. Shao, Z. Xu, and X. Cai. 2010. "Analysis of parameter uncertainty in semi-distributed hydrological models using bootstrap method: A case study of SWAT model applied to Yingluoxia watershed in northwest China." J. Hydrol. 385 (1–4): 76–83. https://doi.org/10.1016/j.jhydrol.2010.01.025.

Mendonça, F., R. P. De Oliveira, and F. F. Mauad. 2018. "Lumped versus distributed hydrological modeling of the Jacaré-Guaçu Basin, Brazil."

J. Environ. Eng. 144 (8): 1–13. https://doi.org/10.1061/(ASCE)EE.1943-7870.0001397.

Moore, R. J. 1985. "The probability-distributed principle and runoff production at point and basin scales." Hydrol. Sci. J. 30 (2): 273–297. https://doi.org/10.1080/02626668509490989.

Nanda, T., B. Sahoo, H. Beria, and C. Chatterjee. 2016. "A wavelet-based non-linear autoregressive with exogenous inputs (WNARX) dynamic neural network model for real-time flood forecasting using satellite- based rainfall products." J. Hydrol. 539 (Aug): 57–73. https://doi.org/10.1016/j.jhydrol.2016.05.014.

Nourani, V., A. H. Baghanam, J. Adamowski, and O. Kisi. 2014. "Applications of hybrid wavelet-artificial intelligence models in hydrology: A review." J. Hydrol. 514 (June): 358–377. https://doi.org/10.1016/j.jhydrol.2014.03.057.

Özger, M. 2009. "Comparison of fuzzy inference systems for streamflow prediction." Hydrol. Sci. J. 54 (2): 261–273. https://doi.org/10.1623/hysj.54.2.261.

Özger, M., A. K. Mishra, and V. P. Singh. 2012. "Long lead time drought forecasting using a wavelet and fuzzy logic combination model: A case study in Texas." J. Hydrometeorol. 13 (1): 284–297. https://doi.org/10.1175/JHM-D-10-05007.1.

Pramanik, N., R. K. Panda, and A. Singh. 2011. "Daily river flow forecasting using wavelet ANN hybrid models." J. Hydroinf. 13 (1): 49–63. https://doi.org/10.2166/hydro.2010.040.

Quan, Z., J. Teng, W. Sun, T. Cheng, and J. Zhang. 2015. "Evaluation of the HYMOD model for rainfall-runoff simulation using the GLUE method." Proc. IAHS 368: 180–185. https://doi.org/10.5194/piahs-368-180-2015.

Singh, S. K., and N. Marcy. 2017. "Comparison of simple and complex hydrological models for predicting catchment discharge under climate change." AIMS Geosci. 3 (3): 467–497. https://doi.org/10.3934/geosci.2017.3.467.

Song, X., F. Kong, C. Zhan, and J. Han. 2012. "Hybrid optimization rainfall-runoff simulation based on Xinanjiang model and artificial neural network." J. Hydrol. Eng. 17 (9): 1033–1041. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000548.

Sudheer, C., R. Maheswaran, B. K. Panigrahi, and S. Mathur. 2014. "A hybrid SVM-PSO model for forecasting monthly streamflow." Neural Comput. Appl. 24 (6): 1381–1389. https://doi.org/10.1007/s00521-013-1341-y.

Tiwari, M. K., and C. Chatterjee. 2011. "A new wavelet–bootstrap–ANN hybrid model for daily discharge forecasting." J. Hydroinf. 13 (3): 500–519. https://doi.org/10.2166/hydro.2010.142.

Vrugt, J. A., C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson. 2008. "Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation." Water Resour. Res. 44 (12): 1–15. https://doi.org/10.1029/2007WR006720.

Wang, K., and A. Altunkaynak. 2012. "Comparative case study of rainfall- runoff modeling between SWMM and fuzzy logic approach." J. Hydrol. Eng. 17 (2): 283–291. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000419.

Wang, Y., S. Guo, L. Xiong, P. Liu, and D. Liu. 2015. "Daily runoff fore- casting model based on ANN and data preprocessing techniques." Water 7 (8): 4144–4160. https://doi.org/10.3390/w7084144.

Wi, S., Y. C. E. Yang, S. Steinschneider, A. Khalil, and C. M. Brown. 2015. "Calibration approaches for distributed hydrologic models in poorly gaged basins: Implication for streamflow projections under climate change." Hydrol. Earth Syst. Sci. 19 (2): 857–876. https://doi.org/10.5194/hess-19-857-2015.

Wu, M. C., G. F. Lin, and H. Y. Lin. 2014. "Improving the forecasts of extreme streamflow by support vector regression with the data extracted by self-organizing map." Hydrol. Processes 28 (2): 386–397. https://doi.org/10.1002/hyp.9584.

Yang, J., P. Reichert, K. C. Abbaspour, J. Xia, and H. Yang. 2008. "Comparing uncertainty analysis techniques for a SWAT application to the Chaohe Basin in China." J. Hydrol. 358 (1–2): 1–23. https://doi.org/10.1016/j.jhydrol.2008.05.012.

Yang, J., P. Reichert, K. C. Abbaspour, and H. Yang. 2007. "Hydrological modelling of the Chaohe Basin in China: Statistical model formulation and Bayesian inference." J. Hydrol. 340 (3–4): 167–182. https://doi.org/10.1016/j.jhydrol.2007.04.006.

Zeroual, A., M. Meddi, and A. A. Assani. 2016. "Artificial neural network rainfall-discharge model assessment under rating curve uncertainty and monthly discharge volume predictions." Water Resour. Manage. 30 (9): 3191–3205. https://doi.org/10.1007/s11269-016-1340-8.

Zhang, J., Y. Li, G. Huang, X. Chen, and A. Bao. 2016. "Assessment of parameter uncertainty in hydrological model using a Markov-chain- Monte-Carlo-based

multilevel-factorial-analysis method." J. Hydrol. 538 (Jul): 471–486. https://doi.org/10.1016/j.jhydrol.2016.04.044.

Zhang, L., C. He, X. Bai, and Y. Zhu. 2017. "Physically based adjustment factors for precipitation estimation in a large arid mountainous water- shed, northwest China." J. Hydrol. Eng. 22 (11): 04017047. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001576.

Zheng, Y., and F. Han. 2016. "Markov chain Monte Carlo (MCMC) uncertainty analysis for watershed water quality modeling and management." Stochastic Environ. Res. Risk Assess. 30 (1): 293–308. https:// doi.org/10.1007/s00477-015-1091-8

Chapter 3

PROPAGATION OF PARAMETER UNCERTAINTY IN SWAT: A PROBABILISTIC FORECASTING METHOD BASED ON POLYNOMIAL CHAOS EXPANSION AND MACHINE LEARNING

ABSTRACT

Soil and Water Assessment Tool (SWAT) is one of the most widely used semi-distributed hydrological models. Assessment of the uncertainties in SWAT outputs is a popular but challenging topic due to the significant number of parameters. The purpose of this study is to investigate the use of Polynomial Chaos Expansion (PCE) in assessing uncertainty propagation in SWAT under the impact of significant parameter sensitivity. Furthermore, for the first time, a machine learning technique (i.e., artificial neural network, ANN) is integrated with PCE to expand its capability in generating probabilistic forecasts of daily flow. The traditional PCE and the proposed PCE-ANN methods are applied to a case study in the Guadalupe watershed in Texas, USA to assess the uncertainty propagation in SWAT for flow prediction during the historical and forecasting periods. The results show that PCE provides similar results as the traditional Monte-Carlo (MC) method, with a coefficient of determination (R^2) value of 0.99 for the mean flow, during the historical period; while the proposed PCE-ANN method reproduces MC output with a R^2 value of 0.84 for mean flow during the forecasting period. It is also indicated that PCE and PCE-ANN are as reliable as but much more efficient than

MC. PCE takes about 1% of the computational time required by MC; PCE-ANN only takes a few minutes to produce probabilistic forecasting, while MC requires running the model for dozens or hundreds, even thousands, of times. Notably, the development of the PCE-ANN framework is the first attempt to explore PCE's probabilistic forecasting capability using machine learning. PCE-ANN is a promising uncertainty assessment and probabilistic forecasting technique, as it is more efficient in terms of computation time, and it does not cause loss of essential uncertainty information. https://doi.org/10.1016/j.jhydrol.2020.124854

Keywords: Daily streamflow simulation Data-driven modeling, Polynomial chaos expansion (PCE), Soil and water assessment tool (SWAT) Uncertainty analysis, Probabilistic forecasting

3.1. INTRODUCTION

In the water resources sector, designers, researchers, and decision-makers rely intensively on hydrological models. Hydrological models have become widely applied and highly evolved along with advancements in computing technology. Researchers have developed many hydrological models with different structures, assumptions, and processes. The variety of hydrological models makes it difficult for users to decide which model best represents a watershed's hydrological processes. The Soil and Water Assessment Tool (SWAT) has gained a lot of attention since Arnold et al. first proposed it in 1998. SWAT has been widely used for watershed modeling for several reasons: 1) SWAT is an open source model, which allows researchers and engineers to debug and improve the algorithm for higher accuracy for specific study areas (Eckhardt et al., 2002; Fu et al., 2014; D. Zhang et al., 2016); 2) SWAT is a semi-distributed model, which enables it to better balance the tradeoff between simulation efforts and accuray compared to both lumped and fully-distributed and models when simulating large non-homegenous watersheds (Eckhardt et al., 2002; Fu et al., 2014; D. Zhang et al., 2016); 3) SWAT has the ability to perform not only hydrological modeling but also simulate sediment transport and water quality processes (Arabi et al., 2008; Dagnew et al., 2016; Debele et al., 2008; Hallouz et al., 2018; Iudicello et al., 2013); 4) SWAT is compatible with geographic information system software (GIS) which can present the results in a more informative and interactive manner to assist decision makers

in interpreting different scenarios (Olivera et al., 2006; Shen et al., 2013; Suliman et al., 2015).

SWAT, like many other hydrological models, needs to be calibrated and validated. The calibration process is performed on parameters that are either not directly measured or very sensitive or missing; the values of these parameters are adjusted to minimize the deviation between simulated results and observations. One of the most widely used SWAT calibration toolkits is the SWAT Calibration and Uncertainty Procedures (SWAT-CUP) (Arnold et al., 2012). The SWAT-CUP toolbox provides several techniques for uncertainty analysis, including generalized likelihood uncertainty estimation (GLUE), sequential uncertainty fitting-2 (SUFI-2), parameter solution (Para-Sol), and Markov chain Monte Carlo (MCMC). The computational time for all of these methods is relatively long, and it increases with respect to watershed size and the number of parameters. Several approaches have been developed and tested to enhance calibration efficiency and to reduce computational time. For example, Zhang et al. (2009) used a genetic algorithm (GA) to calibrate the SWAT model. Using Bayesian Model Averaging algorithm (BMA) with the best model ensembles, the GA-BMA model was able to address prediction uncertainty using interval estimation while requiring less computational time compared to MCMC. Comparing the computation time of this method to MCMC, almost one order of magnitude in hours was reduced while producing comparable results. Li et al. (2010) used a bootstrap algorithm with BMA to calibrate a SWAT model. The bootstrap algorithm was used to calibrate the model

first and to find the best set of parameters from a prior distribution. Then, the model was recalibrated by generating new 'observation' data from adding some residuals to the actual observation. This process was repeated for several iterations to produce a posterior distribution for each parameter. The bootstrap method used less computational time than the GA-BMA method, but its performance in terms of uncertainty quantification varied significantly from case to case and thus was not always reliable. These previous studies indicate that there is a need for more reliable and efficient uncertainty analysis methods for SWAT.

Recently, polynomial chaos expansion (PCE) has been proposed as a new method for uncertainty quantification. Based on previous studies, PCE requires much less computational time than MCMC when used for analyzing the propagation of parameter uncertainties. This algorithm has been used in many fields such as transportation (Stavropoulou and Muller, 2015), chemical process (Paffrath and Wever, 2007; Villegas et al., 2012), and aerodynamics (Wu et al., 2018). It has also been used in water-related areas, such as groundwater flow and contaminant transport (Deman et al., 2016; Laloy et al., 2013; Li and Zhang, 2007; Rupert and Miller, 2007) and computational fluid dynamics (Hosder, 2010; Najm, 2009; Tagade and Choi, 2014). More recently, PCE has been introduced to surface water modeling to quantify model uncertainty. Fan et al. (2014) and Wang et al. (2015) used the PCE with the lumped model HYMOD to test the potential of PCE for assessing model uncertainty. Fan et al. (2014) investigated the use of second- and third-order PCE to address the parameter uncertainties in HYMOD. The

uncertainties of two parameters were analyzed, and one year of synthetic data was used to compare the results generated by Monte-Carlo (MC) simulation, second-order PCE, and third-order PCE. Both second- and third-order PCE provided similar results as MC. The third-order PCE generated results closer to those generated by MC, but was more computationally demanding relative to second-order PCE. Wang et al. (2015) used PCE and reduced PCE as well as MC with latent hypercube sampling (LHS) to conduct the uncertainty analysis for a case study of the Xiangxi River Watershed in China. The HYMOD model was used to simulate two years of flow. The results showed that PCE and reduced PCE gave similar results and almost the same Nash–Sutcliffe model efficiency (NSE) as MC-LHS. However, the time required to perform MC-LHS was 55 seconds, whereas it only took 4.8 and 3.5 seconds for PCE and reduced PCE, respectively. Both studies demonstrated the promising application of PCE in replacing MC for uncertainty analysis in hydrological modeling. PCE has the potential for application as an efficient technique that can save time and computational resources in uncertainty quantification, and its advantages would only be more significant when applied to more sophisticated hydrological models. However, the HYMOD tested in these two studies is only a simple conceptual model, which cannot demonstrate PCE's applicability and reliability for more sophisticated hydrological models. As such, the application of PCE on more complicated models, such as a distributed or semi-distributed model, requires further investigation. Besides, the existing PCE method relies on observations to quantify the propagation of parameter uncertainties with a

model, which prevents it from providing hydrological forecasts under uncertainty. Therefore, PCE's capability for probabilistic forecasting should also be explored.

The objective of this study is to investigate the capability of PCE in building a surrogate of SWAT and demonstrating that PCE could quantify the SWAT's parameter uncertainty and provide probabilistic forecasting in a much more efficient way, compared to the traditional MC method. To enable PCE to generate hydrological forecasts under uncertainty with SWAT, an artificial neural network machine learning algorithm (ANN) will be integrated with PCE. The new PCE-ANN method will be used to build a surrogate model for SWAT, which can forecast daily flow and quantify the uncertainties associated with the obtained forecasts efficiently. The proposed PCE-ANN is applied to a case study in Guadalupe River Watershed in Texas, USA. This paper is divided into seven sections. Section 3.2 describes the setup of the SWAT model, as well as its automatic calibration toolbox SWAT-CUP. Section 3.3 illustrates the framework and development processes of PCE and PCE-ANN. Section 3.4 describes the study area and the data used in this paper. Section 3.5 discusses the results of this study. Section 3.6 discusses the advantages and limitations of the PCE-ANN method. Finally, section 3.7 summarizes the conclusions of this study.

3.2. SWAT MODEL

SWAT is a semi-distributed model that has been widely used for hydrological and environmental modeling. SWAT has gained much attention as it is open-source

software and many databases, documentation, and publications are available to the public. Additionally, several software products (such as ArcSWAT, QSWAT) are available for SWAT to provide users with a user-friendly interface and to present results as intuitive and informative maps. SWAT is standardized for US topographic and weather conditions and all input data required to build a SWAT model is integrated within the SWAT database (White et al., 2016, 2017). As shown in Figure 3-1, SWAT is fed by multiple databases, including the digital elevation models (DEMs), streamlines, soil data, and land uses.

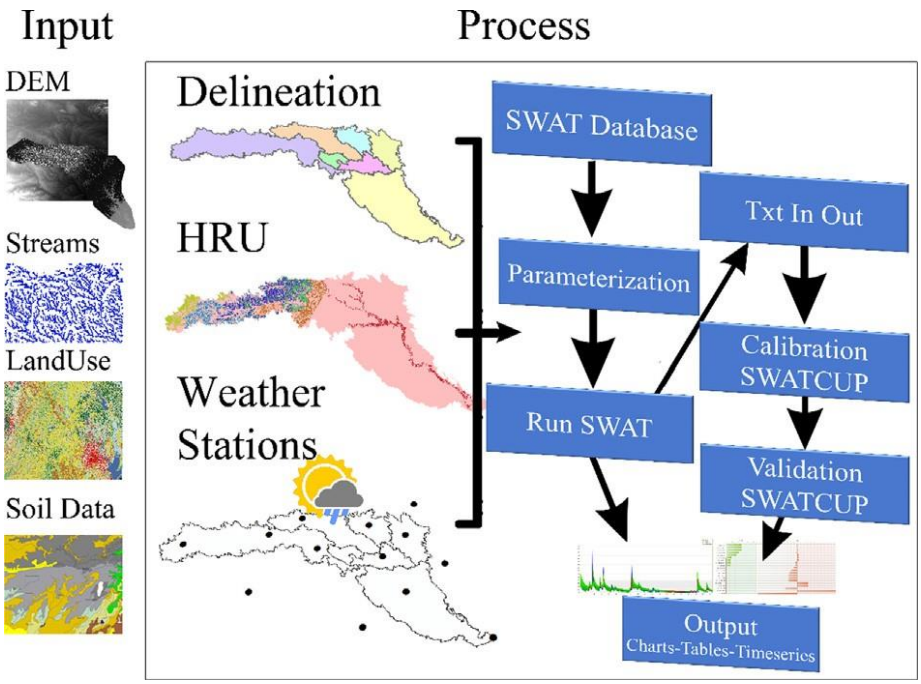

Figure 3-1: SWAT model schematic.

To set-up a SWAT model, streamlines are first burnt to the DEM to make sure that automatic delineation is accurate. Second, outlet points are selected at the desired flow gages. Third, the delineation process is carried out based on the chosen

outlets. Fourth, the study area is classified according to land-use, soil characteristics, and slope to determine the hydrologic response units (HRUs) for each sub-catchment. Fifth, all weather data are fed into the model as a time-series table. Sixth, model parameters are estimated based on the slope, land use, soil data, and weather conditions. Finally, SWAT is ready for calibration and validation, preferably with a warmup period to avoid any initialization error. There are several different approaches for model calibration. One of the most widely used toolkits is SWAT-CUP due to its ease of use and the variety of techniques to be used for calibration.

There are three automatic calibration algorithms embedded in SWAT-CUP: SUFI, GLUE, and Para-Sol. In this study, SUFI is used for model calibration for several reasons. SUFI is the fastest algorithm, as it uses the LHS technique to cover all the range. SUFI depends on re-running the model several times with a narrower parameter range each time. Additionally, in previous studies, SUFI has been found to have a slightly better performance compared to the other two algorithms (Khatun et al., 2018; Singh et al., 2013). The calibration can be done after defining the uniform distribution of each parameter, the number of iterations, and the objective function. The parameter ranges used in this study are determined based on the previous studies and are shown in Table 3-1. Parameters that have different values spatially are changed by multiplying the original value by a ratio in order to keep the spatial distribution relatively consistent. The Curve Number (*CN2*) is an example of the spatial parameters, so it is defined as relative change "R". Other

parameters that have a specific value assigned across the catchment are changed by randomly selecting a value from the range distribution. These parameters are defined as replacement change "V". The objective of the calibration is to maximize the NSE value. After running the automatic calibration tool, the best parameter set is chosen, and the model can be re-run in the SWAT-CUP for validation or in SWAT for further simulation.

SWAT-CUP can provide a global sensitivity analysis report during the calibration process. The sensitivity report shows the most sensitive parameters that should be considered as the target parameters for uncertainty analysis. The sensitivity report can be in the form of scatter plots or in a statistical format (*p*-test and *t*-test). If the scatter plot is uniformly distributed along the range, then the corresponding parameter is not sensitive. If there is a clear trend in the scatter points, then this parameter is sensitive. The statistical format is preferred over the scatter plot format in order to avoid human bias. The parameter is considered sensitive if the *p*-value is lower than 0.05, which also indicates that the *t*-test values are high (Khatun et al., 2018; Yesuf et al., 2016). When two parameters have the same p-value, a *t*-test may be used for differentiating them. In this study, the five most significant parameters identified using this process are used for further uncertainty analysis.

Table 3-1: Model parameters.

Parameter ID	Rule	Min	Max	Parameter description
R__CN2.mgt	Ratio	-0.2	0.2	SCS runoff curve number
V__CH_W2.rte	Replace	0	1000	Average width of main channel.
V__CH_L2.rte	Replace	-0.05	500	Length of main channel.
V__CH_K2.rte	Replace	5	130	Hydraulic conductivity in main channel
R__SOL_AWC(..).sol	Ratio	-0.2	0.4	Soil available water content
R__SOL_BD(..).sol	Ratio	-0.5	0.6	Soil moist bulk density
V__ALPHA_BF.gw	Replace	0	1	Baseflow alpha factor (days)
V__ESCO.hru	Replace	0.8	0.95	Soil evaporation compensation factor
V__GW_DELAY.gw	Replace	30	450	Groundwater delay (days)
R__SOL_K(..).sol	Ratio	-0.8	0.8	Saturated hydraulic conductivity
V__CH_D.rte	Replace	0	30	Average depth of main channel.
V__TLAPS.sub	Replace	-10	10	Temperature lapse rate
V__TIMP.bsn	Replace	0.05	0.9	Snowpack temperature lag factor
V__GWQMN.gw	Replace	0	2	Threshold depth of water in the shallow aquifer required for return flow to occur (mm)
V__SMFMX.bsn	Replace	-5	5	Maximum melt rate for snow during year (occurs on summer)
V__GW_REVAP.gw	Replace	0	0.2	Groundwater "revap" coefficient
V__REVAPMN.gw	Replace	0	500	Threshold depth of water in the shallow aquifer for "revap" to occur (mm)
V__CH_S2.rte	Replace	-0.001	10	Average slope of main channel.
V__SMFMN.bsn	Replace	-5	5	Minimum melt rate for snow during year (occurs on winter)
V__SNO50COV.bsn	Replace	0	0.9	Snow water equivalent that corresponds to 50% snow cover
V__CANMX.hru	Replace	0	100	Maximum canopy storage
V__SNOCOVMX.bsn	Replace	0	500	Minimum snow water content that corresponds to 100% snow cover
V__CH_N2.rte	Replace	0	0.3	Manning's "n" value for the main channel
V__SOL_ALB(..).sol	Replace	0	0.25	Moist soil albedo
V__SURLAG.bsn	Replace	0.1	20	Surface runoff lag time
V__SFTMP.bsn	Replace	-5	5	Snowfall temperature
V__EPCO.bsn	Replace	0	0.9	Plant uptake compensation factor

3.3. UNCERTAINTY QUANTIFICATION

3.3.1. POLYNOMIAL CHAOS EXPANSION (PCE)

PCE is a statistical method which describes the uncertainties in a system using normally distributed random inputs. The statistical process is a composition of independents centered normalized Gaussian random variables presented by Hermite polynomial (Fan et al., 2014; Wang et al., 2015). The PCE equation is written as:

$$y = a_0 + \sum_{i=1}^{n} a_i \Gamma_1(\zeta_i) + \sum_{i=1}^{n}\sum_{j=1}^{i} a_{i,j} \Gamma_2(\zeta_i, \zeta_j) + ... \quad (3\text{-}1)$$

Where y is the output and $\Gamma_p(\zeta_1, \zeta_2, ...\zeta_p)$ is the polynomial chaos of order p. In previous studies, an approximation has been made to truncate PCE to a lower level to reduce the computational time while maintaining reasonable accuracy (Fan et al., 2014; Wang et al., 2015). The number of PCE terms (N) is a function of the PCE order (P) and the number of random variables used for uncertainty analysis (M):

$$N = \frac{(M+P)!}{M!P!} \quad (3\text{-}2)$$

In this study, a second-order Hermite polynomial is used to quantify uncertainties associated with five parameters, giving a total of 21 PCE terms. Thus, the equation of the output can be written as follows:

$$y = a_0 + a_1\zeta_1 + a_2\zeta_2 + a_3\zeta_3 + a_4\zeta_4 + a_5\zeta_5 + a_6\zeta_1\zeta_2 + a_7\zeta_1\zeta_3 + a_8\zeta_1\zeta_4 + a_9\zeta_1\zeta_5 + a_{10}\zeta_2\zeta_3 + a_{11}\zeta_2\zeta_4 + a_{12}\zeta_2\zeta_5 + a_{13}\zeta_3\zeta_4 + a_{14}\zeta_3\zeta_5 + a_{15}\zeta_4\zeta_5 + a_{16}(\zeta_1^2 - 1) + a_{17}(\zeta_2^2 - 1) + a_{18}(\zeta_3^2 - 1) + a_{19}(\zeta_4^2 - 1) + a_{20}(\zeta_5^2 - 1) \quad (3\text{-}3)$$

3.3.2. SELECTION OF COLLOCATION POINTS

The main idea of selecting the collocation points is to have the PCE output for the random input be the same as the model output at those selected points. Once the equations are established, the coefficients in Equation 3-3 can be obtained. This coefficient estimation method is called the probabilistic collocation method (PCM). The collocation points can be selected using the combination of the higher Hermite polynomial roots. Thus, for the second-order Hermite polynomial, the collocation points are the combination of the three roots $(-\sqrt{3}, 0, \sqrt{3})$ for each ζ value. In this study, there are five analyzed parameters for a total of 243 collocation points. Since there are only 21 unknowns and 243 equations (realizations), this system of equations is overdetermined unless there is redundancy in the equations. Solving the system of equations using linear regression is feasible in the historical period, as the simulation output of the model is known (i.e., observations are available). As shown in Figure 3-2a, the model output can be expressed as a set of Hermite orthogonal polynomials in terms of standard normal random variables using PCE, which can be used to assess the propagation of parameter uncertainty. The PCE coefficients can be estimated using observation data during the historical period. However, for the forecasting period, the coefficients are unknown due to the lack of observation data, as shown in Figure 3-2b. Thus, the uncertainties associated

with the forecasts cannot be quantified using the traditional PCE method. In this study, a machine learning technique, i.e., artificial neural network (ANN), is introduced to the traditional PCE framework to enable the analysis of uncertainties associated with simulated future time series and to generate probabilistic forecasts.

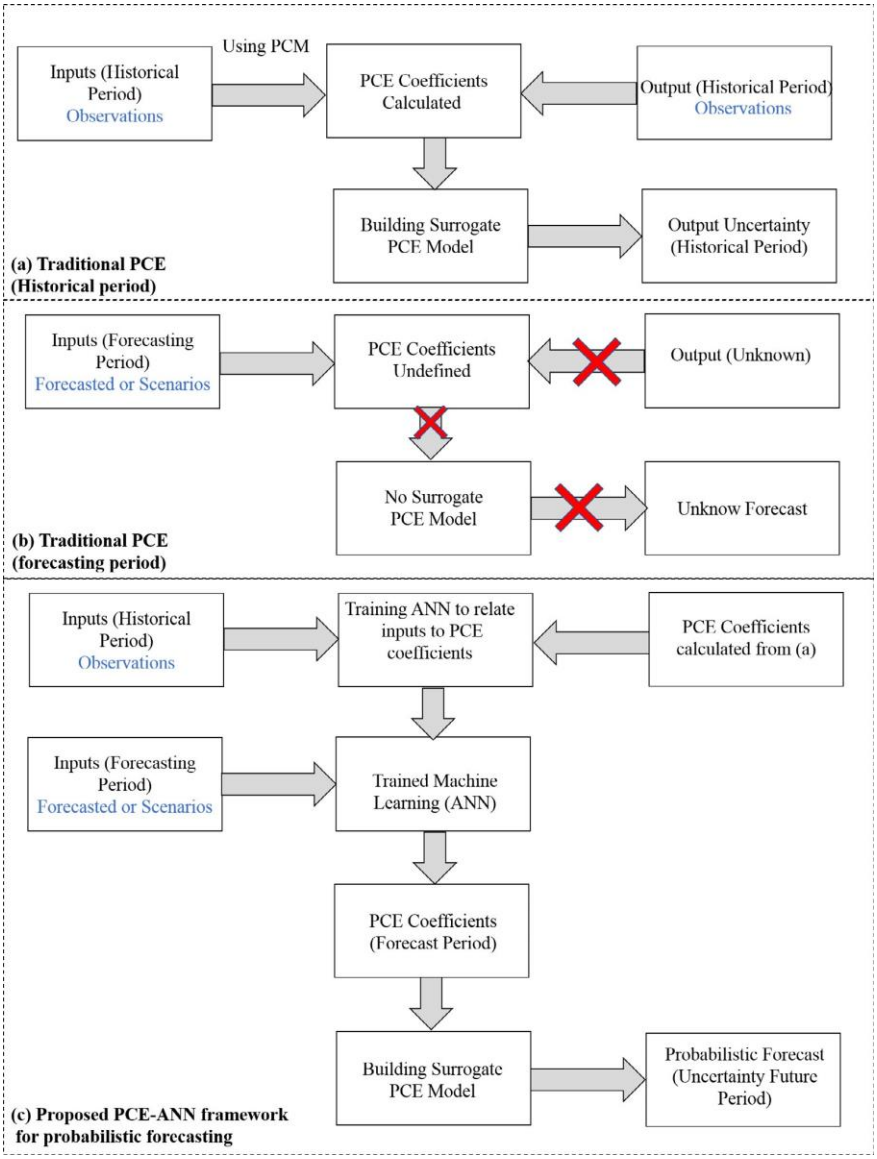

Figure 3-2: Flowchart PCE Framework (a) Traditional PCE in historical period (b) Traditional PCE in forecasting period (c) Proposed PCE-ANN in forecasting period.

3.3.3. ESTIMATION OF PCE COEFFICIENTS BASED ON MACHINE LEARNING

To address the aforementioned issue, a machine learning technique is implemented to estimate the PCE coefficients for the forecasting period where observed flow data is unavailable. In the forecasting period, the coefficients and outputs are unknown as shown in Figure 3-2b. As shown in Figure 3-2c, ANN is used to build the relationships between meteorological data and PCE coefficients in the historical (training) period. The trained ANN model is then used to estimate the PCE coefficients in the forecasting period, which are directly used as the coefficients of the PCE terms. Finally, the Hermite polynomials can be obtained to build a PCE surrogate model, and thus, the uncertainties associated with the future flow can be quantified through probabilistic forecast. The ANN model built in this study consists of 18 input variables, including multi-day cumulative precipitation (denoted as P1, P2, P3, P4, P5, P6, P7, P15, P30, P45, P60, P90, P120, P150, and P180), maximum daily temperature, minimum daily temperature, and wind speed. The output is the 21 PCE coefficients. The ANN model is built using one hidden layer, which consists of 25 neurons. The number of hidden layers and neurons is determined based on a series of sensitivity analysis. A backpropagation algorithm is used to find the weights and coefficients during the training period of the ANN model. Since the ANN model produces different weights and coefficients every time the model is trained, the model is trained for hundred realizations, and the average is taken as the model output.

3.3.4. THE UNCERTAINTY FRAMEWORK

Figure 3-3 summarizes the process of uncertainty analysis for SWAT using both MC and PCE-ANN. First, a SWAT model is built for the study area. Then, the SWAT model is calibrated using an automatic calibration tool SWAT-CUP. The goal of this process is to find the most sensitive parameters from the global sensitivity report as well as the best values for the non-sensitive parameters. After selecting the sensitive parameters, both MC and PCE-ANN can be performed. To perform MC, the chosen parameters are changed randomly along its physical range for 10,000 realizations to find the uncertainty in the prediction period. On the other hand, to perform PCE-ANN uncertainty analysis, the chosen parameters are assumed to be independent and are transformed to standard normal distributions. Collocation points are then selected from the normal distribution for each parameter. Then, the SWAT model is run with all the possible combinations of the parameter values at the collocation points, and one output value is obtained for each combination of parameter points at each time step. Subsequently, a linear equation, where the PCE coefficients are the unknowns, can be established for each parameter combination at each time step. The system of linear equations can be solved to find the values of the PCE coefficients. With the obtained PCE coefficients, the PCE surrogate model is built for the calibration period. In order to enable PCE to quantify the uncertainties in the future period, machine learning is used. A machine learning technique (i.e., ANN) is trained to find the relationships between metrological data and PCE coefficients during the historical period so that the PCE

coefficients can be found during the forecasting period based on meteorological forecasts. Finally, a PCE equation can be obtained to quantify the uncertainties at each time step of the forecasting period without running the SWAT Model or having flow observations.

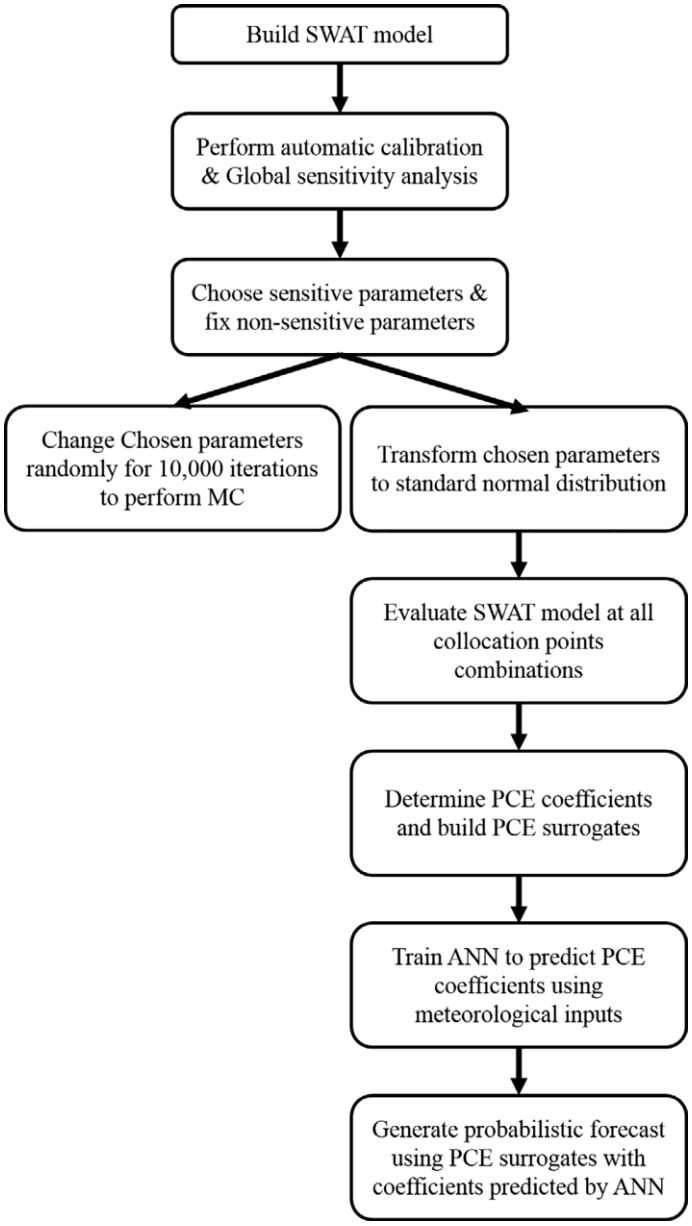

Figure 3-3: Uncertainty analysis framework.

3.4. STUDY AREA AND DATA COLLECTION

A case study of the Guadalupe basin in Texas, USA is used to illustrate the applicability of the proposed PCE-ANN method. The Guadalupe basin, as shown in Figure 3-4, is located in southeastern Texas with an outflow into the Gulf of Mexico. The study area is the upper sub-catchment that is gauged at Spring Branch. The Spring Branch watershed is approximately 3,500 km^2 and has an average slope of 0.12% with a 345 m elevation drop along the longest 290 km stream path. Accordingly, the watershed can be considered a medium to large basin with a mild slope.

Four years of rainfall, temperature, and flow data for the Spring Branch catchment were used in this case study. Weather data is from the National Oceanic and Atmospheric Administration (NOAA), flow data is from the United States Geological Survey (USGS), soil data of a 30 m resolution is obtained from State Soil Geographic USD, land-cover data of a 30 m resolution is from the National Land-Cover Data Sets (NLCD), and DEM of 90 m resolution is from SRTM V4.1 data that is derived from the USGS/NASA SRTM data with a square grid of a size 5 degrees. The data cover the period between 1988 to 1997 with daily resolution. The data were divided into three parts for modeling purposes. First year (1988) was used as a warmup period; three years (1989-1991) of data were used for calibration and sensitivity analysis as this period is assumed to be historical; six years (1992-1997) of data were used for flow forecasting as well as uncertainty analysis,

assuming it is the forecasting (i.e., 'future') period. The basin lies in a moderate temperature zone with an overall average maximum temperature of 27 ºC and an average minimum of 16 ºC throughout the year. The basin is in a semi-arid region where the average rainfall is 2.48 mm/day, resulting in a relatively low average streamflow most of the year of less than 17.60 m^3/s. However, there are few days with extreme events causing high peak flow that could reach to approximately 1,600 m^3/s during the calibration period and 1,900 m^3/s during the forecasting period.

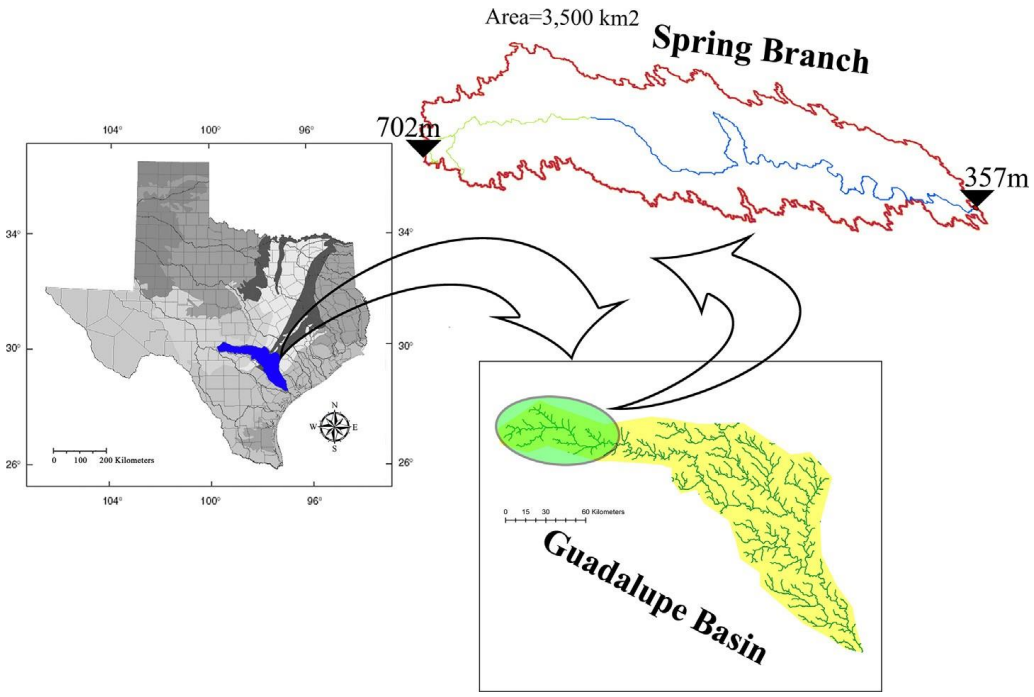

Figure 3-4: Case study map.

3.5. RESULTS

3.5.1. CALIBRATION AND PARAMETER SENSITIVITY

Calibration of the SWAT model was performed for a total of 27 parameters over the historical period from 1989 to 1991. During the automatic calibration process, four parameters were updated by a particular ratio to maintain spatial consistency, while the remaining parameters were replaced using values randomly chosen from the uniform distributions over their physical ranges shown in Table 3-1. After running an automatic calibration process for 2,000 iterations, the fitted values were extracted as shown in Table 3-2. The calibration process produces results with an NSE value of 0.77. This result implies that SWAT can provide satisfactory daily flow simulation results for the study area (Li et al., 2010; Yang et al., 2008, 2007; J. Zhang et al., 2016). It is worth mentioning that the model performance was also tested using six years of data for calibration and three years for validation. There was no significant change in model performance, when the length of calibration period is changed, which further demonstrates SWAT's capability of simulating the Spring Branch catchment.

The sensitivity analysis results as the *t*- and *p*-values are also presented in Table 3-2. As shown in Table 3-2, there are 11 sensitive parameters based on the significance level of the *p*-value (*p*-value less than 0.05). *CN* is the most sensitive parameter based on the *t*-stat. The other 10 sensitive parameters can be classified into three groups: the first group describes channel properties, which can be

measured directly through stream and watershed survey and thus were not chosen as uncertain parameters. The second group contains soil moisture and density parameters, which are the most essential sensitive parameters as they have a significant impact on infiltration, runoff, and evapotranspiration processes. The third group of sensitive parameters is related to baseflow, which determines the amount of infiltrated water that recontributes to streamflow, as well as the travel time to reach the watershed outlet. In this study, the propagation of the uncertainties of five parameters were assessed. The five parameters include *CN* and two parameters from both the second and third groups. From the second group, soil evaporation compensation factor (*ESCO)* and available water content (*Sol_AWC*) were selected to represent the evaporation process and the amount of infiltration, respectively. From the third group, Baseflow alpha factor (*alpha-bf)* and Groundwater delay (*GW_Delay*) were selected to determine the ratio of groundwater contributing to streamflow, and the time it will take to reach the stream. It is worth pointing out that changing the calibration algorithm and/or the length of calibration period could result in different parameter values and different sensitive parameter sets; however, the PCE-ANN framework to be developed would remain the same._The proposed model is capable of analyzing the uncertainties of all sensitive parameters, but only the most representative five parameters were chosen for demonstration purposes.

Table 3-2: Calibrated parameter values.

Parameter ID	Fitted	t-Stat	p-Value
R__CN2.mgt	0.01	-47.36	0
V__CH_W2.rte	294.75	-17.92	0
V__CH_L2.rte	85.58	-15.54	0
V__CH_K2.rte	27.22	-12.94	0
R__SOL_AWC(..).sol	0.16	8.47	0
R__SOL_BD(..).sol	-0.29	-7.69	0
V__ALPHA_BF.gw	0.01	7.48	0
V__ESCO.hru	0.88	-7.45	0
V__GW_DELAY.gw	271.18	6.48	0
R__SOL_K(..).sol	-0.7	-6.3	0
V__CH_D.rte	5.8	2.24	0.03
V__TLAPS.sub	-2.97	-1.57	0.12
V__TIMP.bsn	0.8	1.26	0.21
V__GWQMN.gw	0.53	-1.21	0.23
V__SMFMX.bsn	2.53	-1.18	0.24
V__GW_REVAP.gw	0.12	1.17	0.24
V__REVAPMN.gw	276.88	-1.12	0.26
V__CH_S2.rte	2.98	1.05	0.3
V__SMFMN.bsn	-1.09	-1	0.32
V__SNO50COV.bsn	0.22	-0.96	0.34
V__CANMX.hru	84.58	0.92	0.36
V__SNOCOVMX.bsn	14.38	0.89	0.37
V__CH_N2.rte	0.29	-0.81	0.42
V__SOL_ALB(..).sol	0.08	-0.46	0.65
V__SURLAG.bsn	7.58	-0.33	0.74
V__SFTMP.bsn	-4.62	-0.22	0.83
V__EPCO.bsn	0.55	0.21	0.84

3.5.2. BUILDING A PCE SURROGATE FOR SWAT

Before exploring and improving PCE's capability of addressing forecasting uncertainties, the traditional PCE's performance for quantifying the uncertainties during the historical period was evaluated. Based on the calibration results,

optimized values of SWAT's parameters were found. However, as shown in the sensitivity analysis results, some sensitive parameters lead to significant uncertainty. The uncertainty analysis for SWAT was performed by setting all the parameters to fixed values (as in Table 3-1) and only changing the chosen five sensitive parameters. To perform the PCE on the historical period (1989-1991), the SWAT model was run for 243 times at the collocation point combinations. After performing those runs, the PCE coefficients were calculated, and the parameter uncertainty that propagated to the model output was described using Hermite Polynomials of the standard Gaussian variable, as shown in Equation 3-1.

In order to assess PCE's reliability in terms of quantifying SWAT's parameter uncertainty, the PCE's results were compared to those of a traditional uncertainty quantification technique, Monte Carlo (MC) simulation. 10,000 ensembles were taken from the parameter distribution and used for SWAT simulation for MC uncertainty. The ensembles were taken from the normal distribution to be used in the PCE equation (surrogate model). From the 10,000 ensembles, the mean flow and the variance were calculated. As shown in Figure 3-5 and Figure 3-6, the mean daily flow is almost the same for both MC and PCE. As shown in Figures 3-7 and 3-8, the variance tends to be higher in the MC Results than PCE, because the surrogate PCE model was built based on using the collocation points which only accounts for approximate 90% of the parameter range. The results demonstrate that PCE can be considered as a reliable alternative of MC for quantifying SWAT's parameter uncertainty. By building an efficient surrogate of SWAT, PCE can

generate uncertainty analysis results similar to MC while reducing the required simulation number from thousands to 243. However, it should be noted that the PCE approach can only be used when observations are known to be able to find the coefficients in the surrogate model. So, the current PCE cannot be used for generating forecasts under uncertainties. A machine learning algorithm was introduced to extend the PCE capabilities in probabilistic flow forecasting.

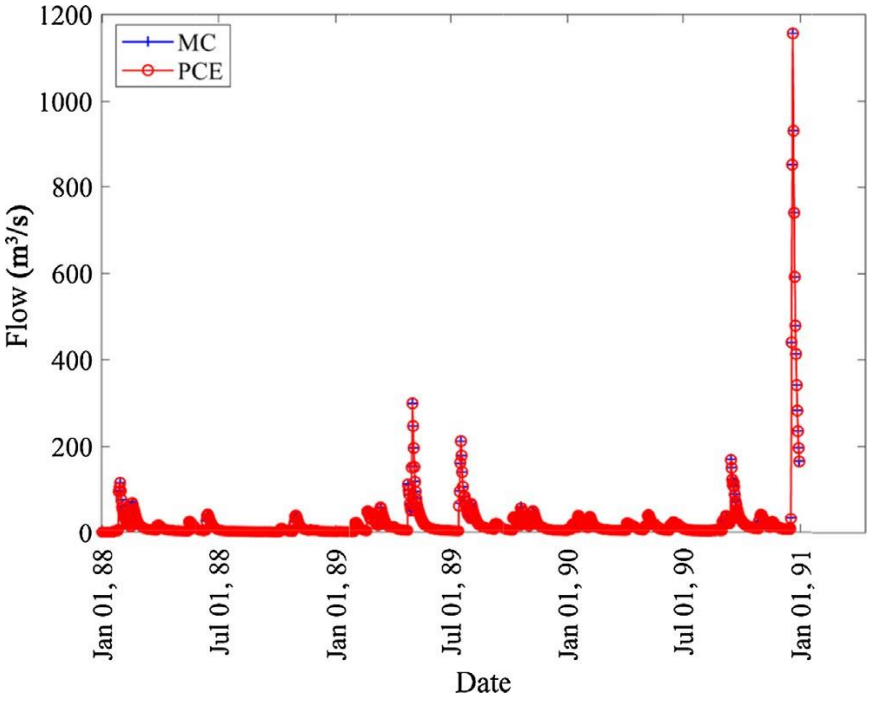

Figure 3-5: Comparison of mean flow time series generated by MC and PCE for the calibration period.

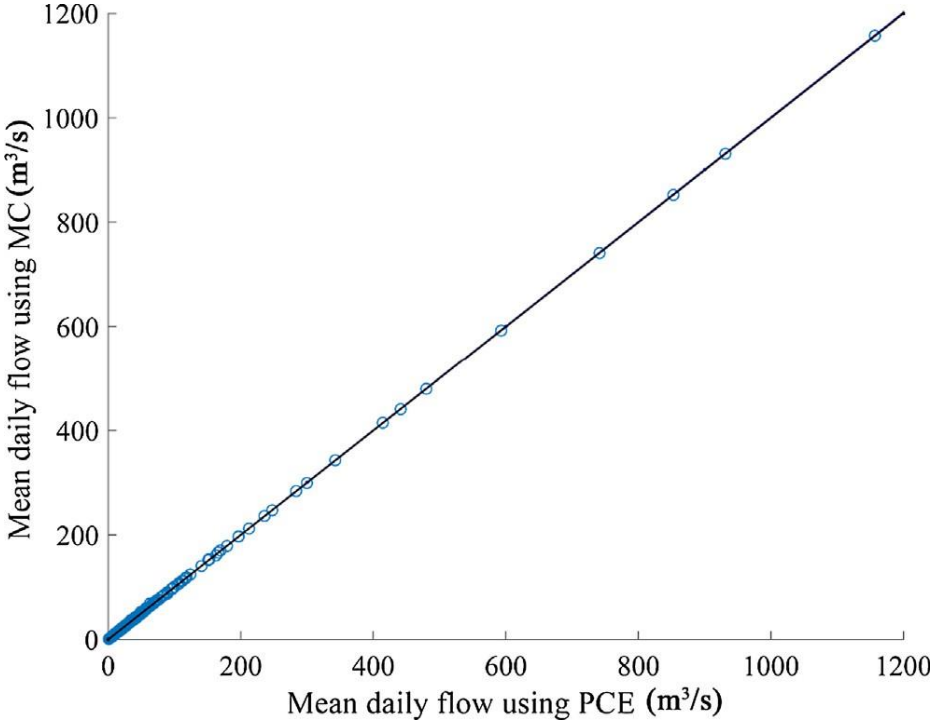

Figure 3-6: Scatter plot of MC and PCE mean flow for the calibration period.

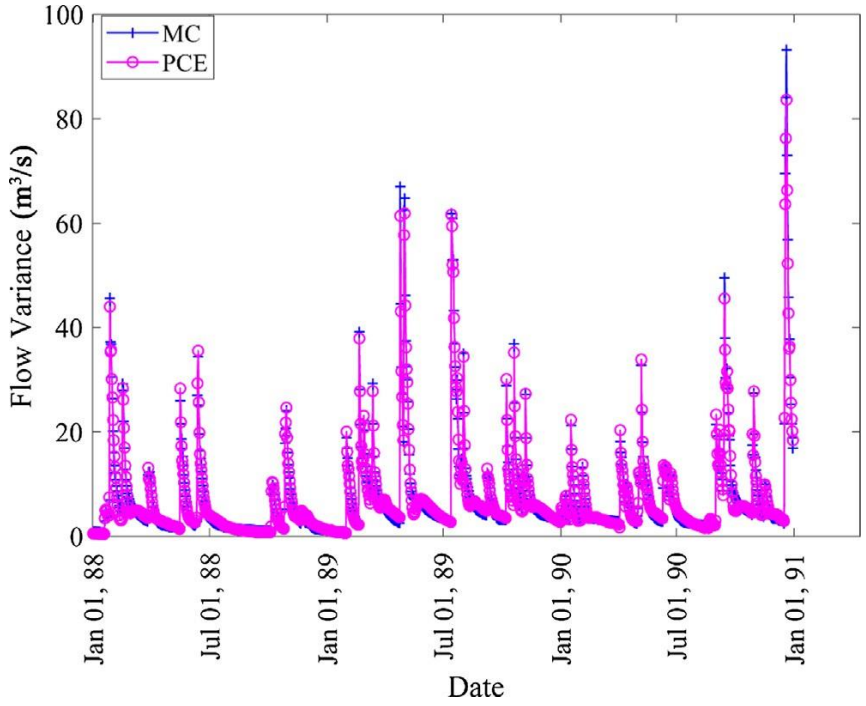

Figure 3-7: Time series of MC and PCE flow variation during the calibration period.

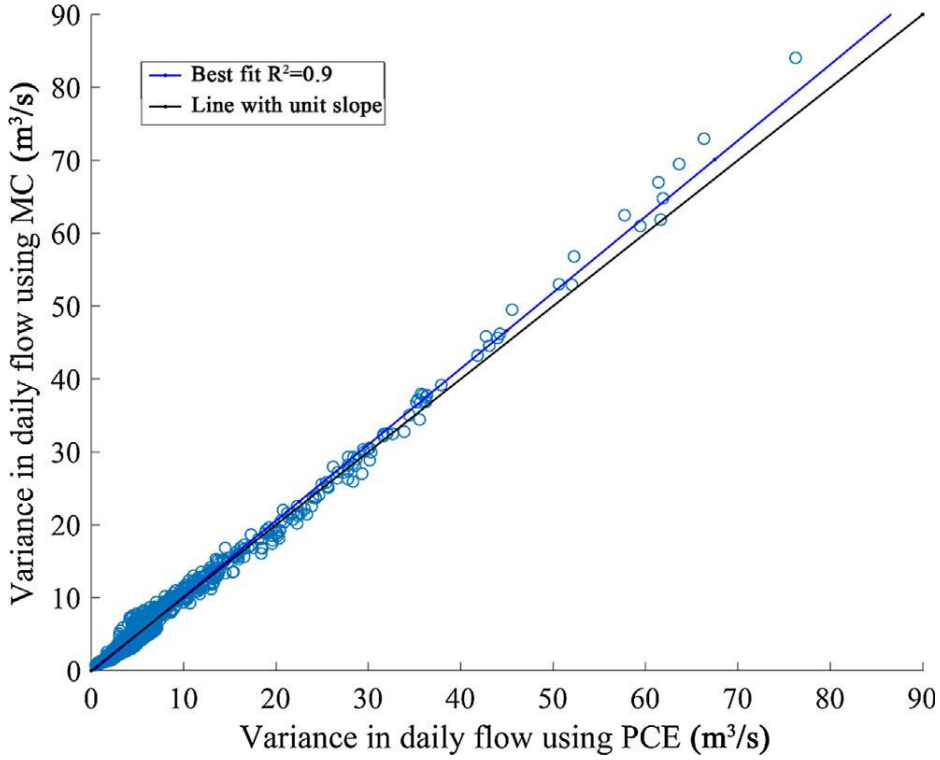

Figure 3-8: Scatter plot of MC and PCE flow variation for the calibration period.

3.5.3. EXPLORING PCE'S FORECASTING CAPABILITY

To enable PCE for generating flow forecasts under parameter uncertainties, ANN was used to find the relationships between metrological conditions and the corresponding PCE coefficients. The ANN model was trained several times to ensure the stability of the results. The mean square error for the model on average was 0.01 with a standard deviation of 0.08 and average coefficient of determination (R^2) value of 0.90 for the 21 coefficients, which indicates that the model is reliable. Once the model was validated, it could be used to find the coefficient in the forecasting period (1992-1997).

To perform uncertainty analysis for the forecasting period (1992-1997) using the proposed PCE-ANN framework, SWAT was first run for 243 times at the collocation point combinations for the period 1989-1991. Then, the ANN model was built to estimate PCE coefficients in the forecasting period. Finally, the uncertainties associated with the forecasted flow can be decoded using a composition of independents centered normalized Gaussian random variables presented by Hermite polynomial. For comparison purposes, a MC simulation of SWAT was conducted with a range of values for each of the five sensitive parameters, and randomly selected values for the normal distribution were plugged in the surrogate PCE model to generate samples of the PCE output. From the 10,000 runs of the MC simulation and the surrogate PCE-ANN results, mean flow, variance, as well as the 25th and 75th percentiles, were calculated. The mean flow predicted by both PCE-ANN and MC is shown in Figure 3-9. Although the difference between PCE-ANN and MC during the forecasting period is slightly higher than that during the historical period, the two mean flow time series share a very similar pattern. There is only a small overestimation for a few values in 1992, 1995 and 1996. This overestimation may be a result of the error propagation, which started during the process of PCE coefficient estimation based on ANN. Figure 3-10 shows the linear relationship between PCE-ANN and MC mean values. The correlation coefficient R^2 is 0.84, which indicates that there is a good fit between the PCE-ANN results and the MC results with only some minor deviations. The slope of the best fit line is slightly less than 1, which implies an overall

overestimation by the PCE-ANN framework compared to MC. When the mean flow is compared with observation, MC has an NSE value of 0.52, while PCE-ANN has an NSE value of 0.55. This indicates that both MC and PCE-ANN performed well for flow forecasting. The variability of flow generated from both PCE-ANN and MC is represented as the standard deviation in Figure 3-11. The variation is higher in MC than PCE-ANN most of the time which agrees with the deduction made in the historical period. It is worth mentioning that the PCE-ANN output has a higher variation for the peak flow in July 1997. Figure 3-12 shows the linear relationship between the standard deviation values of PCE-ANN and MC. Figure 3-12 shows that the standard deviation is similar at low uncertainty conditions, whereas MC variation is higher at high uncertainty conditions. In general, the slope of best fit is greater than one, which indicated that the MC output has an overall higher variance as apprehended from the previous section. Figure 3-12 also shows that the standard deviation has an acceptable fit with a R^2 value of 0.65. Figure 3-13 shows more details for uncertainty results by presenting the 25^{th} and 75^{th} percentiles from both methods. To further compare the PCE-ANN and MC results, the distributions of flow on four selected days are presented in Figure 3-14. The histograms show that the shapes of probability distributions generated by MC and PCE-ANN are very similar. The histogram of MC shows more variance than PCE, which strengthens the other analysis conducted from previous figures. Based on a thorough comparison, it can be concluded that PCE-ANN is able to provide

probabilities forecasts close to MC's with much less computational time and resources.

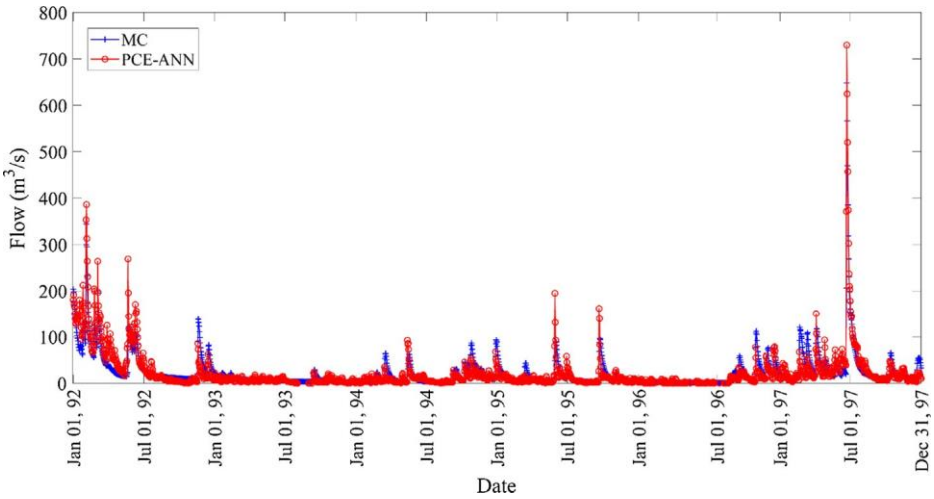

Figure 3-9: Comparison of mean flow time series generated by MC and PCE for the Validation period.

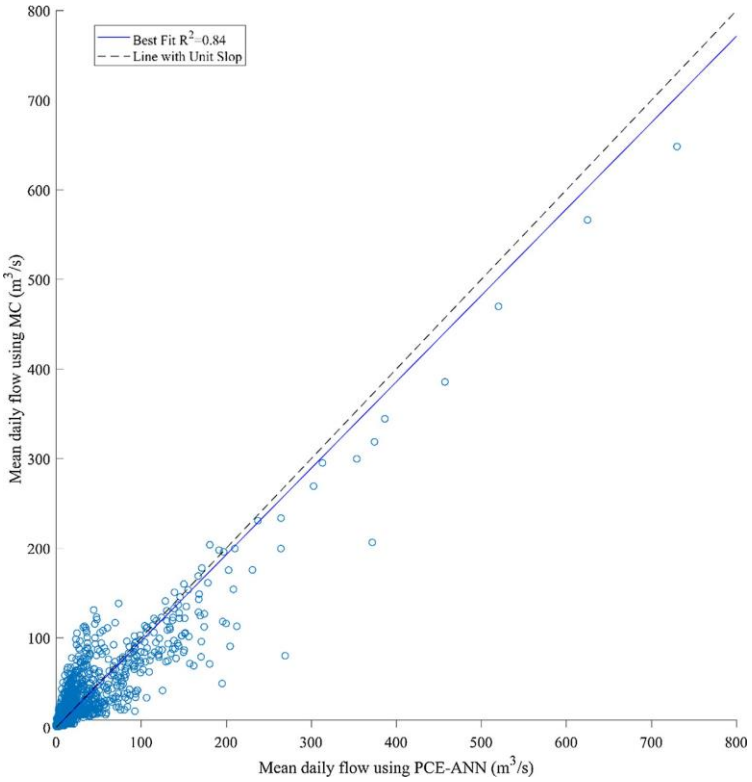

Figure 3-10: Scatter plot of MC and PCE mean flow for the validation period.

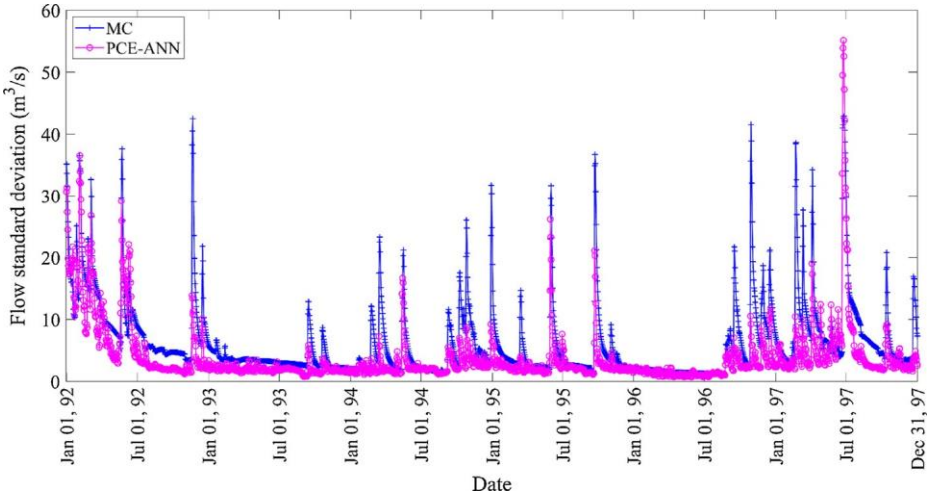

Figure 3-11: Time series of MC and PCE flow variation during the validation period.

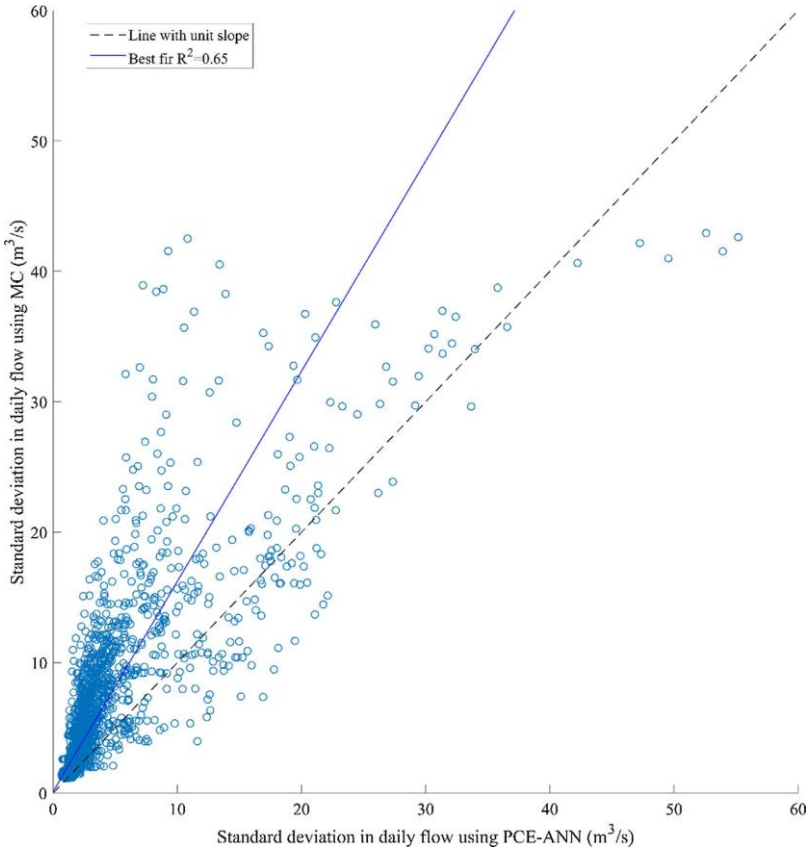

Figure 3-12: Scatter plot of MC and PCE flow standard deviation for the validation period.

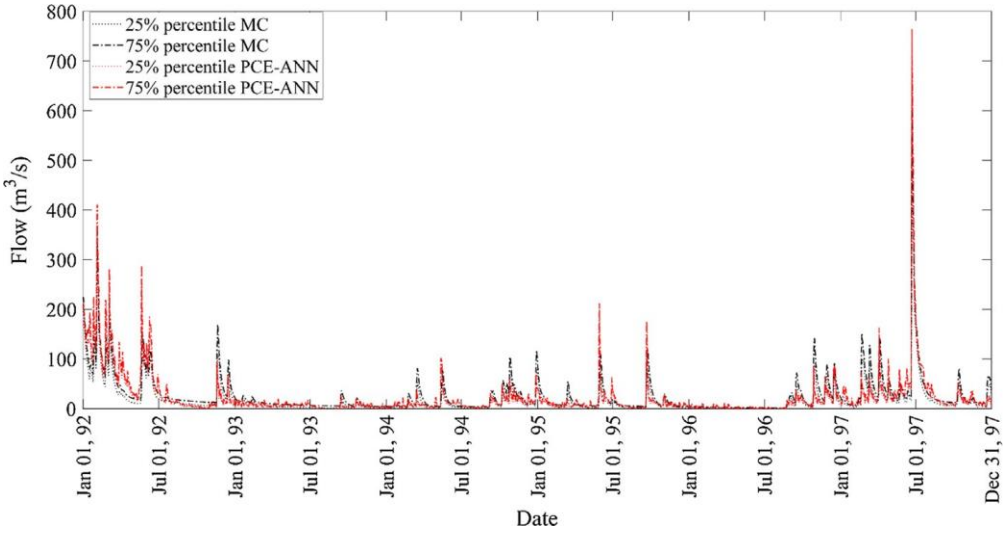

Figure 3-13: Time series of MC and PCE flow percentiles during the validation period.

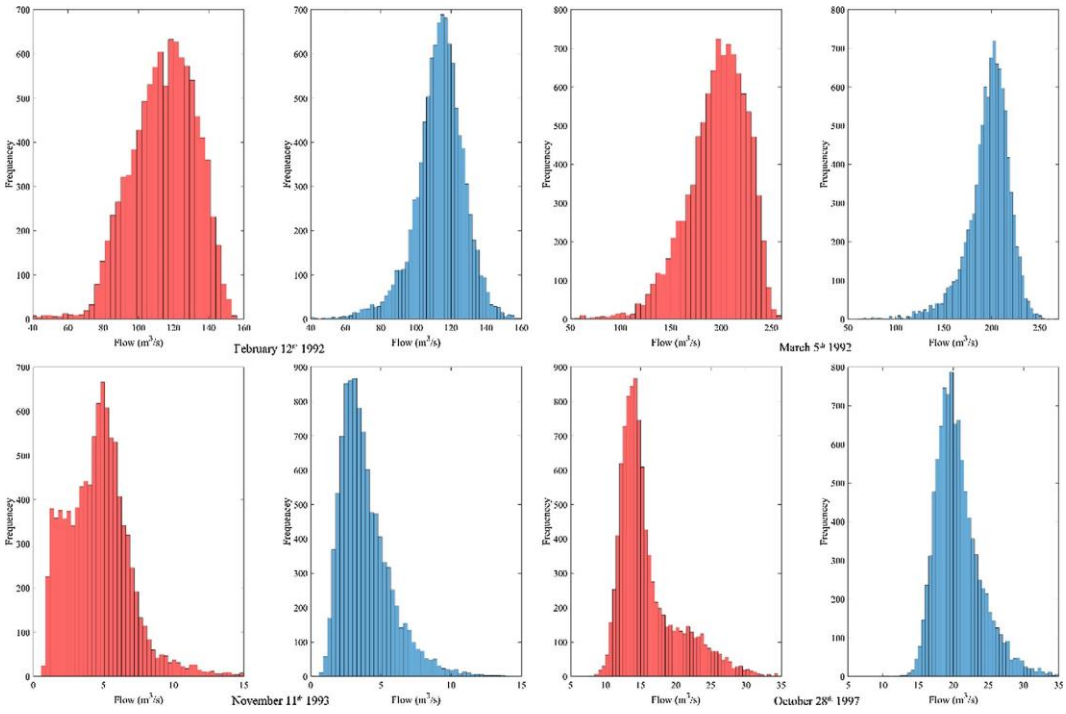

Figure 3-14: Examples of flow histograms. Red: MC output, Blue: PCE-ANN output.

3.6. DISCUSSION

The use of PCE to build a surrogate model for the semi-distributed SWAT model was successful. The PCE produced very similar uncertainty analysis results during the historical period as the MC method. This proves that PCE can be used to simplify the SWAT model (semi-distributed hydrological model) to a surrogate one. To enable PCE for probabilistic forecasting, a machine learning algorithm was integrated to estimate PCE coefficients for the forecasting period, where flow observation data is not available. ANN succeeded to mimic the relationship between weather input and PCE coefficients. The provided PCE-ANN method generated similar probabilistic outputs compared to MC simulation, in terms of mean and variation (with R^2 being 0.84 and 0.64, respectively). Based on previous studies, the R^2 value of 0.84 for mean flow can be considered a good fit. For example, Thavhana et al. (2018) obtained an R^2 of 0.65 for SWAT calibration and 0.523 for validation, and Li et al. (2010) obtained an R^2 of 0.70 for daily streamflow forecasting using SWAT and 0.86 for monthly forecasting. In terms of computation efficiency, the MC simulation took approximately 50 continuous hours, while it took only 30 minutes to build the PCE. Moreover, it took PCE-ANN several minutes to generate probabilistic flow forecasts, while MC required another 50 hours to rerun the model for the future period. These results demonstrate that the use of PCE could save computation time and resources as running a surrogate model (PCE) thousands of realizations will take only a couple of minutes.

ANN is a widely used machine learning method, and its performance is very satisfactory for this particular case study. This is why only ANN was tested and integrated into the PCE framework in this study. Nevertheless, other statistical or machine learning methods can be further investigated for future studies. The selection of candidate method depends on the quality and quantity of input data, as well as the ability of the method itself. Ideally, a candidate method should have a structure well suited for multi-target prediction, in order to support the estimation of multiple PCE coefficients (Ibrahim and Karakurt, 2013; Mosavi et al., 2018).

One limitation of PCE is that the uncertain model parameters are assumed independent. PCE is not applicable if the model has significant dependency between parameters. There are a few studies that attempted to address the issue of correlated parameters; however, only the parameter correlations in simple models can be tackled in those existing studies (Paulson et al., 2017; Rahman, 2018). There is no existing PCE framework that can quantify the effects of correlated parameters on the output of complex hydrological models. Further research is required to enable PCE to support uncertainty analysis for hydrological models with dependent parameters. Another assumption in this study is that the parameters all have a uniform distribution defined by its range, whereas in reality, the distribution varies depending on the catchment morphology. In future research, the use of the traditional GLUE method for finding more accurate parameter distributions can be further investigated. (Yang et al., 2008). Introducing GLUE into the PCE framework will increase the required computational time; however, it would still

be worth investigating as GLUE have the potential to address the issue of parameter dependency as well.

3.7. CONCLUSIONS

In this study, a new PCE-ANN approach for quantifying parameter uncertainties and generating probabilistic flow forecasts was developed. The proposed method can quantitatively describe the propagation of parameter uncertainty in a modeling system and was applied to the hydrological modeling of the Guadalupe basin in Texas, US. Previously, PCE has only been used to build surrogates for simple, lumped hydrological models such as HYMOD. In this study, PCE was used for a more complex, semi-distributed model SWAT for the first time.

The traditional PCE framework requires observed flow as input to assess the propagation of parameter uncertainties and thus could not be used to analyze the uncertainties in future time series. By innovatively integrating data-driven techniques into the traditional PCE framework, the proposed PCE-ANN approach enables the PCE to generate probabilistic flow forecasts. The probabilistic forecasting results of PCE-ANN were compared to those of MC simulation. The results demonstrated that PCE-ANN was able to provide probabilities forecasts close to MC's with much less computational time and resources.

PCE-ANN's advantage in terms of computational efficiency is expected to be more significantly beneficial as the model complexity increases. In the future, the PCE-ANN should be tested for other complex hydrological models, as well as for more watersheds of different sizes and types. Other advanced data-driven

techniques can be further investigated to improve the prediction of PCE coefficients. Also, one of PCE's fundamental assumption is that the uncertain parameters are independent. Further research is required to address the common parameter dependency issue in hydrological modeling.

3.8. ACKNOWLEDGEMENT

This study was supported by Natural Sciences and Engineering Research Council of Canada (NSERC). We would like to thank Dr. Wendy Huang for her comments on the original version of this manuscript. We are also grateful to the editor and anonymous reviewers for comments that greatly improved the manuscript.

3.9. REFERENCES

Arabi, M., Frankenberger, J.R., Engel, B.A., Arnold, J.G., 2008. Representation of agricultural conservation practices with SWAT 3055, 3042–3055. https://doi.org/10.1002/hyp

Arnold, J.G., Moriasi, D.N., Gassman, P.W., White, M.J., 2012. SWAT: Model use, calibration, and validation. Am. Soc. Agric. Biol. Eng. 55, 1491–1508.

Dagnew, A., Philip, T., Silvia, W.G., 2016. Modeling Agricultural Watersheds with the Soil and Water Assessment Tool (SWAT): Calibration and Validation with a Novel Procedure for Spatially Explicit HRUs. Environ. Manage. 57, 894–911. https://doi.org/10.1007/s00267-015-0636-4

Debele, B., Srinivasan, R., Parlange, J., 2008. Coupling upland watershed and downstream waterbody hydrodynamic and water quality models (SWAT and CE-QUAL-W2) for better water resources management in complex river basins 135–153. https://doi.org/10.1007/s10666-006-9075-1

Deman, G., Konakli, K., Sudret, B., Kerrou, J., Perrochet, P., Benabderrahmane, H., 2016. Using sparse polynomial chaos expansions for the global sensitivity analysis of groundwater lifetime expectancy in a multi-layered hydrogeological model. Reliab. Eng. Syst. Saf. 147, 156–169. https://doi.org/10.1016/j.ress.2015.11.005

Eckhardt, K., Haverkamp, S., Fohrer, N., Frede, H., 2002. SWAT-G, a version of SWAT99. 2 modified for application to low mountain range catchments 27, 641–644.

Fan, Y.R., Huang, W., Huang, G.H., Huang, K., Zhou, X., 2014. A PCM-based stochastic hydrological model for uncertainty quantification in watershed systems. Stoch. Environ. Res. Risk Assess. 29, 915–927. https://doi.org/10.1007/s00477-014-0954-8

Fu, C., James, A.L., Yao, H., 2014. SWAT-CS: Revision and testing of SWAT for Canadian Shield catchments. J. Hydrol. 511, 719–735. https://doi.org/10.1016/j.jhydrol.2014.02.023

Hallouz, F., Meddi, M., Mahé, G., Alirahmani, S., 2018. Modeling of discharge and sediment transport through the SWAT model in the basin of Harraza (Northwest of Algeria). Water Sci. 32, 79–88. https://doi.org/10.1016/j.wsj.2017.12.004

Hosder, S., 2010. Point-Collocation Nonintrusive Polynomial Chaos Method for Stochastic Computational Fluid Dynamics 48, 2721–2730. https://doi.org/10.2514/1.39389

Ibrahim, H., Karakurt, O., 2013. Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms. J. Hydrol. 477, 119–128. https://doi.org/10.1016/j.jhydrol.2012.11.015

Iudicello, J.J., Chin, D.A., Asce, F., 2013. Multimodel, Multiple Watershed Examination of In-Stream Bacteria Modeling 139, 719–727. https://doi.org/10.1061/(ASCE)EE.1943-7870.0000670.

Khatun, S., Sahana, M., Jain, S.K., Jain, N., 2018. Simulation of surface runoff using semi distributed hydrological model for a part of Satluj Basin: parameterization and global sensitivity analysis using SWAT CUP. Model. Earth Syst. Environ. 4, 1111–1124. https://doi.org/10.1007/s40808-018-0474-5

Laloy, E., Rogiers, B., Vrugt, J.A., Mallants, D., Jacques, D., 2013. Efficient posterior exploration of a high-dimensional groundwater model from two-stage Markov chain Monte Carlo simulation and polynomial chaos expansion 49, 2664–2682. https://doi.org/10.1002/wrcr.20226

Li, H., Zhang, D., 2007. Probabilistic collocation method for flow in porous media: Comparisons with other stochastic methods 43, 1–13. https://doi.org/10.1029/2006WR005673

Li, Z., Shao, Q., Xu, Z., Cai, X., 2010. Analysis of parameter uncertainty in semi-distributed hydrological models using bootstrap method: A case study of SWAT model applied to Yingluoxia watershed in northwest China. J. Hydrol. 385, 76–83. https://doi.org/10.1016/j.jhydrol.2010.01.025

Mosavi, A., Ozturk, P., Chau, K., 2018. Flood Prediction Using Machine Learning Models: Literature Review. Water 10, 1536–1576. https://doi.org/10.3390/w10111536

Najm, H.N., 2009. Uncertainty Quantification and Polynomial Chaos Techniques in Computational Fluid Dynamics. SIAM J. SCI. Comput. 37, A2535–A2557. https://doi.org/10.1146/annurev.fluid.010908.165248

Olivera, F., Valenzuela, M., Srinivasan, R., Choi, J., 2006. A RC GIS-SWAT: A GEODATA MODEL AND GIS INTERFACE FOR SWAT 1 77845.

Paffrath, M., Wever, U., 2007. Adapted polynomial chaos expansion for failure detection 226, 263–281. https://doi.org/10.1016/j.jcp.2007.04.011

Paulson, J.A., Buehler, E.A., Mesbah, A., Joel, A., Paulson, A., Edward, A., Buehler, A., 2017. ScienceDirect Arbitrary Polynomial Chaos for Arbitrary Polynomial Chaos for Arbitrary Polynomial of Chaos for Uncertainty Propagation Correlated Uncertainty Propagation of Correlated Uncertainty Propagation of Correlated Random Variables in Dynamic Systems Random Variables in Dynamic Random Variables in Dynamic Systems. IFAC-PapersOnLine 50, 3548–3553. https://doi.org/10.1016/j.ifacol.2017.08.954

Rahman, S., 2018. A polynomial chaos expansion in dependent random variables ☆. J. Math. Anal. Appl. 464, 749–775. https://doi.org/10.1016/j.jmaa.2018.04.032

Rupert, C.P., Miller, C.T., 2007. An analysis of polynomial chaos approximations for modeling single-fluid-phase flow in porous medium systems 226, 2175–2205. https://doi.org/10.1016/j.jcp.2007.07.001

Shen, Z.Y., Chen, L., Liao, Q., Liu, R.M., Huang, Q., 2013. A comprehensive study of the effect of GIS data on hydrology and non-point source pollution modeling. Agric. Water Manag. 118, 93–102. https://doi.org/10.1016/j.agwat.2012.12.005

Singh, V., Bankar, N., Salunkhe, S.S., Bera, A.K., Sharma, J.R., 2013. Hydrological stream flow modelling on Tungabhadra catchment: parameterization and uncertainty analysis using SWAT CUP. Curr. Sci. 104, 1187–1199.

Stavropoulou, F., Muller, J., 2015. PARAMETRIZATION OF RANDOM VECTORS IN POLYNOMIAL CHAOS EXPANSIONS VIA OPTIMAL TRANSPORTATION. SIAM J. Sci. Comput. 37, 2535–2557. https://doi.org/10.1137/130949063

Suliman, A.H.A., Jajarmizadeh, M., Harun, S., Zaurah, I., Darus, M., 2015. Comparison of Semi-Distributed, GIS-Based Hydrological Models for the Prediction of Streamflow in a Large Catchment 3095–3110. https://doi.org/10.1007/s11269-015-0984-0

Tagade, P.M., Choi, H., 2014. A Generalized Polynomial Chaos-Based Method for Efficient Bayesian 22, 602–624.

Thavhana, M.P., Savage, M.J., Moeletsi, M.E., 2018. SWAT model uncertainty analysis, calibration, and validation for runoff simulation in the Luvuvhu River catchment, South Africa. Phys. Chem. Earth 105, 115–124. https://doi.org/10.1016/j.pce.2018.03.012

Villegas, M., Augustin, F., Gilg, A., Hmaidi, A., Wever, U., 2012. Application of the Polynomial Chaos Expansion to the simulation of chemical reactors with uncertainties. Math. Comput. Simul. 82, 805–817. https://doi.org/10.1016/j.matcom.2011.12.001

Wang, S., Huang, G.H., Huang, W., Fan, Y.R., Li, Z., 2015. A fractional factorial probabilistic collocation method for uncertainty propagation of hydrologic model parameters in a reduced dimensional space. J. Hydrol. 529, 1129–1146. https://doi.org/10.1016/j.jhydrol.2015.09.034

White, M., Gambone, M., Yen, H., Daggupati, P., Bieger, K., Deb, D., Arnold, J., 2016. DEVELOPMENT OF A CROPLAND MANAGEMENT DATASET 52, 269–274. https://doi.org/10.1111/1752-1688.12384

White, M.J., Gambone, M., Haney, E., Arnold, J., Gao, J., 2017. Development of a Station Based Climate Database for SWAT and APEX Assessments in the US 1–9. https://doi.org/10.3390/w9060437

Wu, X., Zhang, W., Song, S., Ye, Z., 2018. Sparse grid-based polynomial chaos expansion for aerodynamics of an airfoil with uncertainties. Chinese J. Aeronaut. 31, 997–1011. https://doi.org/10.1016/j.cja.2018.03.011

Yang, J., Reichert, P., Abbaspour, K.C., Xia, J., Yang, H., 2008. Comparing uncertainty analysis techniques for a SWAT application to the Chaohe Basin in China. J. Hydrol. 358, 1–23. https://doi.org/10.1016/j.jhydrol.2008.05.012

Yang, J., Reichert, P., Abbaspour, K.C., Yang, H., 2007. Hydrological modelling of the Chaohe Basin in China: Statistical model formulation and Bayesian inference. J. Hydrol. 340, 167–182. https://doi.org/10.1016/j.jhydrol.2007.04.006

Yesuf, H.M., Melesse, A.M., Zeleke, G., Alamirew, T., 2016. Streamflow prediction uncertainty analysis and verification of SWAT model in a tropical watershed. Environ. Earth Sci. 75, 1–16. https://doi.org/10.1007/s12665-016-5636-z

Zhang, D., Chen, X., Yao, H., 2016. Science of the Total Environment SWAT-CS: Enhancing SWAT nitrate module for a Canadian Shield catchment. Sci. Total Environ. 550, 598–610. https://doi.org/10.1016/j.scitotenv.2016.01.109

Zhang, J., Li, Y., Huang, G., Chen, X., Bao, A., 2016. Assessment of parameter uncertainty in hydrological model using a Markov-Chain-Monte-Carlo-based multilevel-factorial-analysis method. J. Hydrol. 538, 471–486. https://doi.org/10.1016/j.jhydrol.2016.04.044

Zhang, X., Srinivasan, R., Bosch, D., 2009. Calibration and uncertainty analysis of the SWAT model using Genetic Algorithms and Bayesian Model Averaging. J. Hydrol. 374, 307–317. https://doi.org/10.1016/j.jhydrol.2009.06.023

Chapter 4

UNCERTAINTY ANALYSIS FOR HYDROLOGICAL MODELS WITH INTERDEPENDENT PARAMETERS: AN IMPROVED POLYNOMIAL CHAOS EXPANSION APPROACH

ABSTRACT

The use of polynomial chaos expansion (PCE) has gained a lot of attention due to its capability to efficiently estimate the effects of parameter uncertainty on model outputs. The traditional PCE technique requires the studied parameters to be independent. In hydrological modeling, although model parameters are often assumed to be independent for simplicity of computation, such an assumption is not always valid. Neglecting parameter correlations could significantly affect the analysis of uncertainty, leading to distorted modeling results. In this study, an improved PCE approach is proposed to address this issue and support the uncertainty analysis for hydrological models with correlated parameters. The proposed approach is based on the integration of principle component analysis (PCA) and PCE, where PCA is used to transform correlated parameters into orthogonal independent components. To demonstrate the applicability of this approach, SWAT model is developed for the Guadalupe River Watershed in Texas, US, and the integrated PCA-PCE framework is used to assess the uncertainty of SWAT's dependent parameters. A traditional Monte-Carlo (MC) simulation is also used to address the uncertainty in the developed SWAT model. The results show

that PCA-PCE could generate similar probabilistic flow results compared to MC, while maintaining a very high computational efficiency. The coefficients of determination (R^2) for the mean and variance are 0.998 and 0.973, respectively, and the computational requirement is reduced by 99% using the developed PCA-PCE approach. It is shown that the PCA-PCE approach is reliable and efficient in assessing the uncertainties in hydrological models with correlated parameters.

Key points:

- An improved PCE approach was proposed to assess the propagation of uncertainties associated with interdependent parameters.
- The proposed method reduces computational requirements for uncertainty analysis by 99% compared to traditional Monte-Carlo simulation.

4.1. INTRODUCTION

Hydrological models are an essential tool for designers and decision makers to accommodate future water demands and assess the associated risks (Halldin, 2005; Karlsson et al., 2016; Lin et al., 2007). In hydrological models, there are parameters and coefficients that need to be tuned in order for the model to accurately represent the watershed characteristics. The exact precise values of these parameters that match the model outputs with field observations are difficult to find, which necessitates the analysis of parameter uncertainty for hydrological modeling.

Before uncertainty analysis, sensitivity analysis can be carried out for a hydrological model to find the most crucial parameters that affect the output responses and to determine the parameter ranges (Study, 2019; H. Wu & Chen, 2015). The uncertainty of the most sensitive parameters can then be analyzed (Almeida, Pereira, & Pinto, 2018). The computational requirement for uncertainty analysis depends on the method chosen, as well as the structure and number of parameters of the hydrological model (Devak & Dhanya, 2017).

There are a number of widely used methods for uncertainty analysis, such as generalized likelihood uncertainty estimation (GLUE), Monte Carlo (MC) simulation, and sequential uncertainty fitting algorithm (SUFI-2) (Tolessa et al., 2015; Y. Wu & Liu, 2012; Xie & Lian, 2013). In previous studies, it is commonly assumed that the uncertain parameters are independent regardless of the uncertainty analysis method used. The assumption of parameter independency is usually made

to overcome computational complexity (Paulson et al., 2017). In fact, in hydrological models there could be many parameters that are heavily dependent on each other (Christiaens & Feyen, 2002; Q. Wu, Liu, Cai, Li, & Jiang, 2017; Yang, Reichert, Abbaspour, Xia, & Yang, 2008a). For example, Yang et al. (2008) performed a GLUE analysis for the soil & water assessment tool (SWAT) model and found that there are considerable correlations between several parameters, such as available water content (Sol_AWC) and curve number (CN). In the past two decades, there have been limited attempts to address the uncertainty of independent parameters (Longland, 2017; F. Wu & Tsang, 2004; Xu, 2013; Xu & Gertner, 2008). One of the most straightforward and popular methods is MC simulation. Given the distributions of and correlations between uncertain parameters, MC simulation allows for the identification of all the possible modeling outcomes of events, making it easy to quantify the effects of uncertain parameters. However, MC is time and resource demanding, especially when the complexity of model structure and the number of uncertain parameters increases (Longland, 2017; F. Wu & Tsang, 2004).

More recently, polynomial chaos expansion (PCE) has been applied as a promising approach to improve the computational efficiency of uncertainty analysis for complex modeling systems. PCE is a non-sampling-based method that builds polynomial chaos surrogates to determine the evolution of parameter uncertainty in a dynamic modeling system. PCE was first introduced to account for only Gaussian random variables (Norbert Wiener, 1938). A generalized polynomial chaos

framework was later proposed to generalize the method to other distributions (Xiu, 2010). It has been proven in many applications that PCE is computationally superior to Monte Carlo simulation. However, the traditional PCE method can only be used to analyze the uncertainty of independent parameters (Fan, Huang, Huang, Huang, & Zhou, 2014; Wang, Huang, Huang, Fan, & Li, 2015). To overcome this drawback, a limited number of studies have been carried out to account for parameter dependency in PCE. Navarro et al. (2014) introduced an arbitrary PCE approach that can handle dependent parameters using Gram-Schmidt orthogonalization transformation. Rahman (2018) explored the use of whitening transformation for generating orthonormal polynomials to be used in analytical solutions. Both studies used a transformation method to find orthogonal independent variables, which successfully solved the parameter dependency problem for PCE; however, the transformation method could only be used for simple equation-based models. For more complex modeling systems like most hydrological models, there is currently no existing solution to this problem. An improved PCE approach that could tackle the uncertainty associated with dependent parameters in an efficient manner is needed.

The objective of this study is to investigate how the traditional PCE framework could be improved to assess the uncertainty of dependent parameters in hydrological models. An improved approach based on the integration of principle component analysis (PCA) and PCE is developed and tested. PCA and the Johnson transformation are used to decouple correlated hydrological model parameters and

obtain orthonormal independent variables, which are further used for establishing PCEs for uncertainty analysis. A SWAT model developed for a watershed in the Guadalupe River basin in Texas, USA is used as case study to demonstrate the applicability of the proposed method. In addition to the uncertainty analysis based on the proposed PCA-PCE approach, a parallel analysis is conducted using MC simulation. The results are compared to show the advantage of the proposed approach. This paper is divided into five sections. Section 4.2 describes PCE, SWAT, PCA, Johnson transformations, and the integrated PCA-PCE framework. Section 4.3 describes the study area and data used in this paper. Section 4.4 discusses the results of this study. Finally, Section 4.5 summarizes the conclusions of this study.

4.2. METHODOLOGY

4.2.1. POLYNOMIAL CHAOS EXPANSION

PCE is an efficient tool for assessing the effects of parameter uncertainty on the output of a simulation model. Traditionally, such effects can be measured using the MC method, which requires inputting random parameter values into the model for a large number of simulations runs. PCE analyzes the propagation of the random behavior in parameters through hypergeometric orthogonal polynomials in the Askey scheme (Martinez, Crestaux, & Le Maitre, 2009; Xiu & Karniadakis, 2003). The general PCE equation can be written as follows:

$$y = a_0 + \sum_{i=1}^{n} a_i P_1(\zeta_i) + \sum_{i=1}^{n}\sum_{j=1}^{i} a_{ij} P_2(\zeta_i, \zeta_j) + \sum_{i=1}^{n}\sum_{j=1}^{i}\sum_{k=1}^{j} a_{ijk} P_3(\zeta_i, \zeta_j, \zeta_k) + ... \quad (4\text{-}1)$$

where y is the model output, $\zeta_i, \zeta_j,$ and ζ_k are the random variables, $P_n(\zeta_1, \zeta_2, ...\zeta_n)$ is the polynomial chaos of order *p*, *n* is the number of PCE variables (i.e., uncertain parameters of the simulation model), and a_0, a_i, $a_{ij,}$ and a_{ijk} are PCE coefficients.

Although there are a number of different Askey-scheme hypergeometric orthogonal polynomials, the Hermite polynomial first used in PCE by Wiener (1938) is the most suitable and generalized (Fan et al., 2014; Fenfen, Shishi, & Ying, 2014; Wang et al., 2015). Hermite polynomials converge quickly and represent the randomness in Gaussian independent variables well (Funahashi & Kijima, 2012). The Hermite polynomial is expressed as follows:

$$H_n(\zeta_i, \zeta_j, ...) = e^{\frac{\zeta^T\zeta}{2}} (-1)^n \frac{\partial^n}{\partial\zeta_i \partial\zeta_j .....} e^{\frac{\zeta^T\zeta}{2}} \quad (4\text{-}2)$$

The terms of expansion from Hermite polynomials are determined based on the order of PCE and the number of uncertain variables. In the Weiner PCE, the random variables ($\zeta_i, \zeta_j, \zeta_k \ldots$) are independent and have zero mean and unit variance. The dimension of the Hermite polynomial is determined based on *T*. For the one-dimensional Hermite polynomials of orders 0, 1, 2, and 3 are given as below:

$$H_0 = 1 \quad (4\text{-}3)$$

$$H_1 = \zeta \tag{4-4}$$

$$H_2 = \zeta^2 - 1 \tag{4-5}$$

$$H_3 = \zeta^3 - 3\zeta \tag{4-6}$$

By substituting the polynomials in Equation 4-1 with one-dimensional Hermite polynomials with regard to standard normal random variables, generalized Weiner PCEs can be established.

Subsequently, PCE coefficients (a_0, a_i, $a_{ij,}$ and $a_{ijk}\ldots$) can be determined using a probabilistic collocation method (PCM) (Fan et al., 2014). In PCM, a set of collocation points are first derived from the roots of the Hermite polynomial of one order higher than the chosen order. In this study, the second-order Hermite polynomial is used to establish PCEs. Thus, the roots of the third-order Hermite polynomial (6) are obtained as the collocation points. By running the simulation model at one possible combination of the three collocation points ($-\sqrt{3}$, 0, and $\sqrt{3}$) for each parameter, the corresponding output y can be obtained. The equations of all possible combinations of the collocation points form a system of equations, which can then be solved for the unknown PCE coefficients. Thus, a surrogate model consisting of one PCE equation for each time step can be established to accelerate the quantification of output uncertainty. In the existing PCE approach, all uncertain parameters are assumed to be independent.

4.2.2. SWAT

The SWAT model is a semi-distributed model. SWAT has been widely used for hydrological and water quality modeling as it can simulate hydrological processes, as well as the transport of sediment, nutrients and pesticide (Jeffrey G Arnold, Moriasi, Gassman, & White, 2012). The hydrological simulation in SWAT involves over thirty parameters. These parameters can be estimated using watershed data such as Digital Elevation Models (DEM), land use, and soil characteristics. The exact parameter values are challenging to determine, but parameter ranges can be obtained through sensitivity and uncertainty analysis based on calibration with observation data.

The SWAT-CUP tool is an automatic calibration, validation, and uncertainty analysis engine for SWAT. There are several algorithms that are embedded within SWAT-CUP, including SUFI-2, GLUE, MC, and parameter solution (Para-Sol) (Almeida et al., 2018; Zhang, Jin, He, & Zhang, 2016). Regardless of which algorithm one chooses, the parameters related to the dominant processes of interest should be calibrated, and their ranges and distributions should be defined. Typically, a termination process defined by the number of iterations or a threshold of the error function is used by the automatic calibration algorithm. Based on the calibration results, a global sensitivity analysis can be carried out and sensitive parameters can be identified. This process is adopted in this study to identify the most sensitive parameters for uncertainty analysis.

4.2.3. PRINCIPLE COMPONENT ANALYSIS

PCA is a widely used statistical procedure that uses orthogonal transformation to present the variability of a large number of correlated variables using a smaller number of uncorrelated variables. The produced uncorrelated variables are orthogonal and are called principal components (PCs) (Ma & Dai, 2011; Wold, Esbensen, & Geladi, 1987). Each of the PCs produced is a linear combination of the original variables. To avoid bias in the PC, all the variables have to be scaled before transformation. The first PC accounts for the most variability in the data. Then each PC added accounts for additional information that is not explained in previous components. If the number of PCs are the same as the number of variables, then all the information in the original data is explained with uncorrelated components. PCA is widely used to reduce number of variables and thus to reduce the computational requirements for handling the data. In this study, PCA is incorporated into the PCE framework to tackle parameter correlations and generate dependent variables for PCE analysis.

There are different methods to perform PCA. In this study, the eigenvector technique is used. First, all the variables in the original data set are normalized. Second, the covariance matrix is calculated for the normalized centered data. The covariance matrix diagonal contains the variance of the variables, whereas the off-diagonal elements are covariances between two variables. The covariance matrix is a generalized and unnormalized version of variable correlations. Third, the eigen-decomposition of the covariance matrix is calculated. The results of the eigen-decomposition consist of eigenvalues and eigenvectors. The eigen-values represent

the magnitudes for the dominant directions. The eigenvalue with the highest values represents the first direction for the most important PC. The eigenvectors are then sorted based on the eigenvalues in descending order to provide their rank as well as the cut off threshold. If the eigenvalues are the same value, then this means that the original data are uncorrelated and do not require PCA transformation. Finally, the original variables are projected on the chosen eigenvectors. To produce the PCs, the original data are multiplied with the transpose of each eigenvector in the ranking order. The produced PCs are orthogonal and uncorrelated. It has been found in many previous studies such as (Ma & Dai, 2011; Wold et al., 1987) that most variances in a dataset can be explained in the first few components.

4.2.4. TRANSFORMATION TO STANDARD NORMAL

The uncertain variables in the PCM, to obtain PCE coefficients, are assumed to be independent and normally distributed. Although the projected PCs from PCA are independent, they rarely follow normal distributions. Therefore, the projected PCs have to be transformed using a generalized and reliable technique.

There are several commonly used transformation functions such as Box-Cox, power transformations, exponential, and logarithmic (Mach, Thuring, & Šámal, 2006; Mateu, 1997). The Box-Cox function is a type of power transformation which works with many distributions as the lambda value varies from -5 to 5 which allows it to cover different forms of equations. The power transformation is a simpler

method than Box-Cox, as it is not as comprehensive, but it still gives commendable results. The exponential transformation is widely used with data that have a near lognormal distribution. The logarithmic transformation is only applicable to exponential and lognormal distributions. In this study, there are certain requirements for the transformation function: 1) it should be nonlinear to overcome the asymmetry in the data and transform it towards normality; 2) it should be monotonous to maintain the order of the data before and after the transformation process; 3) it should be reversible to convert the results back to its original distribution; and 4) there should not be any limiting assumption for the distribution of the original data, since the PC data do not always fit to a specific distribution.

The Johnson distribution (Johnson, 1949) is a transformation function that meets all of the abovementioned requirements. It is a family of three distributions that can fit any data, which makes it a generalized transformation method (Yu, 1994). The Johnson distribution has four parameters, including two shape parameters (γ and δ), a scale parameter (λ), and a location parameter (ζ). δ and λ are always positive. The three distributions of the Johnson family are S_B, S_L, and S_U, where B stands for bounded data, L stands for bounded from below or lognormal, and U stands for unbounded. Table 4-1 shows the equations to transform the three distributions to standard normal. In this study, the S_B family is used as the data is bounded. For a domain of $\zeta \leq x \leq \zeta + \lambda$, the general probability density function is shown as below:

$$f(x) = \frac{\delta}{\lambda\sqrt{2\pi}z(1-z)} \exp\left(-\frac{1}{2}\left(\gamma + \delta\, ln\left(\frac{z}{1-z}\right)\right)^2\right) \quad (4\text{-}7)$$

where $z = \frac{x-\zeta}{\lambda}$.

Table 4-1: Johnson Distributions.

Family	Transformation equation	Parameter Conditions	X condition
S_B	$Z = \gamma + \delta * \ln\left(\frac{x-\zeta}{\lambda+\zeta-x}\right)$	$\delta, \lambda > 0, -\infty < \gamma < \infty, -\infty < \zeta < \infty$	$\zeta < x < \zeta + \lambda$
S_L	$Z = \gamma + \delta * \ln(x-\zeta)$	$\delta > 0, -\infty < \gamma < \infty, -\infty < \zeta < \infty$	$x > \zeta$
S_U	$Z = \gamma + \delta * \sinh^{-1}\left(\frac{x-\zeta}{\lambda}\right)$	$\delta, \lambda > 0, -\infty < \gamma < \infty, -\infty < \zeta < \infty$	$-\infty < x < \infty$

The transformation process is summarized as follows: 1) obtain the data of independent PC variables from the PCA of interdependent SWAT parameters; 2) transform the PC variables to independent normally distributed variables using the Johnson S_B distribution; 3) test the normality of the transformed data. The maximum likelihood method is used to fit the data to the Johnson distribution.

4.2.5. PCA-PCE FRAMEWORK

The traditional PCE can only deal with independent parameters. First, the independent distributions of the parameters are transformed to independent standard normal distributions, and a number of the collocation points are extracted from each of the independent standard normal distributions. Second, the hydrological model is run at all combinations of the collocation points to obtain the corresponding model outputs. Third, the PCE coefficients at each simulation time

step are determined by solving a set of linear equations based on the collocation points and the obtained model output. Finally, once the coefficients are found, the PCE equation can be used as a surrogate model to assess the effects of independent random parameters on the model output.

Since the traditional PCE cannot handle dependent parameters and parameter dependency does exists in many hydrological models, the PCA-PCE framework is proposed to address this issue. Figure 4-1 compares the difference between the traditional PCE and the proposed PCA-PCE framework. The essence of PCA-PCE is to decouple the correlations among dependent parameters and generate independent standard normal distributions for further development of PCE. In the PCA-PCE framework, the parameter correlations are defined either by assumptions based on previous studies or by generating correlated samples through an uncertainty analysis technique such as GLUE. A smaller number of steps are conducted prior to the selection of collocation points. First, a posterior distribution for each parameter is generated based on the defined correlations. Second, PCA is used to obtain two or three projected PCs with orthogonal distributions, which can explain the variations in and correlations among the selected parameters. Third, the PCs generated are fitted to Johnson S_B distributions using the maximum likelihood method. Fourth, a standard normal distribution can be obtained from each of the fitted S_B distributions using the S_B transformation equation in Table 4-1, and the uncertainty associated with the selected interdependent parameters can be reflected

using standard normal distributions. Then, collocation points can be chosen from the transformed normal distributions using PCM.

To run the simulation model at the chosen collocation points, the corresponding original parameter values can be found through the reverse of the abovementioned transformation. First, Johnson inverse transformation is used to find the corresponding values of the collocation points on the PC distributions. Second, the original values corresponding to the collocation points can be achieved by multiplying the PCs by the eigenvectors and adding the mean. It is worth mentioning that the corresponding parameter value may lie outside its predefined range. In this case, the collocation points set including this parameter should be eliminated. After running the hydrological model at all of the chosen collocation points, the corresponding model output can be obtained to calculate the PCE coefficients at each simulation time step. With the coefficients, one PCE equation can be established for each time step to generate probabilistic output for uncertainty quantification (Figure 4-1).

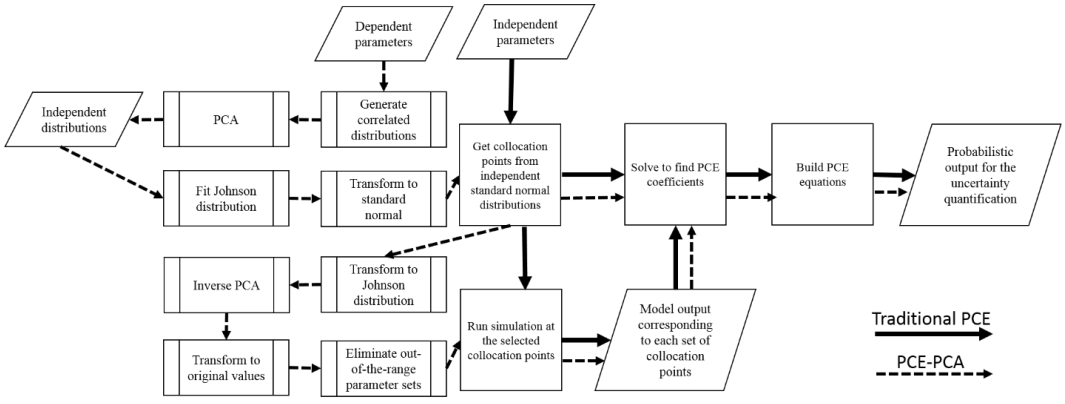

Figure 4-1: PCE-PCA framework

4.3. STUDY AREA AND DATA COLLECTION

A SWAT model for the Guadalupe River basin in Texas, US is developed and used as a case study to illustrate the applicability of the PCA-PCE framework. The Guadalupe basin is the fourth largest river basin in Texas. In this study, only the upper sub-catchment defined by the Spring Branch gage (Figure 4-2) is modeled. The Spring Branch catchment has a total area of 3,500 km^2 and an average slope of 0.12%. Four years (1989 to 1992) of daily rainfall, temperature, and flow data are used for uncertainty analysis. The weather and flow data were collected from the National Oceanic and Atmospheric Administration (NOAA) and the United States Geological Survey (USGS), respectively. The Spring Branch catchment is a semi-arid region with moderate temperature. The average minimum temperature during the four years is 16 ºC, and the average maximum temperature is 27 ºC. The average rainfall is 2.48 mm/day, and the streamflow rate is typically less than 17.6 m^3/s. There are some extreme rainfall events during the study period. The highest peak flow was 1,574 m^3/s, occurred on 21st December 1991.

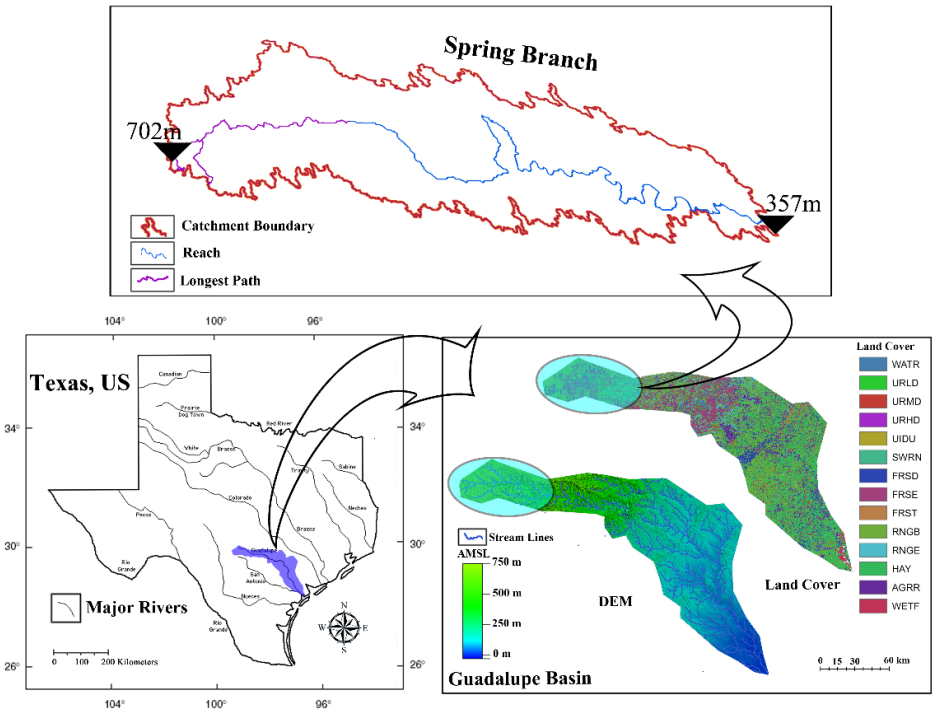

Figure 4-2: The Spring Branch in the Guadalupe River Basin, Texas, US. AMSL stands for above mean sea level. The land cover abbreviations can be found in Table A2 in Arnold et al. 2012

Soil data with a resolution of 30 m is collected from the State Soil Geographic (STATSGO) dataset from the United States Department of Agriculture (USDA). Land-cover data with a resolution of 30 m is obtained from the National Land-Cover Data Sets (NLCD), and the 90-m digital elevation model (DEM) is obtained from the Shuttle Radar Topography Mission 4.1 (SRTM V4.1) database, which is derived from the National Aeronautics and Space Administration (NASA) SRTM data with a grid size of 5°.

The five most sensitive SWAT parameters, including *CN*, soil evaporation compensation factor (*ESCO*), *Sol_AWC*, baseflow alpha factor (*alpha-bf*), and groundwater delay (*GW_Delay*), are selected for demonstration. The ranges of these five selected parameters are determined based on previous studies (Ghaith & Li, 2020; Yang et al., 2008) and the distributions are assumed to be uniform as shown in Table 4-2. The correlations among these sensitive parameters are chosen based on (Yang et al., 2008), and are summarized in Table 4-3.

Table 4-2: The Selected Five SWAT Parameters.

Parameter ID	Rule	Min	Max	Parameter description
R__CN2.mgt	Ratio	-0.2	0.2	SCS runoff curve number
V__ESCO.hru	Replace	0.8	0.9	Soil evaporation compensation factor
R__SOL_AWC(..).sol	Ratio	-0.2	0.4	Soil available water content
V__ALPHA_BF.gw	Replace	0	1	Baseflow alpha factor (days)
V__GW_DELAY.gw	Replace	30	450	Groundwater delay (days)

Table 4-3: Covariance Matrix.

	CN2	**ESCO**	**SOL_AWC**	**ALPHA_BF**	**GW_DELAY**
CN2	1				
ESCO	0	1			
SOL_AWC	0.44	0.56	1		
ALPHA_BF	0	0	0	1	
GW_DELAY	0	0	0	0	1

4.4. RESULTS

4.4.1. DATA GENERATION AND PREPARATION

To generate samples for the three dependent parameters (*CN2*, *ESCO*, and *Sol_AWC*), three dependent standard normal distributions are defined first based on

the correlations in Table 4-3. Then the standard normal distributions are transformed to uniform distributions, and the uniform distributions are translated and stretched/squeezed to fit the parameter ranges in Table 4-2. A total of 10,000 sets of samples are generated for *CN2*, *ESCO*, and *Sol_AWC*. Figure 4-3 shows the histograms of and correlations between the three parameters. Samples of the other two independent parameters (*Alpha_BF* and *GW_Delay*) are also generated based on the uniform distributions defined by Table 4-2.

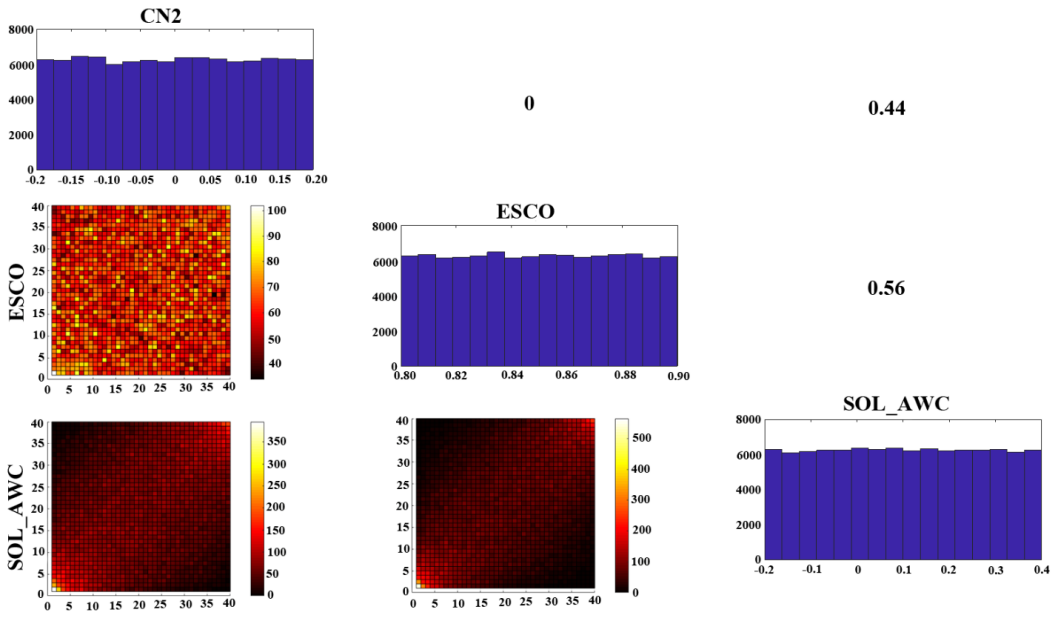

Figure 4-3: Generated samples of the interdependent parameters.

The correlated distributions are transformed to independent distributions using principle component analysis (PCA). Three orthogonal PCs, which explain 100% of the overall parameter variability are obtained and shown in Figure 4-4 (a)-(c). Johnson S_B distribution is used to fit the PC data. As shown in Figure 4-4 (a)-(c), each of the fitted lines matches the corresponding PC histogram very well, which

demonstrates the accuracy of the Johnson S_B distribution. After obtaining the fitted Johnson S_B distributions and their distribution parameters, the S_B equation in Table 4-1 is used to transform PC distributions to standard normal distributions. As shown in Figure 4-4 (d)-(f), three independent standard normal distributions that represent each of the PCs generated.

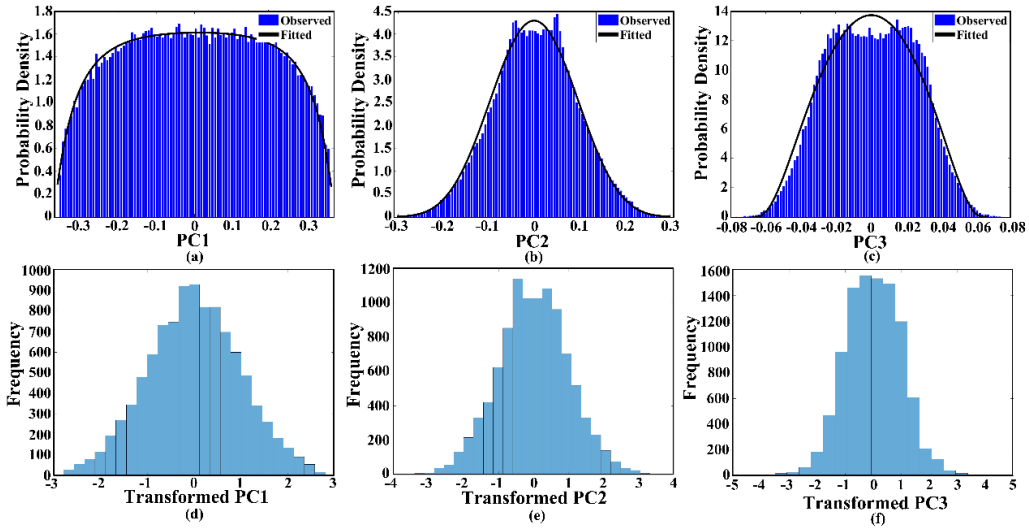

Figure 4-4: Histograms before and after the Johnson distribution transformation. Johnson distributions: (a) - (c); transformed standard normal distributions: (d) - (f).

As previously mentioned in Section 4.2.1, collocation points are chosen from the standard normal distributions at (-√3,0, √3). For the two independent parameters, Alpha_BF and GW_Delay, the corresponding parameter values are obtained through a direct transformation from a standard normal distribution to a uniform distribution. The three collocation points of Alpha_BF and GW_Delay are (0.037, 0.450, 0.863) and (47.5, 240.0, 432.5), respectively. For the three

interdependent parameters, *CN2*, *ESCO*, and *Sol_AWC*, 27 (3^3) sets of collocation points are obtained from the PCs' transformed standard normal distributions (Table 4-4). Following the procedure of eliminating out-of-range collocation points described in Section 4.2.5, 11 collocation point sets are selected for the three interdependent parameters. Figure 4-5 shows the predefined and reconstructed distributions of the three parameters. The three reconstructed distributions are flattened normal distributions with no tails.

A total of 99 collocation point combinations, 3 (*Alpha_BF*) × 3 (*GW_Delay*) × 11 (*CN2*, *ESCO*, and *Sol_AWC*), are obtained for the development of the PCEs.

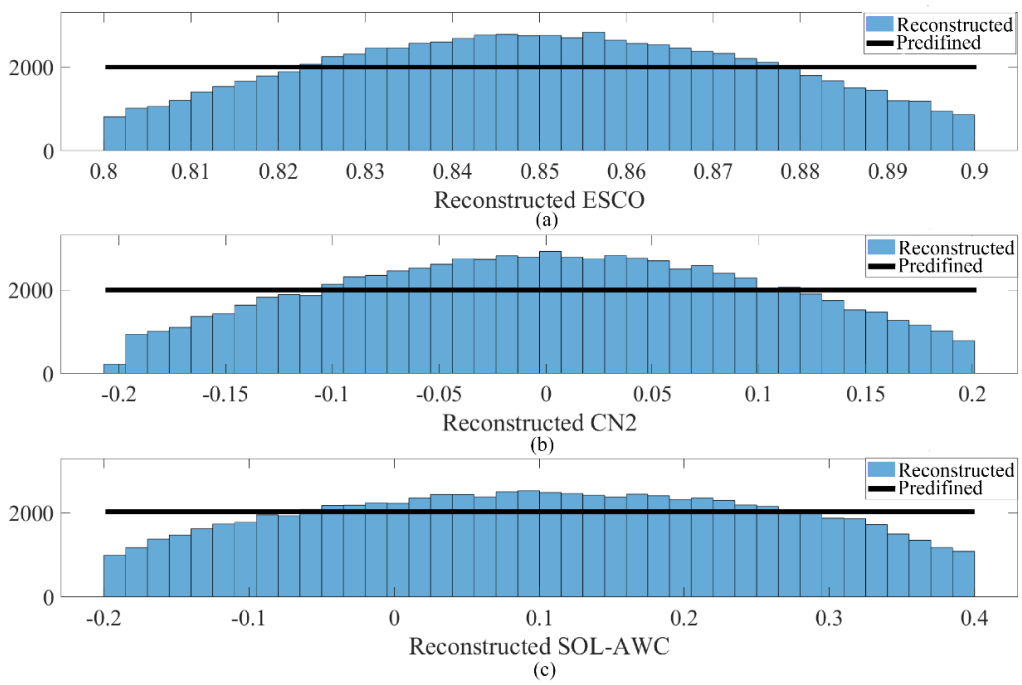

Figure 4-5: Reconstructed distributions.

Table 4-4: Collocation Points.

Standard Normal Distribution			Johnson Distribution			Actual Distribution			Within Range
PC1	PC2	PC3	PC1	PC2	PC3	ESCO	CN2	SOL AWC	
0.00	0.00	0.00	-0.31	-0.16	-0.04	0.81	-0.28	-0.10	No
0.00	0.00	1.73	-0.31	-0.16	0.00	0.85	-0.27	-0.11	No
0.00	0.00	-1.73	-0.31	-0.16	0.04	0.89	-0.27	-0.11	No
0.00	1.73	0.00	-0.31	0.00	-0.04	0.79	-0.14	-0.17	No
0.00	1.73	1.73	-0.31	0.00	0.00	0.83	-0.13	-0.18	Yes
0.00	1.73	-1.73	-0.31	0.00	0.04	0.88	-0.13	-0.18	Yes
0.00	-1.73	0.00	-0.31	0.16	-0.04	0.77	0.00	-0.24	No
0.00	-1.73	1.73	-0.31	0.16	0.00	0.82	0.00	-0.24	No
0.00	-1.73	-1.73	-0.31	0.16	0.04	0.86	0.01	-0.25	No
1.73	0.00	0.00	0.00	-0.16	-0.04	0.83	-0.14	0.17	Yes
1.73	0.00	1.73	0.00	-0.16	0.00	0.87	-0.14	0.17	Yes
1.73	0.00	-1.73	0.00	-0.16	0.04	0.91	-0.14	0.16	No
1.73	1.73	0.00	0.00	0.00	-0.04	0.81	0.00	0.10	Yes
1.73	1.73	1.73	0.00	0.00	0.00	0.85	0.00	0.10	Yes
1.73	1.73	-1.73	0.00	0.00	0.04	0.89	0.00	0.10	Yes
1.73	-1.73	0.00	0.00	0.16	-0.04	0.79	0.14	0.04	No
1.73	-1.73	1.73	0.00	0.16	0.00	0.83	0.14	0.03	Yes
1.73	-1.73	-1.73	0.00	0.16	0.04	0.88	0.14	0.03	Yes
-1.73	0.00	0.00	0.31	-0.16	-0.04	0.84	-0.01	0.45	No
-1.73	0.00	1.73	0.31	-0.16	0.00	0.88	-0.01	0.44	No
-1.73	0.00	-1.73	0.31	-0.16	0.04	0.93	0.00	0.44	No
-1.73	1.73	0.00	0.31	0.00	-0.04	0.83	0.13	0.38	Yes
-1.73	1.73	1.73	0.31	0.00	0.00	0.87	0.14	0.37	Yes
-1.73	1.73	-1.73	0.31	0.00	0.04	0.91	0.14	0.37	No
-1.73	-1.73	0.00	0.31	0.16	-0.04	0.81	0.27	0.31	No
-1.73	-1.73	1.73	0.31	0.16	0.00	0.85	0.27	0.31	No
-1.73	-1.73	-1.73	0.31	0.16	0.04	0.89	0.28	0.30	No

4.4.2. COMPARISON OF PCA-PCE AND MC

After running SWAT at all of the 99 collocation point combinations, a set of 99 equations is obtained to calculate the PCE coefficients at each time step. Then a time series of PCE equations is established as a surrogate model to quantify output uncertainties and generate probabilistic outputs. Meanwhile, another set of probabilistic output is generated through 10,000 runs of MC simulation. The probabilistic output from PCA-PCE and MC are compared based on their mean and standard deviation values. Figure 4-6a compares the mean values of probabilistic flow rates generated from PCA-PCE and MC. The mean flow results show a near-perfect fit with a coefficient of determination (R^2) of 0.998. The time series of mean values for both MC and PCE are presented in Figure 4-7. The two time-series are very similar, showing no significant differences. The flow deviation at each time step is also calculated and illustrated in Figures 6b and 8. It is shown in Figure 4-6b that there is a very good fit and the R^2 value is as high as 0.973. Figure 4-8 also implies that the standard deviations of MC output are well replicated by PCE. The results demonstrate that the proposed PCA-PCE approach which requires as few as 99 simulation runs can produce probabilistic outputs very similar to that of 10,000 runs of MC simulation. This implies that PCA-PCE can be effective as MC simulation in assessing the uncertainties associated with interdependent parameters while significantly reducing the computational time (by 99% in this study). Also, the use of improved PCE with interdependent parameters for complex hydrological model, and not just for simple equations is a novel contribution of this work.

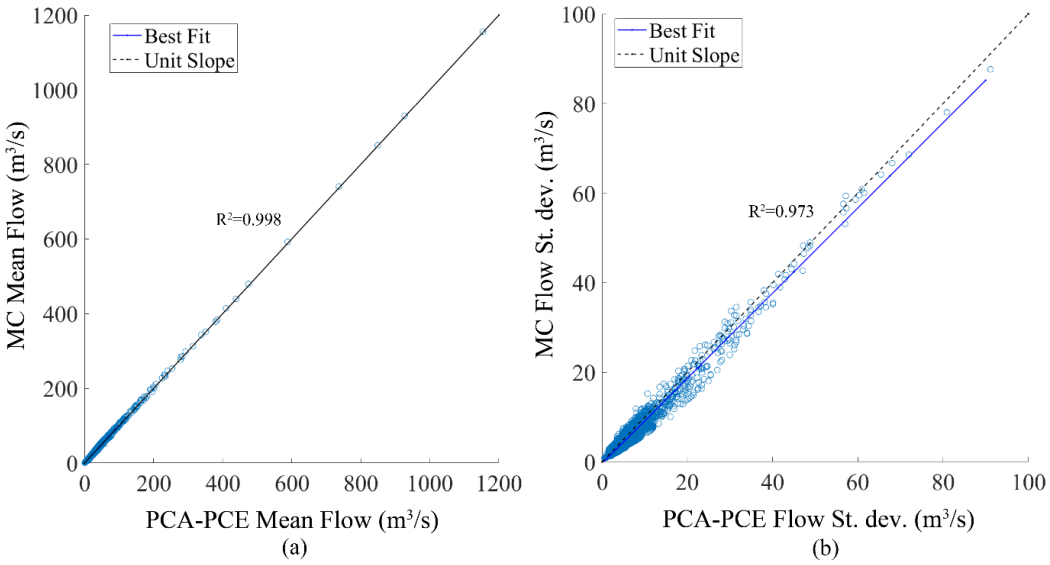

Figure 4-6: Scatter plot of MC against PCA-PCE: (a) Mean (b) Standard deviation.

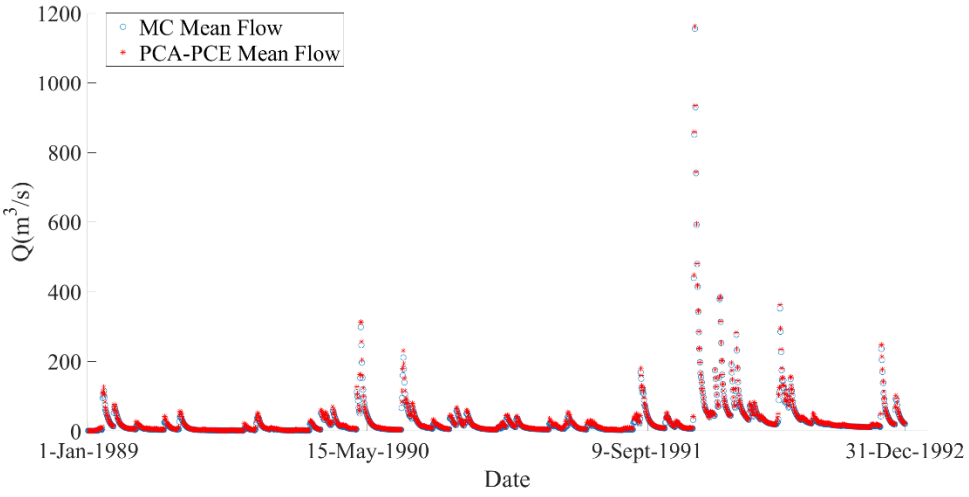

Figure 4-7: Comparison of mean flow time series generated by MC and PCA-PCE.

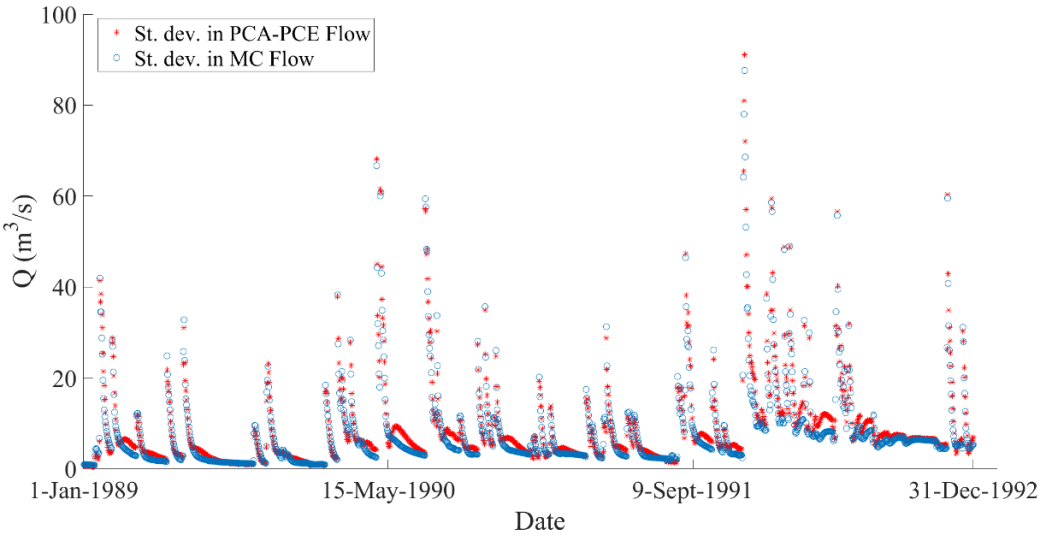

Figure 4-8: Time series of flow deviation.

4.4.3. SENSITIVITY OF PARAMETER INTERDEPENDENCY

In this study, the correlations among *CN2*, *ESCO*, and *Sol_AWC* are assumed based on a previous study (Yang et al., 2008b). The actual values of the correlation coefficients could change from one case study to another, depending on catchment characteristics. To illustrate the impacts of parameter correlations and to demonstrate the importance of addressing parameter interdependency, a sensitivity analysis of the correlation coefficients of the three parameters was carried out. A baseline scenario with independent parameters, where the parameter correlation coefficients are zero, was first built. Then, 42 additional scenarios with medium, high, and very high parameter interdependencies (which correspond to correlation coefficients of 0.5, 0.75, and 0.9, respectively) were analyzed. The obtained 43 scenarios were then run using the PCA-PCE framework. A time series of the daily

mean flow was generated for each of the 43 scenarios. The standard deviation value of mean flow at each time step was calculated and a histogram was plotted as shown in Figure 4-9. Figure 4-9 indicates that the standard deviation of mean flow due to parameter correlation can be as high as 18.5 m^3/s. It is noted that the frequency of low deviation is high, but that is mainly because the corresponding mean flow is also low. The correlation of variation (CV) is also calculated, and the histogram of the CV is plotted in Figure 4-10. Figure 4-10 shows that on average there is a CV of approximately 13%. The high CV are for the low flow event while the low CV are for high mean flow events. The results show a clear variation among different scenarios, which demonstrates the necessity to address parameter interdependency.

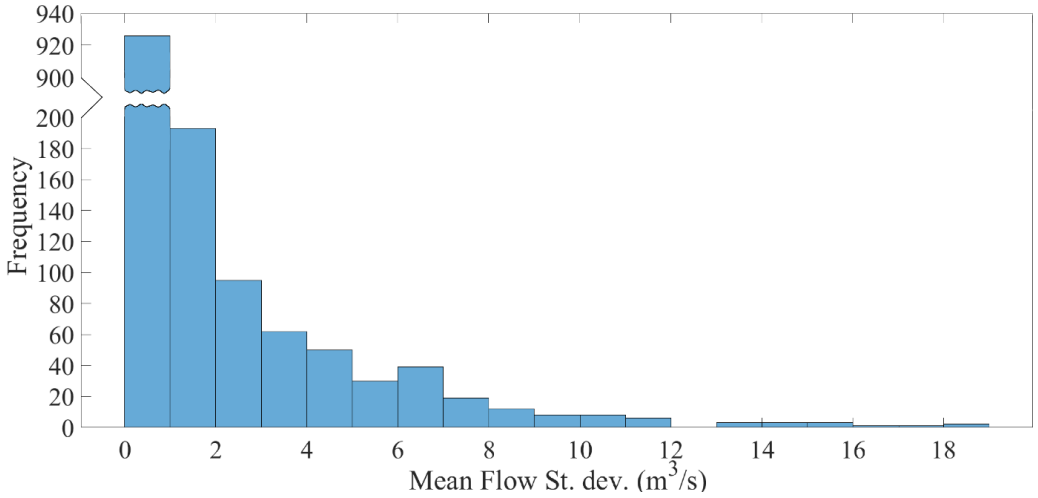

Figure 4-9: Histogram of standard deviation for mean flow due to different correlations.

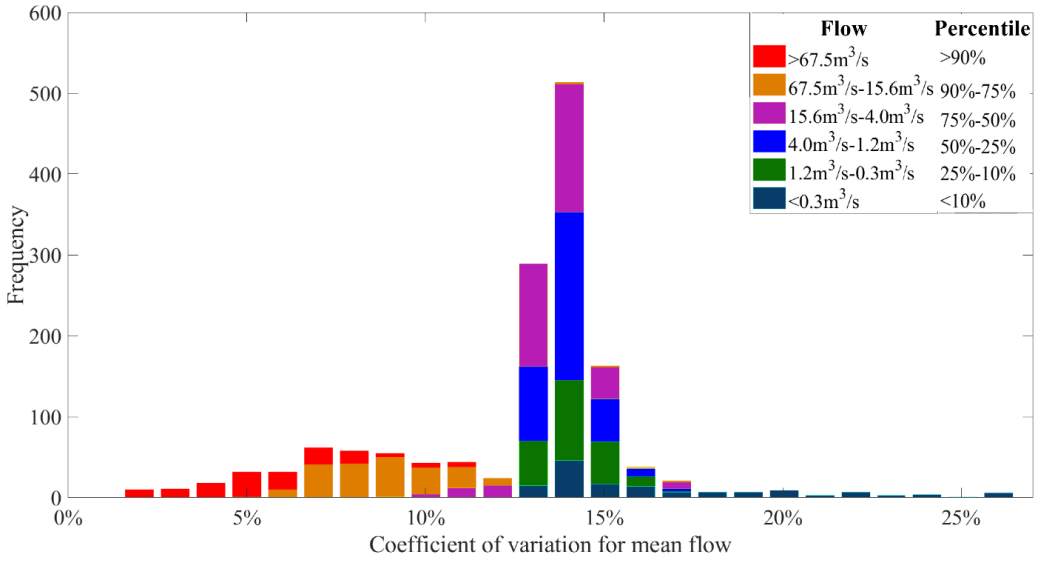

Figure 4-10: Coefficient of variation for mean flow due to different correlations scenarios.

To further analyze how the impacts of correlation coefficients, three days with 25th, 50th and 75th percentiles of flow were selected from the mean flow time series of the baseline scenario. The three days are July 4th, 1989, April 30th, 1991, and February 6th, 1990, respectively. The box plots of mean flow from all 43 scenarios on these days are shown in Figure 4-11. Figure 4-11 shows that the flow deviation increases with respect to the mean flow value. This means the flow deviation due to parameter interdependency could be particularly high during extreme flow events, such as floods. Furthermore, by comparing the results from the 42 scenarios with interdependent parameters to the baseline scenario, it was found that addressing parameter interdependency in SWAT could lead to a total flow change ranging from -10.6% to 10.9% during 1989-1992, which is similar to the

aforementioned results of CV of approximately 13%. The results once again demonstrate the importance to analyze parameter interdependency in hydrological modeling.

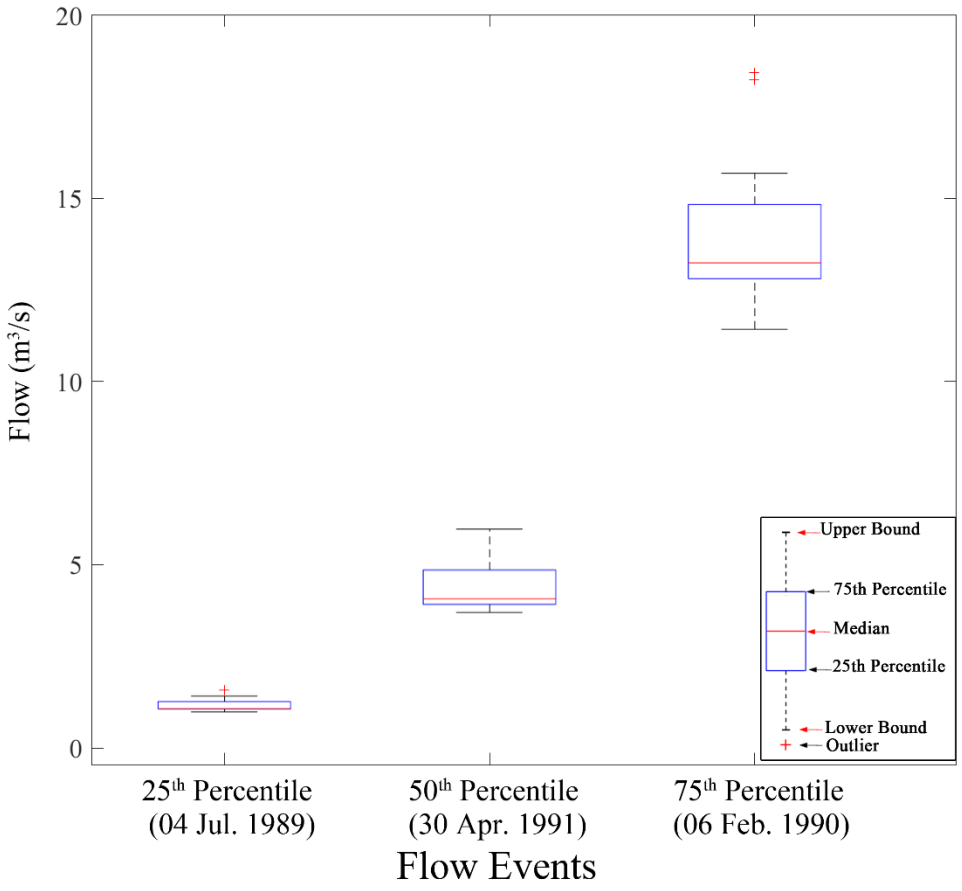

Figure 4-11: Histogram of mean flow corresponding to different correlations for different flow.

Since the model output is most sensitive to the change in parameter correlations when the mean flow is high, the peak flow event on December 22nd, 1991 was chosen for detailed sensitivity analysis. All the mean flow of the 43 scenarios for the peak day are calculated and plotted against each of the correlation values, as

shown in Figure 4-12. Figure 4-12a shows how the peak flow changes with respect to C1 (correlation between *CN2* and *ESCO*) while C2 (correlation between *CN2* and *Sol_Awc*) and C3 (correlation between *ESCO* and *Sol_Awc*) are constant. From Figure 4-12a, it can be deducted that from having no correlation to medium correlation for C1, the flow will decrease, whereas the change from high to extreme correlation has no effect. When C2 is higher or equal to C1 and C3 the flow will be lower than the relative base case scenario when both C2 and C3 equal to zero, otherwise, the peak flow will be higher. The outcome of this analysis can be summarized that C2 and C3 values have an interactive effect on the peak flow, while the C1 value effect is binary either no correlation or correlation. From Figure 4-12c, it is clearly visible that the effect of C3 also is binary, the exact value does not affect the mean peak flow while having both C1 and C2 constant; however, as the C1 or C2 values increase the flow values tends to decrease. When the C1 values is higher than the C2 value, the flow value will be higher than for the same C1 value with equivalent or higher C2. From Figure 4-12b, it can be deduced from the strong to extreme correlation value for C2, there is no change in the mean peak flow. Thus, for C1, C2 and C3 values being extreme or strong, correlation won't be significantly impacted. The uncertainty for the exact value of the correlation won't have a significant effect on the model outputs; however, knowing an approximate correlation value is important. A histogram of all 43 scenarios for the peak flow is plotted as shown in Figure 4-13 showing the minimum, mean and maximum values.

Figure 4-13 shows that the variation from different scenarios in mean flow is different than for minimum and maximum flow.

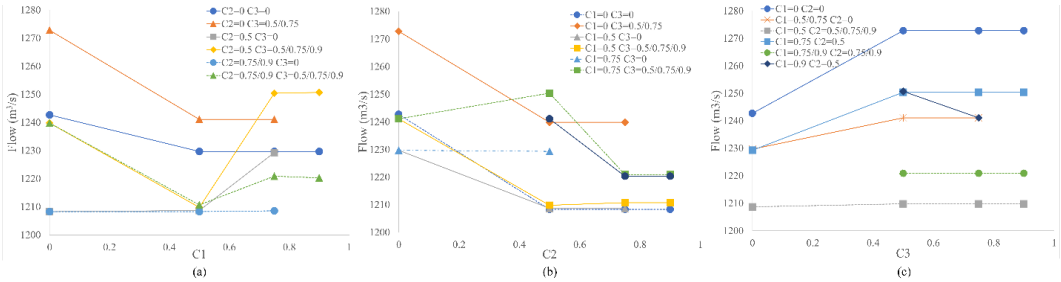

Figure 4-12: Mean flow for the peak flow day change with respect to (a) C1 while C2 and C3 are constant (b) C2 while C1 and C3 are constant (c) C3 while C1 and C2 are constant.

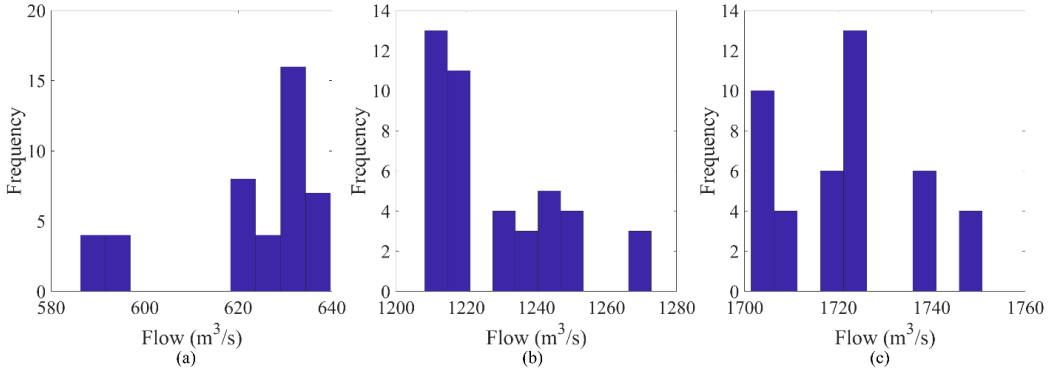

Figure 4-13: *Histogram for the peak flow day on 22nd Dec. 1991 due to different correlation scenarios (a) Min. flow (b) Mean flow (c) Max. flow.*

4.5. CONCLUSIONS

In this study, an integrated principle component analysis polynomial chaos expansion (PCA-PCE) approach is developed to support uncertainty analysis for hydrological models with interdependent parameters. The traditional PCE approach has been proven to be very efficient in quantifying parameter uncertainty; however, it could only be used when the model parameters are independent, which is not

always a valid assumption in hydrological modeling. In this improved PCE approach, PCA is used for orthogonal data transformation to convert interdependent hydrological parameters to independent variables, which are then fed into the traditional PCE framework to assess the propagation of parameter uncertainty in the hydrological model. In this study, the proposed PCA-PCE approach is applied on a semi-distributed hydrological model, Soil & Water Assessment Tool (SWAT). The SWAT model is used to simulate the rainfall-runoff relationship for the Guadalupe River basin in Texas, US.

The results demonstrate that the PCA-PCE approach can be as effective as Monte Carlo (MC) simulation in quantifying the uncertainties associated with interdependent hydrological parameters, while significantly reducing the computational requirements. There is a near perfect fit between the mean flow obtained from PCA-PCE and MC, with an R^2 value of 0.998, and the standard deviation of the flow shows an R^2 value of 0.973. PCA-PCE can reduce the computation time to generate probabilistic flow time series for the period of 1989 to 1992 by as much as 99% compared to MC simulation. Results of the sensitivity analysis on the correlation coefficients demonstrate that it is necessary to address parameter interdependency in hydrological modeling, particularly for the modeling of high flow events.

The proposed PCA-PCE approach provides a reliable, efficient, and promising alternative for analyzing the uncertainties of interdependent hydrological parameters and could provide valuable technical support for hydrological risk

assessment and management. For future work, PCA-PCE can be tested for more correlated parameters in SWAT and for more complex hydrological models.

4.6. ACKNOWLEDGEMENT

This study was supported by Natural Sciences and Engineering Research Council of Canada (NSERC).

The rainfall data is obtained from the National Oceanic and Atmospheric Administration website at https://www.ncdc.noaa.gov/ (last accessed 1 August 2017). The flow gage data is obtained from the United States Geological Survey website at https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/

GHCND: USW00012912/detail (last accessed 1 August 2017). The developed code for the proposed PCA-PCE model is available from the corresponding author by request.

4.7. REFERENCES

Almeida, R. A., Pereira, S. B., & Pinto, D. B. F. (2018). Calibration and Validation of the Swat Hydrological Model for The Mucuri River Basin SWAT-CUP. *Engenharia Agrícola*, *38*(1), 55–63. https://doi.org/http://dx.doi.org/10.1590/1809-4430-Eng.Agric.v38n1p55-63/2018

Arnold, J G, Kiniry, J. R., Srinivasan, R., Williams, J. R., Haney, E. B., & Neitsch, S. L. (2012). Soil & Water Assessment Tool. *Texas Water Resources Institute*, *439*(Appendix A), 566–567.

Arnold, Jeffrey G, Moriasi, D. N., Gassman, P. W., & White, M. J. (2012). SWAT : Model use , calibration , and validation.

Christiaens, K., & Feyen, J. (2002). Use of sensitivity and uncertainty measures in distributed hydrological modeling with an application to the MIKE SHE model. *Water Resources Research*, *38*(9). https://doi.org/10.1029/2001WR000478

Devak, M., & Dhanya, C. T. (2017). Sensitivity analysis of hydrological models : review and way forward. *Journal of Water and Climate Change |*, *8*(4), 557–575. https://doi.org/10.2166/wcc.2017.149

Fan, Y. R., Huang, W., Huang, G. H., Huang, K., & Zhou, X. (2014). A PCM-based stochastic hydrological model for uncertainty quantification in watershed systems. *Stochastic Environmental Research and Risk Assessment*, *29*(3), 915–927. https://doi.org/10.1007/s00477-014-0954-8

Fenfen, X., Shishi, C., & Ying, X. (2014). Dynamic system uncertainty propagation using polynomial chaos. *Chinese Journal of Aeronautics*, *27*(5), 1156–1170. https://doi.org/10.1016/j.cja.2014.08.010

Funahashi, H., & Kijima, M. (2012). A chaos expansion approach for the pricing of contingent claims. *Journal OfComputational Finance*, *18*(3), 27–58.

Ghaith, M., & Li, Z. (2020). Propagation of parameter uncertainty in SWAT: A probabilistic forecasting method based on polynomial chaos expansion and machine learning. *Journal of Hydrology*, *586*(March), 124854. https://doi.org/10.1016/j.jhydrol.2020.124854

Halldin, S. (2005). Modelling Hydrological Consequences of Climate Change — Progress and Challenges. *Advances in Atmospheric Sciences*, *22*(6), 789–797.

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, *36*(1), 149–176. https://doi.org/10.1093/biomet/36.1-2.149

Karlsson, I. B., Sonnenborg, T. O., Christian, J., Trolle, D., Duus, C., Olesen, J. E., … Jensen, K. H. (2016). Combined effects of climate models , hydrological model structures and land use scenarios on hydrological impacts of climate change. *Journal of Hydrology*, *535*(February), 301–317. https://doi.org/10.1016/j.jhydrol.2016.01.069

Lin, Yu-pin;Hong, Nien-ming; Wu, Pei-jung ;Lin, C. (2007). Modeling and assessing land-use and hydrological processes to future land-use and climate change scenarios in watershed land-use planning. *Environmental Geology*,

53(September), 623–634. https://doi.org/10.1007/s00254-007-0677-y

Longland, R. (2017). Astrophysics Correlated uncertainties in Monte Carlo reaction rate calculations. *Astronomy & Astrophysics*, *34*, 1–9. https://doi.org/10.1051/0004-6361/201730911

Ma, S., & Dai, Y. (2011). Principal component analysis based methods in bioinformatics studies. *Briefings in Bioinformatics*, *12*(6), 714–722. https://doi.org/10.1093/bib/bbq090

Mach, P., Thuring, J., & Šámal, D. (2006). Transformation of Data for Statistical Processing. *29th International Spring Seminar on Electronics Technology, IEEE*, *2*(2), 278–282. https://doi.org/10.1109/ISSE.2006.365112

Martinez, J., Crestaux, T., & Le Maitre, O. (2009). Polynomial chaos expansion for sensitivity analysis. *Reliability Engineering and System Safety*, *94*(Nov), 1161–1172. https://doi.org/10.1016/j.ress.2008.10.008

Mateu, J. (1997). Methods of Assessing and Achieving Normality Applied to Environmental Data. *Journal of Environmental Management*, *21*(5), 767–777.

Navarro, M., Witteveen, J., & Blom, J. (2014). Polynomial chaos expansion for general multivariate distributions with correlated variables.

Norbert Wiener. (1938). The Homogeneous Chaos. *American Journal of Mathematics , Vol . 60 (4)*, 897-936. https://doi.org/10.2307/2371268

Paulson, J. A., Buehler, E. A., Mesbah, A., Joel, A., Paulson, A., Edward, A., & Buehler, A. (2017). ScienceDirect Arbitrary Polynomial Chaos for Arbitrary Polynomial Chaos for Arbitrary Polynomial of Chaos for Uncertainty

Propagation Correlated Uncertainty Propagation of Correlated Uncertainty Propagation of Correlated Random Variables in Dynamic System. *IFAC-PapersOnLine*, *50*(1), 3548–3553. https://doi.org/10.1016/j.ifacol.2017.08.954

Rahman, S. (2018). A polynomial chaos expansion in dependent random variables ☆. *Journal of Mathematical Analysis and Applications*, *464*(1), 749–775. https://doi.org/10.1016/j.jmaa.2018.04.032

Study, S. C. (2019). Model Uncertainty Analysis Methods for Semi-Arid Watersheds with Di ff erent Characteristics : A Comparative. *Water*, *11*(1177). https://doi.org/doi:10.3390/w11061177

Tolessa, O., Nossent, J., Velez, C., Kumar, N., Griensven, A. Van, & Bauwens, W. (2015). Environmental Modelling & Software Assessment of the different sources of uncertainty in a SWAT model of the River Senne (Belgium). *Environmental Modelling and Software*, *68*, 129–146. https://doi.org/10.1016/j.envsoft.2015.02.010

Wang, S., Huang, G. H., Huang, W., Fan, Y. R., & Li, Z. (2015). A fractional factorial probabilistic collocation method for uncertainty propagation of hydrologic model parameters in a reduced dimensional space, *529*, 1129–1146. https://doi.org/10.1016/j.jhydrol.2015.09.034

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*, 37–52.

Wu, F., & Tsang, Y. (2004). Second-order Monte Carlo uncertainty / variability

analysis using correlated model parameters : application to salmonid embryo survival risk assessment. *Ecological Modelling*, *177*(Feb), 393–414. https://doi.org/10.1016/j.ecolmodel.2004.02.016

Wu, H., & Chen, B. (2015). Evaluating uncertainty estimates in distributed hydrological modeling for the Wenjing River watershed in China by GLUE , SUFI-2 , and ParaSol methods. *Ecological Engineering*, *76*, 110–121. https://doi.org/10.1016/j.ecoleng.2014.05.014

Wu, Q., Liu, S., Cai, Y., Li, X., & Jiang, Y. (2017). Improvement of hydrological model calibration by selecting multiple parameter ranges. *Hydrology and Earth System Sciences*, *21*(Jan), 393–407. https://doi.org/10.5194/hess-21-393-2017

Wu, Y., & Liu, S. (2012). Environmental Modelling & Software Automating calibration , sensitivity and uncertainty analysis of complex models using the R package Flexible Modeling Environment (FME): SWAT as an example. *Environmental Modelling and Software*, *31*, 99–109. https://doi.org/10.1016/j.envsoft.2011.11.013

Xie, H., & Lian, Y. (2013). Uncertainty-based evaluation and comparison of SWAT and HSPF applications to the Illinois River Basin. *Journal of Hydrology*, *481*, 119–131. https://doi.org/10.1016/j.jhydrol.2012.12.027

Xiu, D. (2010). *Numerical methods for stochastic computations: A spectral method approach* (6th ed.). Princeton, N.J: Princeton University Press.

Xiu, D., & Karniadakis, G. E. (2003). Modeling uncertainty in flow simulations via

generalized polynomial chaos. *Journal of Computational Physics*, *187*, 137–167. https://doi.org/10.1016/S0021-9991(03)00092-5

Xu, C. (2013). Decoupling correlated and uncorrelated parametric uncertainty contributions for nonlinear models. *Applied Mathematical Modelling*, *37*(24), 9950–9969. https://doi.org/10.1016/j.apm.2013.05.036

Xu, C., & Gertner, G. Z. (2008). Uncertainty and sensitivity analysis for models with correlated parameters. *Reliability Engineering and System Safety*, *93*(June), 1563–1573. https://doi.org/10.1016/j.ress.2007.06.003

Yang, J., Reichert, P., Abbaspour, K. C., Xia, J., & Yang, H. (2008a). Comparing uncertainty analysis techniques for a SWAT application to the Chaohe Basin in China, 1–23. https://doi.org/10.1016/j.jhydrol.2008.05.012

Yang, J., Reichert, P., Abbaspour, K. C., Xia, J., & Yang, H. (2008b). Comparing uncertainty analysis techniques for a SWAT application to the Chaohe Basin in China. *Journal of Hydrology*, *358*(1–2), 1–23. https://doi.org/10.1016/j.jhydrol.2008.05.012

Yu, A. (1994). Johnson's SB Distribution Function as Applied in the Mathematical Representation of Particle Size Distributions. Part 1 : Theoretical Background and Numerical Simulation. *Particle & Particle Systems Characterization*, *1*, 291–298.

Zhang, L., Jin, X., He, C., & Zhang, B. (2016). Comparison of SWAT and DLBRM for Hydrological Modeling of a Mountainous Comparison of SWAT and DLBRM for Hydrological Modeling of a Mountainous Watershed in Arid

Northwest China, (October 2017). https://doi.org/10.1061/(ASCE)HE.1943-5584.0001313

Chapter 5

CONCLUSIONS AND RECOMMENDATIONS

This dissertation presents three hybrid approaches to address uncertainty analysis for hydrological modelling. In general, there are three different types of uncertainties, including parameter, structure, and data uncertainties. Both the model parameter and structure uncertainties are addressed in this dissertation. The proposed hybrid approaches are based on the integration of data-driven modeling techniques with traditional hydrological simulation and/or uncertainty quantification methods. Their applicability is demonstrated using a case study of the Spring Branch watershed in the Guadalupe Basin in Texas, USA. These mix methods approaches can generate probabilistic flow forecasts with high accuracy in an efficient way that requires very low computational power, and thus provide valuable decision support for water resources planning and management.

5.1. MAIN FINDINGS IN CHAPTER 2

- A hybrid approach, called the HHDD approach, was developed to integrate a physical process-based model (HYMOD) and a data-driven model (artificial neural network, ANN) for hydrological forecasting.
- The proposed HHDD model showed a better performance than the traditional physically-based HYMOD.
- It was found that the accuracy of data-driven models is influenced by the input data. The performance of the proposed hybrid model could be

significantly enhanced by adding more input data, such as cumulative rainfall, to feed its second layer (i.e., ANN).

- It was found that the addition of input data in data-driven modeling does not guarantee a better performance. The hybrid model performed better with additional cumulative precipitation data but worse with additional cumulative outflow data.
- It was demonstrated that the proposed hybrid model is more robust than both physically-based and data-driven models.

5.2. MAIN FINDINGS IN CHAPTER 3

- Polynomial chaos expansion (PCE) was successfully used to analyze the parameter uncertainty of a complex, semi-distributed hydrological model (the Soil & Water Assessment Tool, SWAT).
- PCE generated similar results as Montel Carlo (MC) simulation for uncertainty analysis with 98% less computational time.
- An innovative approach based on the integration of a data-driven model (ANN) with PCE was developed. The developed PCE-ANN approach can enable the PCE to generate probabilistic forecasts through SWAT.
- It was found that the PCE-ANN approach is more efficient than MC, as it builds a surrogate model for uncertainty quantification using historical data and does not require the re-run of SWAT.

- It was demonstrated that the PCE-ANN approach is reliable for quantifying parameter uncertainty and it has a significant advantage in terms of efficiency, especially for complex hydrological models.

5.3. MAIN FINDINGS IN CHAPTER 4

- An improved PCE approach was developed to overcome the difficulty of addressing parameter interdependency for uncertainty analysis in hydrological modeling. In the developed approach, principal component analysis (PCA) was introduce to the traditional PCE framework for the first time, to decode parameter interdependency.
- A distribution transformation method, named Johnson transformation, was proposed to transform PCA output for the establishment of PCE.
- The developed PCA-PCE approach was applied to quantify the uncertainty of interdependent parameters in SWAT.
- The results showed that parameter interdependency could significantly affect the flow prediction (especially for peak flow events) and should not be neglected during uncertainty analysis.

5.4. RECOMMENDATIONS FOR FUTURE WORK

- More case studies should be conducted to further demonstrate the applicability and advantages of the three proposed approaches.
- Both the HHDD and PCE-ANN approaches could be improved by introducing more advanced data-driven algorithms.

- The performance of HHDD and PCE-ANN could be enhanced by adopting an input variable selection tool.
- The potential of the PCE-ANN and PCA-PCE approaches for quantifying the parameter uncertainty in other hydrological models could be further investigated.