Doctor of Philosophy

TOWARD COMPUTATIONALLY EFFICIENT AND ACCURATE REAL-TIME ENSEMBLE FLOOD FORECASTING USING PHYSICS-INFORMED SURROGATE MODELING WITH EXTRAPOLATING CAPABILITY

The Graduate School

of the University of Ulsan

School of Civil and Environmental Engineering

TRAN NGOC VINH

TOWARD COMPUTATIONALLY EFFICIENT AND ACCURATE REAL-TIME ENSEMBLE FLOOD FORECASTING USING PHYSICS-INFORMED SURROGATE MODELING WITH EXTRAPOLATING CAPABILITY

TRAN NGOC VINH

School of Civil and Environmental Engineering

University of Ulsan, Korea

TOWARD COMPUTATIONALLY EFFICIENT AND ACCURATE REAL-TIME ENSEMBLE FLOOD FORECASTING USING PHYSICS-INFORMED SURROGATE MODELING WITH EXTRAPOLATING CAPABILITY

Supervisor: Associate Professor JONGHO KIM

A Dissertation

Submitted to
the Graduate School of the University of Ulsan
In Partial Fulfillment of the Requirement
for the Degree of

DOCTOR OF PHILOSOPHY
(Civil Engineering)

by

TRAN NGOC VINH

School of Civil and Environmental Engineering

University of Ulsan, Korea

June, 2022

TOWARD COMPUTATIONALLY EFFICIENT AND ACCURATE REAL-TIME ENSEMBLE FLOOD FORECASTING USING PHYSICS-INFORMED SURROGATE MODELING WITH EXTRAPOLATING CAPABILITY

This certifies that the dissertation of
TRAN NGOC VINH is approved by

Committee Chair: Assist. Prof. JONGMUK WON

Committee Member: Assoc. Prof. JONGHO KIM

Committee Member: Assist. Prof. DAESEUNG KYUNG

Committee Member: Assist. Prof. SEONG JIN NOH

Committee Member: Assist. Prof. BYUNGHYUN KIM

School of Civil and Environmental Engineering

University of Ulsan, Korea

June, 2022

For my family

Acknowledgment

> "Surround yourself with people that push you to do better. No drama or negativity. Just higher goals and higher motivation. Good times and positive energy. No jealousy or hate. Simply bringing out the absolute best in each other"
>
> - *(Buffet, W)*

Many long days were spent staring at words, equations, and codes to produce this dissertation. However, none of these pages could have been written without significant support.

My deepest gratitude goes to my advisor, Prof. Jongho Kim, for enabling me to freely explore and develop this research throughout this long journey. His guidance, expertise, confidence, encouragement, and continuous support have been essential for generating this work. His profound knowledge, abundant experience, and integral view on research have been extremely helpful for not only this dissertation, but also my entire academic career. No words can sufficiently express my profound thanks for him for giving me the opportunity to pursue my Ph.D degree at the University of Ulsan, South Korea.

I am also very appreciative of their invaluable advice from my Ph.D. Committee members who take their time to review my dissertation and provide their organized remark and unforeseen angles of view on research I have never considered.

I would love to thank my fellow scholars, Manh Van Doi, Trung Duc Tran, Jihwan Kwon, Dojin Beak, Quynh Thi Nguyen, Lam Huu Phan, Ha Phuong La, Dr. Phat Cong Vo, Dr. Binh

Thanh Pham, and Dr. Dam Thanh Pham for sharing their expertise and friendship. Thanks for having a wonderful Ulsan life together with Vietnamese student association and Vietnam people association.

Successful and satisfactory journey of dissertation cannot be attained without their helps. Prof. Valeriy Y. Ivanov for sharing his knowledge in real-time flood forecasting. Dr. Donghui Xu, Dr. Chase Dwelle and Dr. Khachik Sargsyan for sharing their wisdom in sensitivity analysis, uncertainty quantification, and surrogate model.

I want to acknowledge the people and organizations that shared the data, open source codes, and models that have been used in this thesis. I am also grateful for National Centre for Hydro-Meteorological Forecasting, Vietnam Meteorological and Hydrological Administration for providing hydrometeorological data in Vietnam. I would like to thank the Uncertainty Quantification group, UQLab of ETH Zurich for sharing open-source algorithms.

Finally, my warm acknowledgments to my parents, Chung Tran and Ly Luong and my younger sister, Lan Huong Tran, for unconditionally loving me and supporting me when I decided to leave them in pursuit of my dreams. I have now been abroad, no matter the situation and hardship. Thank you for always being there, sometimes afar but always close.

Table of Contents

CHAPTER

List of Publications

1. **Tran, V. N.**, and J. Kim (2019), Quantification of predictive uncertainty with a metamodel: Toward more efficient hydrologic simulations, *Stochastic Environmental Research and Risk Assessment*. [Chapter II]

2. **Tran, V. N.**, and J. Kim (2021), Toward an Efficient Uncertainty Quantification of Streamflow Predictions Using Sparse Polynomial Chaos Expansion, *Water*. [Chapter II]

3. **Tran, V. N.**, M. C. Dwelle, K. Sargsyan, V. Y. Ivanov, and J. Kim (2020), A novel modeling framework for computationally efficient and accurate real-time ensemble flood forecasting with uncertainty quantification, *Water Resources Research*. [Chapter III]

4. **Tran, V. N.**, and J. Kim (2021), A Robust Surrogate Data Assimilation Approach to Real-Time Forecasting using Polynomial Chaos Expansion, *Journal of Hydrology*. [Chapter IV]

5. **Tran, V. N.**, and J. Kim (2022), Robust and efficient uncertainty quantification for extreme events that deviate significantly from training dataset using Polynomial Chaos-Kriging, *Journal of Hydrology*. [Chapter V]

6. Tran, T. D., **Tran, V. N.**, and J. Kim (2021), Improving the Accuracy of Dam Inflow Predictions Using a Long Short-Term Memory Network Coupled with Wavelet Transform and Predictor Selection, *Mathematics*.

7. Ivanov, V. Y., M. C. Dwelle, D. Xu, K. Sargsyan, D. B. Wright, N. Katopodes, J. Kim, **Tran, V. N.,** A. Warnock, S. Fatichi, P. Burlando, E. Caporali, P. Restrepo, B. F. Sanders, M. M. Chaney, A. M. B. Nunes, F. Nardi, E. R. Vivoni, E. Istanbulluoglu, G. Bisht, and R. L. Bras (2021), Breaking Down the Computational Barriers to Real-Time Urban Flood Forecasting, *Geophysical Research Letters*.

List of Figures

List of Tables

Abstract

Extreme floods occur more frequently than in the past due to climate warming, and they have more profound socio-economic impacts. Flood forecasting is one of the important components of flood risk management and mitigation but is subject to multiple uncertainties caused by meteorological inputs, initial states, model structures, and model parameters. Numerous research efforts investigated the uncertainties in the tasks of flood prediction. However, at present we entirely lack comprehensive studies that can handle long-lasting challenges of computational burden, inaccuracy, and unreliable predictability in real-time ensemble flood forecasting with uncertainty quantification. This dissertation aims to gain comprehensive knowledge of building novel modeling frameworks for computationally efficient and accurate real-time ensemble flood forecasting with uncertainty quantification.

In this dissertation, a series of innovative methodologies have been developed for accurate, robust, and efficient uncertainty quantification of hydrological models in predicting floods. These methods include: (i) a unified modeling framework based on generalized likelihood uncertainty estimation (GLUE) framework coupled with polynomial chaos expansion (PCE) for fast and robust quantifying and understanding the parameter uncertainty of hydrological model in flood predictions; (ii) a novel modeling framework, for computationally efficient and accurate real-time ensemble flood forecasting with uncertainty quantification, which combines three modeling techniques together for the first time: surrogate modeling, parameter inference, and data assimilation; (iii) a novel, robust and efficient surrogate data assimilation approach for real-time flood forecasting using PCE to replace internal processes of Ensemble Kalman filters (EnKFs);

and (iv) a new surrogate model, named polynomial chaos-kriging (PCK), that can provide reliable ensemble results, even for extreme events that deviate significantly from the training data space.

Corresponding major accomplishments of this dissertation are abridged as follows. (i) The PCE surrogate model is firstly integrated into the GLUE framework to offset the computational demands of an uncertainty quantification task. It provides the benefits of an interpretable, probabilistic framework on which to make inferences about the drivers of model behavior, as well as the sensitivities of the model's output to the uncertain inputs. (ii) The novel framework of real-time ensemble flood forecasting embraces the benefits of three modeling techniques together for the first time: (1) PCE surrogates can significantly decrease computational time; (2) Parameter inference (GLUE) allows for model faster convergence, reduced uncertainty, and superior accuracy of simulated results; and (3) EnKFs assimilate errors that occur during forecasting. This framework provides a holistic, robust approach to accounting and understanding the uncertainties of hydrological parameters and vastly reducing the computational burden of ensemble simulations in real-time flood prediction. This modeling framework contributes to a shift in modeling paradigm arguing that complex, high-fidelity hydrologic and hydraulic models should be increasingly adopted for real-time and ensemble flood forecasting. (iii) The power of surrogate approaches is further exploited to develop new surrogate filters by replacing the internal processes of the EnKFs with PCE. A comprehensive investigation into how to configure a surrogate filter indicates that the new partial (replacing part of original filters) and invariant (valid for entire time periods) approaches are preferred in terms of accuracy and efficiency, which helps directly reduce the number of dimensions and bridge the gap between hindcasting and real-time forecasting. This proposed surrogate filter will be a promising alternative tool for performing computationally-intensive data assimilation in high-dimensional problems. And (iv) a new surrogate model named

polynomial chaos-kriging (PCK) is developed by combining the advantages of two well-known surrogate models, PCE and kriging. This combination enabled streamflow prediction for extreme events that deviated significantly from the trained data space, and allowed for quantifying predictive uncertainty robustly and efficiently. This finding will ultimately inspire novel designs toward a potentially more comprehensive surrogate model.

CHAPTER I

Introduction

> "Without some goals and some efforts to reach it, no man can live"
>
> - *(Dostoyevsky, F)*

1.1 Motivation of the research

1.1.1 Significances and challenges of flood forecasting

Floods are one of the most destructive natural hazards and lead to severe social and economic impacts in most corners of the world (Fig. 1.1) [*Dottori et al.*, 2018; *Paprotny et al.*, 2018]. In the last decade, flooding has affected more than two billion people (Fig. 1.2a) with the estimated socio-economic losses that are more than $662 billion with the trend of growth over the globe (Fig. 1.2b) [*CRED-UNISDR*, 2015; *Ward et al.*, 2017]. This trend is mainly induced by a rising number of global extreme floods caused by the growth of mega urban regions, deforestation, and extreme precipitation events [*Tanoue et al.*, 2016; *Paprotny et al.*, 2018]. Where the latter reason is recognized as the most cause of flooding (bar plot in Fig. 1.1) with recent extreme precipitation events that have highlighted the vulnerability of settlements and infrastructures to flooding [*Bloschl et al.*, 2020].

The magnitude and frequency of extreme flooding events are likely to rise due to global warming [*Donat et al.*, 2016; *Prein et al.*, 2016]. Given the worldwide significance of floods, early warning systems and flood forecasting need to be as robust as possible. Adequate warnings allow people to protect themselves and their property from the harmful effects of floods [*WMO*, 2018]. This motivates the need to improve the operational flood risk system, in particular, flood forecasting, in order to serve as a premise for early warning, preparing, and give timely emergency plans to mitigate the damage of floods. [*Cloke and Pappenberger*, 2009; *Thomas E. Adams*, 2016; *WMO*, 2018].

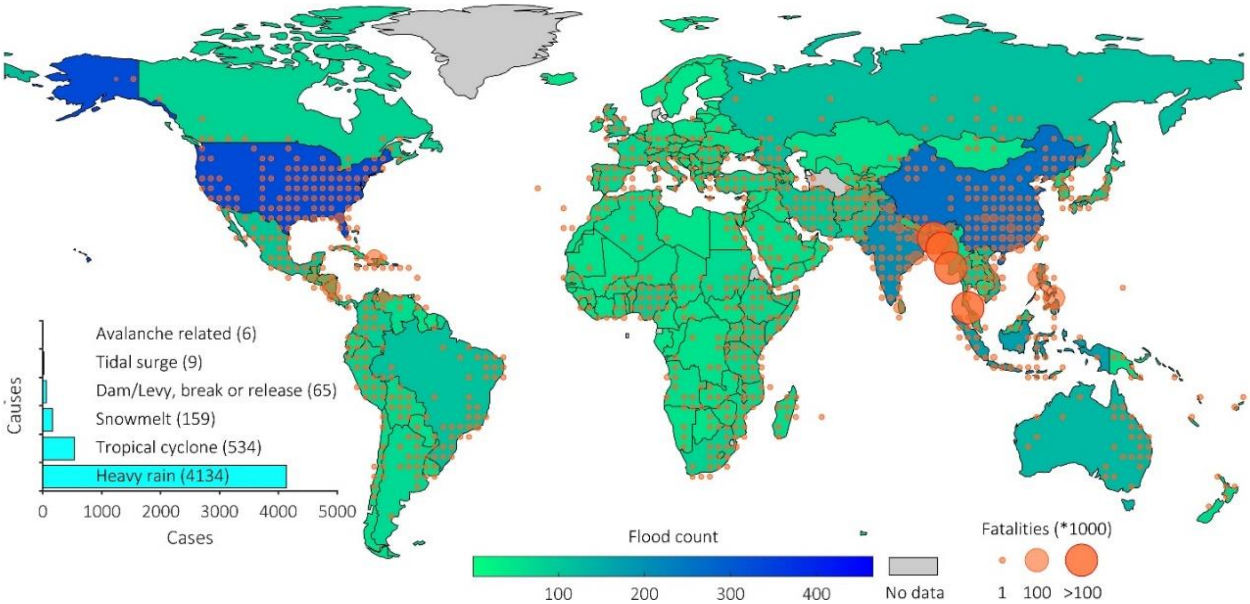


Figure 1.1. Count of flood events (color bar) and fatalities (orange circle) for the period of 1985-2019 for each country and 2.5^{o} × 2.5^{o} latitude-longitude grid, respectively. The bar plot represents the number of causes of flooding. The data sources: Global Active Archive of Large Flood Events.

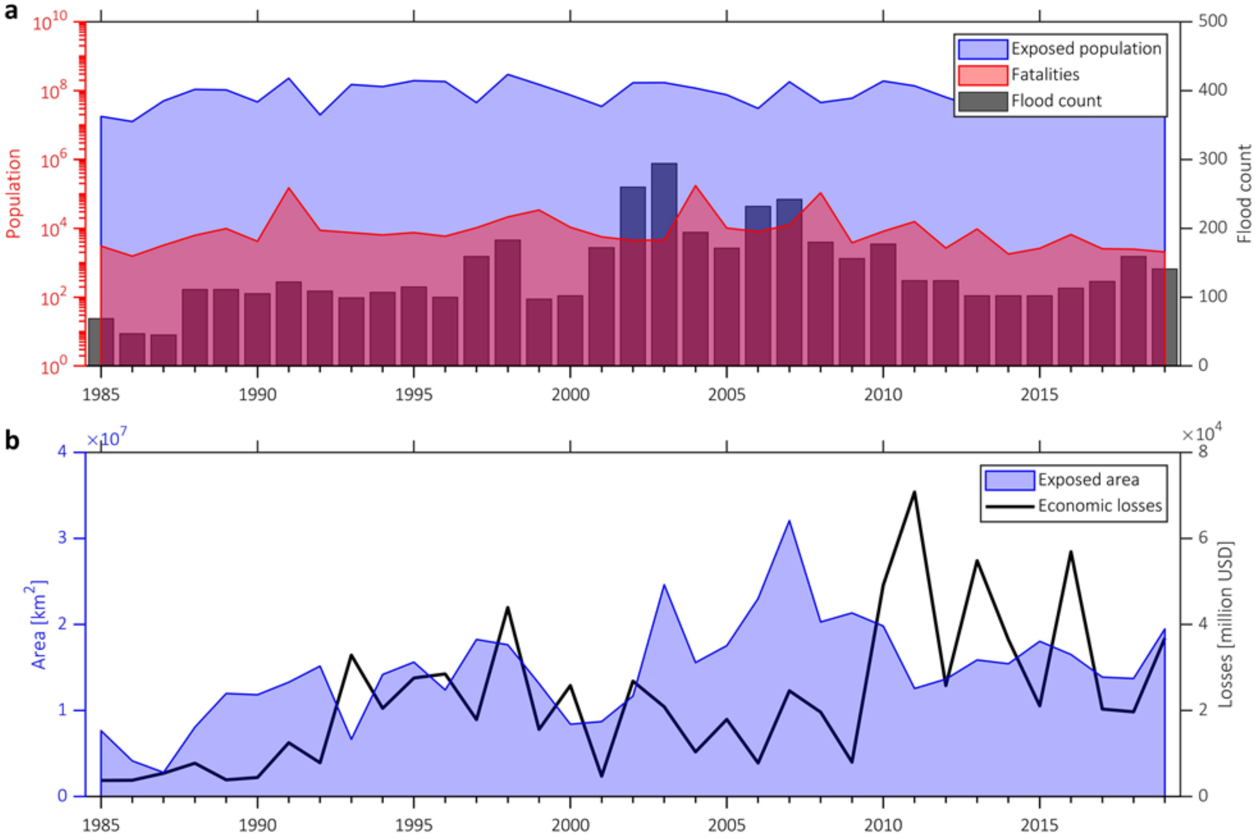


Figure 1.2. (a) Annual counts of global flood events (black bar), exposed population (blue shaded), and fatalities (red shaded) affected flood from 1985 to 2019. (b) Reported global flooded area (blue shaded) and economic loss (black line) from 1985 to 2019. The data sources: Global Active Archive of Large Flood Events.

Hydrologic, hydraulic, or coupled models (called process-based models) are the most commonly used methods to predict flooding phenomena [*Bogner and Pappenberger*, 2011; *Liu et al.*, 2012; *Li et al.*, 2015; *Si et al.*, 2015], although advanced data-driven models currently have been confirmed to produce high accuracy for flood forecasting without in the light of physical laws underlying the rainfall-runoff process [*Bai et al.*, 2016; *Hu et al.*, 2019b; *Jiang et al.*, 2020]. Numerous process-based models have been developed from simplified, lumped, and conceptual to sophisticated, spatially distributed, and highly interrelated processes of water movement, energy, topography, vegetation processes, even anthropogenic impacts in a watershed [*Vrugt et al.*, 2006a]. Each model has adopted various representations of the physical processes for water flow over an entire area presented by mathematical equations [*Smith et al.*, 2004; *Vrugt et al.*, 2005; *Kim et al.*, 2013; *Maxwell et al.*, 2014; *Kim and Ivanov*, 2015].

Although the growing availability of hydrologic observations at fine spatial and temporal scales nowadays helped to improve understanding of the physics and dynamics of the hydrologic system and develop more advanced and sophisticated models [*Reed et al.*, 2004; *Liu and Gupta*, 2007; *Smith et al.*, 2012]. Their outputs (e.g., streamflow) still exists extensive uncertainties due to a lack of knowledge of the involved physical processes and their interactions, the infeasibility of identifying model parameters, and difficulties in measuring (or estimating) initial and current states [*Beven*, 1989; *Butts et al.*, 2004; *Ajami et al.*, 2007; *Moradkhani and Sorooshian*, 2008; *Ivanov et al.*, 2010; *Kim et al.*, 2012a; *DeChant and Moradkhani*, 2014; *Kim and Ivanov*, 2014; *Kim et al.*, 2016a; *Mockler et al.*, 2016]. Furthermore, due to the complexities of natural phenomena represented by equifinality [*Beven and Freer*, 2001; *Beven*, 2006], hysteresis [*Wei and Dewoolkar*, 2006; *Ivanov et al.*, 2010; *Fatichi et al.*, 2015], non-uniqueness [*Beven*, 2000; *McKenna et al.*, 2003; *Kim and Ivanov*, 2014; *Kim et al.*, 2016a], non-linearity [*Kitanidis and Bras*, 1980; *Xie and Zhang*, 2010; *Kim and Ivanov*, 2015], and internal variability [*Nikiema and Laprise*, 2011; *Mondal and Mujumdar*, 2012; *Lafaysse et al.*, 2014; *Kim et al.*, 2016c; *Kim et al.*, 2016b; *Kim et al.*, 2018], perfect predictions using numerical models are infeasible. Thus, quantifying and reducing uncertainties has been a major challenge for researchers in flood prediction and in water planning and supply, sediment management, and reservoir operation [*Faber and Stedinger*, 2001; *Todini*, 2004; *Benke et al.*, 2008; *Saad and Ghanem*, 2009; *Kim et al.*, 2016c; *Kim et al.*, 2016b]. Without consideration of the associated uncertainty, predicted results would be limited value to real-world flood management.

1.1.2 The need for an accurate and computationally efficient ensemble flood forecasting framework

In the past decades, numerous approaches have been developed to quantify the predictive uncertainty of hydrologic responses [*Beven and Binley*, 1992; *Moradkhani et al.*, 2005c; *Vrugt et al.*, 2005; *Weerts and El Serafy*, 2006; *Han et al.*, 2007; *Lohani et al.*, 2014], but this remains a challenge. First, a number of optimization techniques (such as downhill simplex [*Nelder and Mead*, 1965], the shuffled complex evolution (SCE-UA) method [*Duan et al.*, 1992], and the particle swarm optimization [*Kennedy and Eberhart*, 1995]) have been developed to find the single best fitting parameter set. They were successfully used in diverse engineering applications, but the calibration techniques still lack the ability to properly treat the various uncertainties inherent in the system [*Moradkhani et al.*, 2005a]. Furthermore, although a set of parameters obtained from a watershed best represents the behavior of the basin, it may not work for other watersheds [*Beven*, 1989; *Beven and Binley*, 1992; *Moradkhani and Sorooshian*, 2008]. Failure to identify a range of parameter values may increase uncertainty in the model outputs [*Moradkhani and Sorooshian*, 2008].

Second, data assimilation methods such as the ensemble Kalman filter (EnKF) [*Evensen*, 1994] and the particle filter (PF) [*Arulampalam et al.*, 2002] also have received much attention, especially in real-time forecasting. This is because these techniques continuously update model states and parameters whenever new observations are available to improve model predictability [*Vrugt et al.*, 2005; *Liu et al.*, 2012; *Moradkhani et al.*, 2012]. However, some parameters of hydrologic models may not be completely identifiable, and therefore do not show convergence during the assimilation process for complex domains [*Moradkhani et al.*, 2005a; *Moradkhani et al.*, 2012]. Moreover, the potential collapse of EnKF (in which all ensemble members result in a similar value) can be resolved by updating each ensemble member with an independently perturbed observation [*Burgers et al.*, 1998]; in other words, the magnitudes of model states or parameters cannot converge to the values corresponding to observations even after many repeated attempts at the assimilation process. The PF method has an advantage over EnKF in terms of reducing numerical instability because it provides particle weights and uses non-Gaussian state-space models [*Liu et al.*, 2012]. On the other hand, the PF method is computationally more expensive than EnKF, as it generally requires more ensemble members based on the sequential Monte Carlo method [*Moradkhani et al.*, 2005a; *Liu et al.*, 2012].

Alternative probabilistic methods have been developed to deal with uncertainty quantification. They are mostly based on the Monte Carlo (MC) procedure, which provide the posterior distribution of parameters [*Beven*, 2006]. The most common uncertainty quantification methods include: Generalized Likelihood Uncertainty Estimation (GLUE) [*Beven and Binley*, 1992], Bayesian recursive estimation technique (BaRE) [*Thiemann et al.*, 2001], the Metropolis method [*Kuczera and Parent*, 1998], the Shuffled Complex Evolution Metropolis (SCEM-UA) [*Vrugt et al.*, 2003a; *Vrugt et al.*, 2003b], and the DiffeRential Evolution Adaptive Metropolis (DREAM) scheme [*Vrugt et al.*, 2008b]. The GLUE method attempts to identify a variety of parameter sets (namely "behavioral parameter set") given likelihood functions and cutoff threshold values [*Moradkhani and Sorooshian*, 2008]. The BaRE approach can simultaneously perform parameter estimation and hydrologic prediction. Uncertainties associated with parameter estimates are updated recursively, and uncertainty in output predictions becomes smaller when measurements are successively assimilated [*Thiemann et al.*, 2001]. The Metropolis method developed by *Kuczera and Parent* [1998] uses a random walk that adapts to a true probability distribution to describe parameter uncertainty. The SCEM-UA is the extension of the SCE-UA algorithm, which combines the strengths of the Metropolis algorithm, controlled random search, competitive evolution, and complex shuffling to continuously update a proposal distribution and evolve it into a posterior target distribution [*Vrugt et al.*, 2003b]. The DREAM scheme is an adaptation of the SCEM-UA using a novel Markov Chain Monte Carlo (MCMC) sampler. It shows excellent efficiency for complex, highly nonlinear, and multimodal target distributions [*Vrugt et al.*, 2008b]. It is generally necessary to increase the number of repeated model runs to successfully capture the uncertainty of model predictions using these methods [*Beven*, 2006; *Moradkhani and Sorooshian*, 2008]. Consequently, thousands of simulations have to be run to obtain the uncertainty in the outputs because of the high dimensionality of the parameter space and heterogeneity of input fields. The traditional method for uncertainty quantification is not always feasible for the high-fidelity model, which is computationally expensive [*Todini*, 2004; *Beven*, 2006; *Rosenzweig et al.*, 2021]. Even using computational advancements (a workstation, a computational cluster, or cloud computing infrastructure), the computational burden still remains a barrier for the task of real-time ensemble flood forecasting [*Wing et al.*, 2019; *Hosseiny et al.*, 2020; *Xu*, 2020; *Rosenzweig et al.*, 2021].

Recently, surrogate modeling approach gradually becomes an attractive solution to address the computational issue in the task of uncertainty quantification [*Razavi et al.*, 2012b; *Wang et al.*, 2018; *Dwelle et al.*, 2019; *Zhang et al.*, 2020]. This approach substitutes a high-cost deterministic model with a cheap-to-run "surrogate" model that reproduces comparable physical properties but has a lower computational cost [*Razavi et al.*, 2012b; *Asher et al.*, 2015; *Tran et al.*, 2020]. The central premise for a surrogate model to provide results consistent with the original model is to be able to approximate the relationship between input and output similar to the original model [*Wang and Shan*, 2007; *Smith*, 2013; *Asher et al.*, 2015; *Rajabi*, 2019]. The use of a surrogate model is computationally inexpensive compared to the original process-based model, and it can be rigorously sampled for uncertainty propagation, parameter inference, or sensitivity analysis. A full analysis of the uncertainty for any outputs of interest is carried by running a Monte Carlo simulation with the surrogate model. Due to the efficiency of the surrogate, thousands of simulations can be finished in a reasonable time [*Asher et al.*, 2015; *Dwelle et al.*, 2019; *Zhang et al.*, 2020]. Considering abundant published works, Gaussian process or kriging [*Santner et al.*, 2003], and polynomial chaos expansion (PCE) [*Wiener*, 1938] were preferred in statistics and engineering (often computational fluid dynamics) [*Le Maître et al.*, 2002; *Zhao and Xue*, 2010; *Razavi et al.*, 2012b; *Baştuğ et al.*, 2013; *Asher et al.*, 2015; *Fan et al.*, 2016; *Schöbi et al.*, 2017; *Ricciuto et al.*, 2018; *Wang et al.*, 2018; *Dwelle et al.*, 2019; *Rajabi*, 2019; *Wang et al.*, 2020; *Zhang et al.*, 2020]. However, prior studies only stop at using the surrogate model to deal with inverse problems. (e.g., uncertainty quantification or sensitivity analysis), little research has been done with it on real-time flood predictions and, especially, it's extrapolating capability.

1.2 Research scope

As listed in works of literature, at present we still lack solutions of an accurate and computationally efficient ensemble flood forecasting approach with uncertainty quantification. Addressing the corresponding shifts across accuracy, predictability, extrapolating capability, and computational efficiency in flood forecasting is one of the most critical challenges facing society nowadays. This dissertation aims to examine the existing problems in ensemble flood forecasting and suggest novel methodologies to provide solutions to those problems. Research scopes are manifest in the following chapters.

In Chapter 2, an efficient framework of uncertainty quantification is presented, where a well-known generalized likelihood uncertainty estimation (GLUE) framework of *Beven and Binley* [1992] is revisited. It is applied to quantify the parameter uncertainty of a lumped, deterministic rainfall–runoff model (Nedbør–Afstrømnings model, NAM) in hydrologic simulations. Firstly, two new indexes based on the efficiency and accuracy performance of GLUE are formed to optimize the cutoff threshold of likelihood function. The appropriate number of ensemble behavioral sets is then specified to maintain the sufficient range of uncertainty but to avoid any unnecessary computation. To offset the computational cost of quantifying the uncertainty, the GLUE is coupled with a PCE surrogate model, where the size of experimental design and polynomial degree are reasonably determined. Specifically, the least angle regression method is introduced to construct a more accurate surrogate model with a smaller size of training data compared to a surrogate model using a regular regression method. The performance aspects of the developed framework (e.g., accuracy and efficiency) are presented by applying the framework to quantify the parameter uncertainties in hydrologic simulations for 8 and 9 flood events that occurred in the Thu Bon watershed in Vietnam and Hongcheon watershed in South Korea, respectively.

Chapter III introduces a novel modeling framework that simultaneously improves accuracy, predictability, and computational efficiency in real-time ensemble flood forecasting. It embraces the benefits of three modeling techniques integrated together for the first time: surrogate modeling, parameter inference, and data assimilation. The use of PCE surrogates significantly decreases computational time. Parameter inference (using GLUE) allows for model faster convergence, reduced uncertainty, and superior accuracy of simulated results. EnKFs assimilate errors that occur during forecasting. To examine the applicability and effectiveness of the integrated framework, 18 difference approaches are developed according to how surrogate models are constructed, what type of parameter distributions are used as model inputs, and whether model parameters are updated during the data assimilation procedure. The performance in terms of accuracy, predictability, and computational efficiency of 18 approaches is investigated by applying them to forecast floods that occurred in the Vu Gia watershed in Vietnam. In light of the parsimony and good skill of the modeling framework, the novelty and applicability of the developed modeling framework are discussed.

A novel, robust and efficient approach to surrogate data assimilation is presented in Chapter IV with the aim of addressing the computational challenge due to the requirement of repetitive model evaluations in real-time ensemble flood forecasting. A total of eight surrogate filters are developed by replacing the whole or internal processes of EnKFs with PCE surrogates, where the formulation of these surrogate filters can be characterized according to their different surrogate structures, building systems, and assimilating targets. An advanced optimization scheme, named sequential experimental design-polynomial degree (SED-PD), is also advised to compensate for the potential shortcomings of the existing sequential experimental design (SED). Its dual optimization system resolves the issue of SED by which the value of the polynomial degree had to be selected ad-hoc or by trial and error; its multiple stopping criteria ensure convergence even when an accuracy metric does not monotonically decrease over iterations. The chapter also provides investigations of the accuracy and efficiency performances of these surrogate filters and two original filters (i.e., EnKF and Dual EnKF) with both synthetic and real data experiments of data assimilation for flood forecasting in the Vu Gia watershed in Vietnam. Several discussions are then presented to highlight the necessity and transferability of surrogate filters and SED-PD in applications of data assimilation and surrogate construction, respectively.

Chapter V introduces a new surrogate model (called polynomial chaos-kriging, PCK) that merges PCE and Gaussian process with kriging variance. The aim of this combination is to enable the surrogate model to predict streamflow for extreme events that deviated significantly from the trained data space, and allowed for quantifying predictive uncertainty robustly and efficiently. The predictability skill and superiority of PCK compared to PCE and OK (ordinary kriging) are investigated through experiments of quantifying the uncertainty to eight test flood events using a modeling framework that applies GLUE to surrogate models. Additionally, the effects of the acceptance threshold types on the model accuracy and efficiency are discussed. And a new "performance score" is formed to indicate how much better the accuracy and efficiency of the surrogate model are over the original model, thereby providing a guideline for selecting an appropriate surrogate emulator.

The last chapter summarizes this dissertation and addresses perspectives for ongoing and future directions of the research. The major conclusions and critical assumptions of the conducted research are presented, along with the feasibility of expanding the developed modeling frameworks for ensemble flood forecasts associating with complex real-world engineered systems.

CHAPTER II

Efficient uncertainty quantification of hydrologic predictions with a surrogate model

> "You don't have to be great to start, but you have to start to be great"
>
> - *(Ziglar, Z)*

2.1 Introduction

Extreme floods occur more frequently than in the past due to climate warming, and they have more profound socio-economic impacts [*Hirabayashi et al.*, 2013; *Winsemius et al.*, 2015]. Providing highly accurate predictive information and presenting warning messages in a timely manner play a key role in mitigating the risk of floods. Hydrologic, hydraulic, or coupled models are the most commonly used methods to predict flooding phenomena. Each model has adopted various representations of the physical processes for water flow over an entire area [*Smith et al.*, 2004; *Vrugt et al.*, 2005; *Kim et al.*, 2013; *Maxwell et al.*, 2014; *Kim and Ivanov,* 2015]. However, these predictive models always involve uncertainty due to a lack of knowledge of the involved physical processes and their interactions, the infeasibility of identifying model parameters, and difficulties in measuring (or estimating) initial and current states [*Beven*, 1989; *Butts et al.*, 2004; *Ajami et al.*, 2007; *Moradkhani and Sorooshian*, 2008; *Ivanov et al.*, 2010; *Kim et al.*, 2012a; *DeChant and Moradkhani*, 2014; *Kim and Ivanov*, 2014; *Kim et al.*, 2016a; *Mockler et al.*, 2016]. Thus, quantifying and reducing uncertainties has been a major challenge for researchers in flood prediction and in water planning and supply, sediment management, and reservoir operation [*Faber and Stedinger*, 2001; *Todini*, 2004; *Benke et al.*, 2008; *Saad and Ghanem*, 2009; *Kim et al.*, 2016c; *Kim et al.*, 2016b].

Numerous approaches have been proposed to quantify the uncertainty associated with model parameters in hydrologic prediction. First, a number of optimization techniques (such as downhill simplex [*Nelder and Mead*, 1965], the shuffled complex evolution (SCE-UA) method [*Duan et al.*, 1992], and the particle swarm optimization [*Kennedy and Eberhart*, 1995]) have been developed to find the single best fitting parameter set. They were successfully used in diverse engineering applications, but the calibration techniques still lack the ability to properly treat the various uncertainties inherent in the system [*Moradkhani et al.*, 2005a]. Furthermore, although a set of parameters obtained from a watershed best represents the behavior of the basin, it may not work for other watersheds [*Beven*, 1989; *Beven and Binley*, 1992; *Moradkhani and Sorooshian*, 2008]. Failure to identify a range of parameter values may increase uncertainty in the model outputs [*Moradkhani and Sorooshian*, 2008].

Second, data assimilation methods such as the ensemble Kalman filter (EnKF) [*Evensen*, 1994] and the particle filter (PF) [*Arulampalam et al.*, 2002] also have received much attention, especially in real-time forecasting. This is because these techniques continuously update model states and parameters whenever new observations are available to improve model predictability [*Vrugt et al.*, 2005; *Liu et al.*, 2012; *Moradkhani et al.*, 2012]. However, some parameters of hydrologic models may not be completely identifiable, and therefore do not show convergence during the assimilation process for complex domains [*Moradkhani et al.*, 2005a; *Moradkhani et al.*, 2012]. Moreover, the potential collapse of EnKF (in which all ensemble members result in a similar value) can be resolved by updating each ensemble member with an independently perturbed observation [*Burgers et al.*, 1998]; in other words, the magnitudes of model states or parameters cannot converge to the values corresponding to observations even after many repeated attempts at the assimilation process. The PF method has an advantage over EnKF in terms of reducing numerical instability because it provides particle weights and uses non-Gaussian state-space models [*Liu et al.*, 2012]. On the other hand, the PF method is computationally more expensive than EnKF, as it generally requires more ensemble members based on the sequential Monte Carlo method [*Moradkhani et al.*, 2005a; *Liu et al.*, 2012].

Alternative probabilistic methods have been developed to deal with uncertainty quantification. They are mostly based on the Monte Carlo (MC) procedure, which provide the posterior distribution of parameters. The most common probabilistic methods include: Generalized Likelihood Uncertainty Estimation (GLUE) [*Beven and Binley*, 1992], Bayesian recursive estimation technique (BaRE) [*Thiemann et al.*, 2001], the Metropolis method [*Kuczera and Parent*, 1998], the Shuffled Complex Evolution Metropolis (SCEM-UA) [*Vrugt et al.*, 2003a; *Vrugt et al.*, 2003b], and the DiffeRential Evolution Adaptive Metropolis (DREAM) scheme [*Vrugt et al.*, 2008b]. The GLUE method attempts to identify a variety of parameter sets (namely "behavioral parameter set") given likelihood functions and cutoff threshold values [*Moradkhani and Sorooshian*, 2008]. The BaRE approach can simultaneously perform parameter estimation and hydrologic prediction. Uncertainties associated with parameter estimates are updated recursively, and uncertainty in output predictions becomes smaller when measurements are successively assimilated [*Thiemann et al.*, 2001]. The Metropolis method developed by *Kuczera and Parent* [1998] uses a random walk that adapts to a true probability distribution to describe parameter uncertainty. The SCEM-UA is the extension of the SCE-UA algorithm, which combines the strengths of the Metropolis algorithm, controlled random search, competitive evolution, and complex shuffling to continuously update a proposal distribution and evolve it into a posterior target distribution [*Vrugt et al.*, 2003b]. The DREAM scheme is an adaptation of the SCEM-UA using a novel Markov Chain Monte Carlo (MCMC) sampler. It shows excellent efficiency for complex, highly nonlinear, and multimodal target distributions [*Vrugt et al.*, 2008b].

It is generally necessary to increase the number of repeated model runs to successfully capture the uncertainty of model predictions using the probabilistic methods. Therefore, the computational time must increase [*Ballio and Guadagnini*, 2004; *Herman et al.*, 2013]. One of the methods that can offset the increase in time required to calculate uncertainty quantification is to use a parallel computing technique [*Vrugt et al.*, 2006b; *Vrugt et al.*, 2008a]. However, the downside of this is that the computer hardware configuration requirements lead to expensive cost [*Cintra and Velho*, 2018]. A surrogate model (also called metamodel, response surfaces, data-driven model, or model emulator) has been proposed to address the issues that the probabilistic methods involve. Its main objective is to provide nearly equivalent results to those of the original model, quantify the degree of uncertainty more quickly, and effortlessly evaluate the sensitivity of model parameters [*Gerstner*, 1998; *Xiu and Karniadakis*, 2002; *Berveiller et al.*, 2006; *Blatman and Sudret*, 2010; *Blatman and Sudret*, 2011; *Oladyshkin and Nowak*, 2012; *Schobi and Sudret*, 2014]. Arising from a variety of disciplines, various surrogate models have been developed and implemented for water resources problems, such as Gaussian process, artificial neural networks, support vector machines, and polynomial chaos expansion [*Baú and Mayer*, 2006; *Razavi et al.*, 2012a; *Sargsyan et al.*, 2014; *Christelis and Hughes*, 2018; *Wang et al.*, 2018; *Dwelle et al.*, 2019; *Hu et al.*, 2019a; *Rajabi*, 2019; *Tran and Kim*, 2019; *Tran et al.*, 2020; *Wang et al.*, 2020; *Zhang et al.*, 2020]. Among the surrogate models, the polynomial chaos expansion (PCE) method [*Wiener*, 1938; *Ghanem and Spanos*, 1991] has drawn lots of attention, especially in hydrology studies where quantifying the uncertainty of model parameters has been a major issue [*Rajabi*, 2019]. Those studies are performed specifically for flood prediction [*Fan et al.*, 2014; *Wang et al.*, 2015; *Wang et al.*, 2017], subsurface flow [*Laloy et al.*, 2013; *Sochala and Le Maître*, 2013; *Meng and Li*, 2018], and groundwater dynamics [*Laloy et al.*, 2013; *Asher et al.*, 2015]. PCE can mimic the non-linear behaviors, characteristics of the complex physical model and provide global sensitivity analysis easily [*Dwelle et al.*, 2019]. Also, the use of PCE can noticeably offset the computational costs necessary for simulating a great number of ensemble runs, allowing for the uncertainty quantification to take place, even in real-time [*Tran et al.*, 2020]. However, they did not present a way of optimizing their PCE constructions to quantify predictive uncertainty, as well as details of coupling with GLUE [e.g., *Ciriello et al.*, 2012; *Baştuğ et al.*, 2013; *Fan et al.*, 2014; *Wang et al.*, 2015; *Fan et al.*, 2016].

In this Chapter, we aim to propose a unified framework for (i) quantifying the parametric uncertainty of a conceptual rainfall-runoff (CRR) model integrated with the GLUE framework, (ii) quantifying the same uncertainty of a surrogate model built with PCE and identifying changes in efficiency and accuracy as compared to the deterministic model, and (iii) assessing the sensitivity of PCE parameters to identify whether the parameter lower/upper bounds can be further constrained; and then to (iv) discuss challenges existed in the use of PCE and examine/point out an efficient surrogate model for the original model. Wherein, the first three objectives of the unified framework is presented in Section 2.2, while Section 2.3 focuses on the last concerning.

2.2 Uncertainty quantification of hydrologic predictions with a surrogate model

2.2.1 Methods

2.2.1.1 Polynomial chaos expansion

Polynomial chaos expansion (PCE) [*Wiener*, 1938; *Ghanem and Spanos*, 1991] is one of the propagation uncertainty methods and a powerful meta-modeling technique. The PCE method aims to provide a functional approximation of a computational model through its spectral representation with a suitably built basis of polynomial functions. In this dissertation, a surrogate model based on the PCE (so-called PCE model) is constructed to mimic a conceptual hydrologic model and quantify the parametric uncertainty of the model simulation.

The following is a brief summary of PCE theory [*Sudret*, 2008; *Blatman and Sudret*, 2010]. Consider a deterministic rainfall-runoff model denoted by $\boldsymbol{\mathcal{M}}$. The input space of the model is represented by random vectors of input parameters $\boldsymbol{\theta}$, and the model response, y (i.e., streamflow) is:

$$y = \boldsymbol{\mathcal{M}}(\boldsymbol{\theta}) \tag{2.1}$$

The goal of this theory is to approximate the computational model, $\boldsymbol{\mathcal{M}}$ with the PCE model, $\boldsymbol{\mathcal{M}}^{PCE}(\boldsymbol{\theta})$. The latter is computed with a finite sum of orthonormal polynomials for the input parameters.

$$y \approx \boldsymbol{\mathcal{M}}^{PCE}(\boldsymbol{\theta}) = \sum_{\alpha=0}^{N_{\Psi}-1} \varepsilon_{\alpha} \Psi_{\alpha}(\boldsymbol{\theta}) \tag{2.2}$$

where ε_{α} is PCE model coefficients to be determined for all multi-indices; $\Psi_{\alpha}(\boldsymbol{\theta})$ are the corresponding multivariate orthonormal polynomials given as the input parameters; $\boldsymbol{\alpha}$ is a multi-index that identifies the components of the multivariate polynomials; N_P is the number of deterministic model parameters $\boldsymbol{\theta} = \{\theta_j, j = 1, \ldots, N_P\}$; and N_Ψ is the number of PCE coefficients (i.e., the number of polynomial expansion basis terms) determined by N_P and the polynomial degree p as:

$$N_\Psi = \frac{(N_P + p)!}{N_P! p!} \qquad (2.3)$$

The multi-dimensional polynomials are constructed as the product of univariate orthonormal polynomials:

$$\Psi_{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \prod_{j=1}^{N_P} \Psi_{\alpha_j}^{(j)}(\theta_j) \qquad (2.4)$$

Here, $\Psi_{\alpha_j}^{(j)}$ is the univariate orthonormal polynomials of the *j*-th parameter of degree α_j.

Depending on the probabilistic characteristics of $\boldsymbol{\theta}$, different polynomial bases can be used for $\Psi_{\boldsymbol{\alpha}}(\boldsymbol{\theta})$. Polynomial basis functions based on the Weiner-Askey scheme [*Xiu and Karniadakis*, 2002] are illustrated in Table 2.1 for the commonly-used distributions of random variables.

Table 2.1. Polynomial basis functions for probability distributions of uncertain parameters

Distribution	**Orthogonal polynomial family**	**Support**
Uniform	Hermite	(low, up)
Gaussian	Legendre	$(-\infty, \infty)$
Gamma	Laguerre	$[0, \infty)$
Beta	Jacobi	(low, up)

*low and up denote the lower and upper bounds of uncertain parameters, respectively.

When constructing the PCE model, one of the important steps is to compute the PCE coefficients (ε_α). Determining PCE coefficients generally depends on the number of the training set (so called *experiment design*) (N) and the polynomial degree (p) [*Blatman and Sudret*, 2010;

Blatman and Sudret, 2011]. Increasing these numbers requires a lot of computational resources (e.g., Table 2.2), and this tendency will be accelerated if more complicated models are simulated over a complex domain [*Sudret*, 2008].

Table 2.2. Time required for computing PCE coefficients depending on the number of (a) experimental design (*N*) with p of 3, and (b) polynomial degree (*p*) with *N* of 50, for 3 flooding events (in seconds) (for information about flood events, see Section 2.2.2)

(a) for Experiment design (*N*)				**(b) for Polynomial degree (*p*)**			
N	Event 2	Event 5	Event 8	*p*	Event 2	Event 5	Event 8
10	1.3	1.3	7.2	1	1.4	1.4	7.9
50	2.0	2.1	11.3	2	1.4	1.4	9.1
100	4.0	4.1	22.4	3	1.4	1.5	10.8
200	5.5	5.6	30.7	4	2.1	1.8	12.3
500	7.9	7.9	43.2	5	3.2	2.8	17.0
700	9.2	9.4	53.9	7	11.8	11.3	58.1
1000	11.3	12.1	63.1	10	84.2	89.0	415.0
2000	30.9	39.9	165.7	12	328.4	332.8	1455.1

The projection method [*Ghiocel and Ghanem*, 2002; *Le Maître et al.*, 2002] is one of the methods used to compute the PCE coefficients. Therein, N is simply determined from a mathematical equation with the polynomial degree, p, and the number of model parameters, N_P, i.e., N=(p+1)^{N_P}. For example, if p is 3 and N_P is 9, *N* will be equal to 262,144. Repeating the original model based on these numbers takes a considerable amount of time. To reduce such a large computational time, the ordinary least square regression (OLS) method is generally employed in which N can be given by the researcher [*Berveiller et al.*, 2006; *Sudret*, 2008; *Blatman and Sudret*, 2010]. Specifically, the PCE coefficients can be estimated by the regression method as follow:

$$\varepsilon \; = \operatorname{argmin}_{\varepsilon\in\mathbb{R}^{|A|}} \mathbb{E}[(y - \textstyle\sum_{\boldsymbol{\alpha}\in A} \varepsilon_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{\theta}))^2] \qquad (2.5)$$

Given a collection $\boldsymbol{\mathcal{X}} = \{\mathcal{X}^{(1)}, \dots, \mathcal{X}^{(N)}\}$ consisting of the number of *N* sets of the parameters $\boldsymbol{\theta}$ (the set $\boldsymbol{\mathcal{X}}$ is called the *experimental design*), $\boldsymbol{\mathcal{Y}} = \{\boldsymbol{\mathcal{M}}(\mathcal{X}^{(1)}), \dots, \boldsymbol{\mathcal{M}}(\mathcal{X}^{(N)})\}$ is the corresponding model evaluation $\{\mathcal{Y}^{(k)} = \boldsymbol{\mathcal{M}}(\mathcal{X}^{(k)}), k = 1, \dots, N\}$. The estimates of the PCE coefficients are thus given by:

$$\hat{\varepsilon} \; = \; \operatorname{argmin}_{\varepsilon \in \mathbb{R}^{|A|}} \frac{1}{N} \sum_{k=1}^{N} \left(\mathcal{Y}^{(k)} - \sum_{\boldsymbol{\alpha} \in A} \varepsilon_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\mathcal{X}^{(k)}) \right)^2 \tag{2.6}$$

which is equivalent to:

$$\hat{\varepsilon} = (\mathbf{F}^{\mathbf{T}}\mathbf{F})^{-1}\mathbf{F}^{\mathrm{T}}\mathcal{Y} \tag{2.7}$$

where **F** is so-called the *information matrix* of size $N \times |A|$ whose generic term reads:

$$\mathbf{F}_{k,\boldsymbol{\alpha}} = \Psi_{\boldsymbol{\alpha}}(\mathcal{X}^{(k)}) \quad\quad k = 1, \dots, N; \; \boldsymbol{\alpha} = 0, \dots, N_{\Psi} - 1 \tag{2.8}$$

Also, note that the effect of *N* on PCE results has not been investigated in the literature [e.g., *Ciriello et al.*, 2012; *Baştuğ et al.*, 2013; *Fan et al.*, 2014; *Wang et al.*, 2015; *Fan et al.*, 2016] – this will be addressed in Sec. 2.2.3.2 in more detail. Additionally, the p value can be approximated depending on the complexity of model outputs and by the subjectivity of the researcher, where most common values used in literature are 2 or 3 [e.g., *Sochala and Le Maître*, 2013; *Fan et al.*, 2014; *Wang et al.*, 2015; *Wang et al.*, 2017]. In this work, the least squares regression method [*Berveiller et al.*, 2006; *Sudret*, 2008; *Blatman and Sudret*, 2010; *Blatman and Sudret*, 2011] was employed.

2.2.1.2 Deterministic rainfall-runoff model: NAM

Conceptually, lumped rainfall-runoff models treat the whole catchment as a uniform unit so that single representative values are used over the catchment for all inputs and parameters [*Moradkhani and Sorooshian*, 2008]. These conceptual rainfall-runoff models are computationally efficient, and therefore they are widely used in various studies. For this research, the NAM (Nedbør - Afstrømnings Model) was employed. The NAM [*Nielsen and Hansen*, 1973] is one of the widely used deterministic, lumped models that simulates streamflow in the world. It is considered to be a very useful and flexible model and has been applied to many study regions [*Madsen*, 2000; *Butts et al.*, 2004; *Thompson et al.*, 2004; *Liu et al.*, 2007; *Makungo et al.*, 2010; *O'Brien et al.*, 2013; *Mockler et al.*, 2016]. [*Madsen*, 2000; *Butts et al.*, 2004; *Thompson et al.*, 2004; *O'Brien et al.*, 2013; *Mockler et al.*, 2016]. Specifically, its design assumes three different and mutually integrated storages representing a surface zone, lower zone, and routing components that simulate overland flow, interflow, and base flow, respectively. The model requires two input forcing variables (N_I) of spatially averaged precipitation and evapotranspiration, five model states ($N_S = 5$), and nine model parameter values ($N_P = 9$) listed in Table 2.3 [*DHI*, 2014]. The latter states and parameters control the amount of water content and the rates of release from the conceptualized storage compartments of the model. Because evapotranspiration is assumed to be negligible during the rainy season with flooding events, the number of inputs used in this study is 1 ($N_I = 1$). For more detail, readers can refer to *DHI* [2014].

Table 2.3. Description of the NAM model states and parameters

		Unit	Description	Lower Bound	Upper Bound
States	U	mm	Water content in surface storage	0	35
	L	mm	Water content in lower zone/root storage	0	400
	OF	m^3/s	Overland flow	0	$+\infty$
	IF	m^3/s	Inter flow	0	$+\infty$
	BF	m^3/s	Base flow	0	$+\infty$
Parameters	Um	mm	Maximum water content in surface storage	5	35
	Lm	mm	Maximum water content in lower zone/root storage	50	400
	CQOF	[-]	Overland flow coefficient	0	1
	CKIF	hrs	Interflow drainage constant	200	2000
	TOF	[-]	Overland flow threshold	0	0.9
	TIF	[-]	Interflow threshold	0	0.9
	TG	[-]	Groundwater recharge threshold	0	0.9
	CK12	hrs	Time constant for routing interflow and overland flow	3	72
	CKBF	hrs	Time constant for base flow	500	5000

2.2.1.3 Uncertainty quantification: GLUE

The uncertainty of hydrologic models is influenced by various sources such as model input, initial conditions, boundary conditions, and model parameters. In general, it is difficult to estimate all these influencing factors in realistic conditions. Thus, even within the same model, a large number of parameter combinations may provide the same model results [*Beven*, 1989]. Simulation results obtained by models are subject to various uncertainties. This is not uncommon, especially in lumped hydrologic models that use homogeneous parameters for the entire domain. In this work, the GLUE method [*Beven and Binley*, 1992] was chosen to quantify the uncertainty of the lumped NAM caused by the parameter uncertainty.

There are two reasons for selecting this method. Compared to other methodologies, GLUE is straightforward to implement, and it allows flexibility in the definition of the likelihood function used to separate *behavioral* and *nonbehavioral* parameter sets [*Beven*, 2006; *Blasone et al.*, 2008a; *Beven and Binley*, 2014]. The behavioral sets refer to any random combination of parameters that qualify the preset criteria of likelihood functions among all combinations of the parameters. Then, "accepted behavioral runs" are defined as those simulated with the behavior parameter sets. Another advantage of GLUE is that it is designed to be non-intrusive, meaning that one should not need to modify any of the existing source codes in the deterministic models [*Vrugt et al.*, 2008c].

The GLUE method includes several steps. First, a deterministic model was simulated with the parameter sets randomly sampled from prior distributions of parameters. In this study, the initial values of parameters are extracted using the Monte Carlo method from uniform distributions constrained with the potential ranges (see Table 2.3 where upper and lower bounds are illustrated, which were adopted from *DHI* [2014]). Then, the performance of each model run is evaluated by choosing any likelihood function. As a result, only a part of the runs can be selected as the behavioral run. The stricter the likelihood function condition, the more accurate the simulation result, and the smaller the number of behavioral runs. Last, we quantified the uncertainty from the GLUE behavioral runs of both NAM and PCE model (Fig. 2.1).

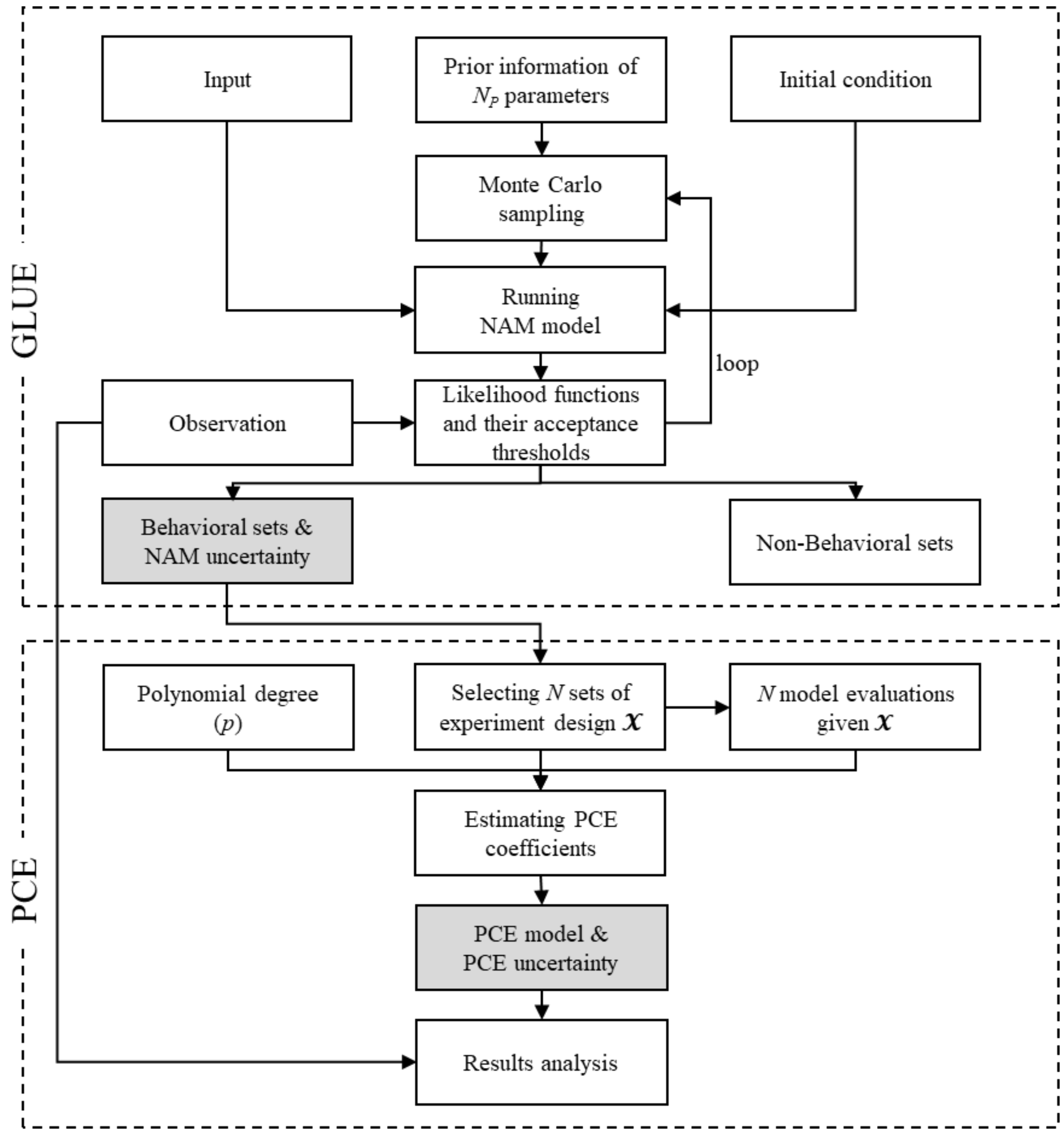

GLUE
Input
Prior information of N_P parameters
Initial condition
Monte Carlo sampling
Running NAM model
loop
Observation
Likelihood functions and their acceptance thresholds
Behavioral sets & NAM uncertainty
Non-Behavioral sets
PCE
Polynomial degree (p)
Selecting N sets of experiment design $\mathcal{X}$
N model evaluations given $\mathcal{X}$
Estimating PCE coefficients
PCE model & PCE uncertainty
Results analysis

Figure 2.1. The unified framework of the hydrologic model uncertainty quantification using PCE and GLUE.

Although the GLUE has been adopted in many studies, the main difficulty of this method is the subjectivity of the selections of the likelihood function and cutoff threshold [*Montanari*, 2005; *Beven*, 2006; *Mantovan and Todini*, 2006; *Freni et al.*, 2008; *Stedinger et al.*, 2008; *Xiong and O'Connor*, 2008; *Freni et al.*, 2009b; *Freni et al.*, 2009a; *Li et al.*, 2010; *Mirzaei et al.*, 2015]. First, the choice of the likelihood function plays a crucial role because it lays the foundation for determining the behavioral parameter sets [*Beven and Binley*, 1992]. Various likelihood functions have been proposed in many works [*Beven and Binley*, 1992; *Romanowicz et al.*, 1994; *Christensen*, 2004; *Montanari*, 2005; *Moriasi et al.*, 2007a], which quantify the closeness between observations and model simulations. The Nash – Sutcliffe efficiency (NSE) has been used most often [*Freer et al.*, 1996; *Gupta et al.*, 1998; *Madsen*, 2000; *Uhlenbrook and Sieber*, 2005; *Kuczera et al.*, 2006; *Freni et al.*, 2008; *Stedinger et al.*, 2008; *Gupta et al.*, 2009; *Franz and Hogue*, 2011]. However, several indices with different functions have been introduced in other studies to evaluate the model simulations including the daily root mean square estimation, the heteroscedastic maximum likelihood estimation for daily mean flow [*Yapo et al.*, 1996], the peak runoff-runoff volume index [*Hossain and Anagnostou*, 2005], or a likelihood measure based on the sum of the absolute errors for discharge [*Choi and Beven*, 2007]. In this work, 3 metrics for the purpose of simulating a flood phenomenon were focused, namely, Nash – Sutcliffe efficiency (NSE), peak error (PE), and volume error (VE), which represent the shape, peak, and volume of the flood hydrograph, respectively [*Kim et al.*, 2012b].

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{\mathrm{T}} (y_t^{obs} - y_t)^2}{\sum_{t=1}^{\mathrm{T}} (y_t^{obs} - \overline{y^{obs}})^2} \tag{2.9}$$

$$\mathrm{PE} = \frac{|y_{max}^{obs} - y_{max}|}{y_{max}^{obs}} \times 100 \tag{2.10}$$

$$\mathrm{VE} = \frac{|V^{obs} - V|}{V^{obs}} \times 100 \tag{2.11}$$

Here, y_t^{obs} and y_t are observed and simulated streamflow at time t, respectively; T is the total duration of a rainfall event; y_{max}^{obs} and y_{max} are observed and simulated streamflow at the peak time of the event, respectively; and V^{obs} and V are the total volume of observed and predicted hydrograph, respectively.

Another weakness of the GLUE is its dependency on the arbitrary selection of acceptance thresholds for the behavioral parameter sets. Acceptance threshold values for the selected likelihood functions should be determined beforehand to be qualified as the behavioral set, and such a selection strongly influences the model results [*Freni et al.*, 2008; *Li et al.*, 2010]. Universal values for behavioral acceptance thresholds accepted in every study have not yet been reported. The values were rather examined separately, often relying on a researcher's experience [*Beven and Binley*, 1992; *Romanowicz et al.*, 1994; *Freer et al.*, 1996; *Zak and Beven*, 1999; *Blasone et al.*, 2008a; *Beven and Binley*, 2014]. In this dissertation, two indices representing an accuracy and efficiency are proposed to identify how to quantitatively determine the cutoff values of likelihood functions including NSE, PE, and VE. The first index, named the accuracy index (AI), calculates the temporal average (U) of the GLUE uncertainty identified here as the range between the 2.5th and 97.5th percentiles of the 1,000 behavioral ensemble outcomes over the entire computation time in the hydrograph, expressed in Eq. (2.12). This computation is repeated for varying threshold values of each likelihood function (i.e., id = 1 to 17 corresponding to the range from 0.5 to 0.9 in 0.025 increments for NSE; id = 1 to 21 from 1 to 51% at an interval of 2.5% for PE and VE). Finally, the accuracy index is computed from Eq. (2.13), which is defined based on the ratio of temporal averages (U_{id}) to the maximum (U_{max}) value.

$$U = \frac{1}{T}\sum_{t=1}^{T}(y_t^{97.5} - y_t^{2.5}) \tag{2.12}$$

$$AI_{id} = 1 - \frac{U_{id}}{U_{max}} \tag{2.13}$$

Here, id is an index corresponding to varying threshold values. $y_t^{97.5}$ and $y_t^{2.5}$ are the discharge values corresponding to the 2.5th and 97.5th percentiles of the uncertain distribution at time t. The other index, named the efficiency index (EI), defined in Eq. (2.14), first counts the number of model runs (Q) required until 1,000 behavioral sets from the GLUE procedure are obtained. Then, Q is computed for each it, i.e., for the identical varying threshold values of each likelihood function. Finally, the efficiency index is estimated by using the ratio to the maximum number of model runs (P_{max}) among the numbers (P_{id}). The values of AI and EI vary from 0 and 1.

$$EI_{id} = 1 - \frac{P_{id}}{P_{max}} \tag{2.14}$$

Two indices proposed above are designed to have opposite tendencies – such a characteristic has an advantage in determining the cutoff threshold values. One could pay attention to the intersection point of these two index curves as an optimal value that meets some degree of accuracy and efficiency at the same time, which is termed *behavioral acceptance threshold*.

2.2.1.4 Sensitivity analysis: Sobol' indices and Morris methods

Sensitivity analysis (SA) is used to evaluate how much each parameter contributes to the output uncertainty and allows for identification of important parameters that dominate model behavior [*Saltelli*, 2002b]. Generally, SA can be categorized into two groups of local and global SA. The local SA computes the changes in the model simulation by changing one parameter while keeping other parameters constant. However, it is often unable to produce meaningful results [*Saltelli et al.*, 2004; *Jiang et al.*, 2015]. On the other hand, the global SA investigates the changes in the model by varying all parameters simultaneously. There are a couple of global SA methods that are widely used, such as Fourier amplitude sensitivity test (FAST) [*Cukier et al.*, 1973], Morris one-at-a-time screening (MOAT) [*Morris*, 1991], Sobol' sensitivity indices [*Sobol'*, 1993], response surface methodology (RSM) [*McKay et al.*, 1979a], etc. In this work, we performed a global sensitivity analysis based on Morris one-at-a-time screening and Sobol' sensitivity indices.

In the abundant literature on sensitivity measures [*Sudret*, 2008], the Sobol' indices have received much attention since they provide accurate information for most models. Sobol' indices are a variance-based sensitivity analysis that identifies parameter sensitivities by evaluating the variance of model output (y) due to the variability of individual parameters and their parameter interactions [*Sobol'*, 2001; *Saltelli*, 2002b; *Crestaux et al.*, 2009]. Instead of the model output y, model performance measures (e.g., NSE, PE, and VE) can be used as an objective function to quantify the sensitivity indices [*Tang et al.*, 2007a]. Sobol' indices correspond to variance-based decomposition, as they measure fractional contributions of each parameter or group of parameters towards the total output variance. Specifically, the total variance, $D(y)$ is decomposed as:

$$D(y) = \sum_{a=1}^{M_P} D_a + \sum_{a<b} D_{ab} + \cdots + D_{1 \ldots M_P} \tag{2.15}$$

where D_a is the variance of y due to the changes of a-th model parameter, $\boldsymbol{\theta}_a$, denoting the first-order contribution to $D(y)$; D_{ab} is the variance of y due to the pairwise interactions of a-th and b-th parameters, referring to the second-order contribution. The first (S_a), total-order ($S_{Total,a}$), and second-order (S_{ab}) Sobol' sensitivity indices can be respectively expressed as:

$$S_a(y) = \frac{D_a(y)}{D(y)} \tag{2.16}$$

$$S_{Total,a}(y) = 1 - \frac{D_{\tilde{a}}(y)}{D(y)} \tag{2.17}$$

$$S_{ab}(y) = \frac{D_{ab}(y)}{D(Ly)} \tag{2.18}$$

where $D_{\tilde{a}}$ is the variance averaged over the contributions resulting from all parameters except for $\boldsymbol{\theta}_a$.

To analyze the sensitivity of NAM parameters, we outline two sensitivity indices, including the first-order and total-order Sobol' indices. The computational requirements to evaluate the first and total order sensitivity indices is $n_s \times (N_P + 2)$ model runs, where n_s denotes the number of samples of each parameter for which the indices are to be calculated [*Saltelli*, 2002a].

Besides, the Morris method is also used to confirm Sobol' indices. The Morris method (MOAT) is designed to work with low computational cost to determine which parameters are (i) negligible, (ii) linear and additive, and (iii) nonlinear or involved in interactions with other parameters [*Jiang et al.*, 2015]. Herein, we used the improved version of *Campolongo et al.* [2007], with n_m samples of each parameter; this method requires a total of $n_m \times (N_P + 1)$ simulations, resulting in two sensitivity measures for each parameter: the mean (μ^*) and standard deviation (σ) values. A high value of μ^* indicates a parameter that has a significant effect on the model output, and the value of σ indicates that either the parameter is interacting with other parameters or it has non-linear effects on the model output.

After the surrogate model (PCE model) was developed, we used the MC method to conduct the sensitivity analysis to input uncertain parameters [*Sobol'*, 2001]. The number of sample of parameters is a key component for both the Morris and Sobol' methods. A rough rule of thumb about the number of model evaluations is that at least 10 × N_P sample points are needed to identify the component factors (i.e., parameters) [*Levy and Steinberg*, 2011]. In the Sobol' method, many previous studies used a large sample size to analyze the sensitivity of model parameters, e.g., 8,192 sets were used to analyze the sensitivity of 18 model parameters [*Tang et al.*, 2007b] while 500,000 sets were used for 14 model parameters [*van Werkhoven et al.*, 2009]. On the other hand, there are several studies using a relatively small sample size, which proved sufficient to maintain the accuracy and repeatability of Sobol' analysis, e.g., a sample size of 2000 was used for 21 [*Fu et al.*, 2012], 13 [*Tang et al.*, 2007a], and 28 model parameters [*Zhang et al.*, 2013]. Therefore, on the basis of prior studies and our experiment, a sample size n_s of 2,000 was used, which requires 2000 × (9 + 2) = 22,000 model runs to analyze the Sobol' indices. For the Morris method, a typical n_m found in the literature is 10 [*Neumann*, 2012; *Jiang et al.*, 2015], and thus a total number of 10 × (9 + 1) = 100 model runs were applied.

2.2.2 Case study

In this research, we chose the 'Thu Bon' basin located in central Vietnam as the study area (Fig. 2.2) because it is known as a region vulnerable to flood. The watershed belongs to a tropical, continental monsoon region. Thus, this region has experienced intense rainfall, severe floods and significant damage. For example, the flood event in 1999 resulted in total damage of 29 million USD, 53 deaths, and 3,500 hectares of damaged fields [*UNDP*, 1999]. This watershed drains into the upstream part of the 'Thu Bon' River, which is one of the sub-basins of the 'Vu Gia-Thu Bon' river basin. The basin has a catchment area of 3,208 km^2, a mainstream length of 105 km, and altitudes ranging from 15 to 2,530 m ('Nong Son' in Fig. 2.2). Since the slope of the terrain is very steep (approximately 22.4% in average), the average annual precipitation is larger than 2,000 mm/year, and most of the rain falls from September to December (rainy season). Floods occur rapidly and frequently in this timeframe. The flow peak in the flooding events is averaged over 5,000 m^3/s, and the largest flow peak observed was 10,000 m^3/s in 1999 at the outlet of 'Nong Son'. Streamflow data used for the outlet was observed hourly at the only hydrometric station within the domain. Rainfall data were also observed hourly and obtained from the four weather stations near the study area. The average rainfall over the basin was calculated using the Thiessen polygon method (Table 2.4).

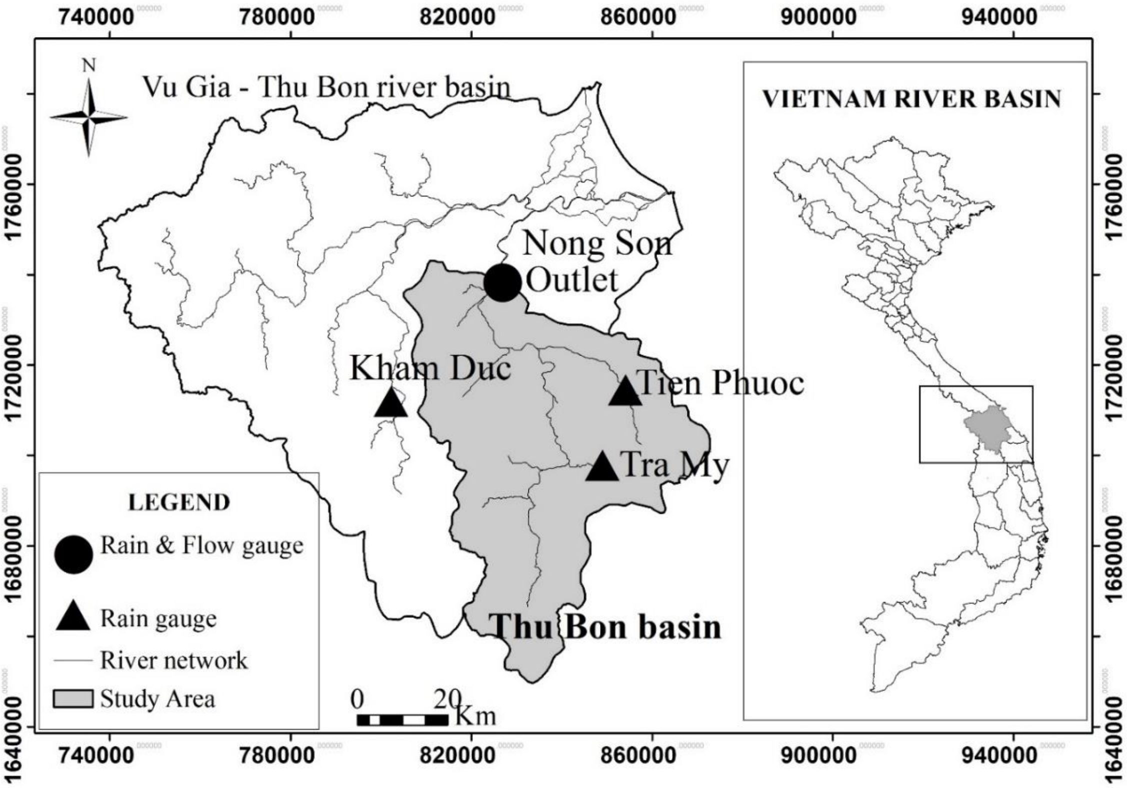


Figure 2.2. Study area: Thu Bon basin

Table 2.4. General information of meteorological and hydrometric stations

Station	Measurement	Coordinate of stations		Areal weights for Thiessen polygons
		Latitude	Longitude	
Tra My	Rain	15°19'60"N	108°15'0"E	0.471
Tien Phuoc	Rain	15°28'60"N	108°17'60"E	0.188
Kham Duc	Rain	15°28'0"N	107°49'0"E	0.158
Nong Son	Rain, Flow	15°41'60"N	108° 3'0"E	0.183

After inspecting the available data, eight flood events were specifically selected in this work (Table 2.5, Figs. 2.3 and 2.4). Those events were chosen to include flood events corresponding to various (low, middle, and high) return periods based on a frequency analysis of flood events (Fig. 2.3). NAM was used to model the Thu Bon watershed for the 8 selected flood events. A warm-up simulation of 10 hours was additionally performed before starting to collect results. Results for only 3 flood events will be illustrated in the main thesis for simplicity, while the rest of the results are included in the Appendix A.

Table 2.5. Characteristics of selected flood events

Event	Time [DD/MM/YYYY]	Flood peak [m^3/s]	Flood frequency [%]	Flood volume [million m^3]	Total rainfall [mm]	Duration [hours]
1	13/09/2015-16/09/2015	1408	92.8	169.8	176.8	84
2	15/10/2015-17/10/2015	893	97.6	135.4	160.3	71
3	01/11/2015-07/11/2015	2508	85.7	645.5	452	162
4	24/11/2015-01/12/2015	1173	95.2	375.8	294.9	170
5	12/09/2016-14/09/2016	3243	73.8	312	297.6	72
6	31/10/2016-06/11/2016	2801	78.6	686	401.1	166
7	31/11/2016-08/12/2016	6730	33.3	1,969.8	786	215
8	10/12/2016-22/12/2016	8169	16.7	2,436.8	848.5	300

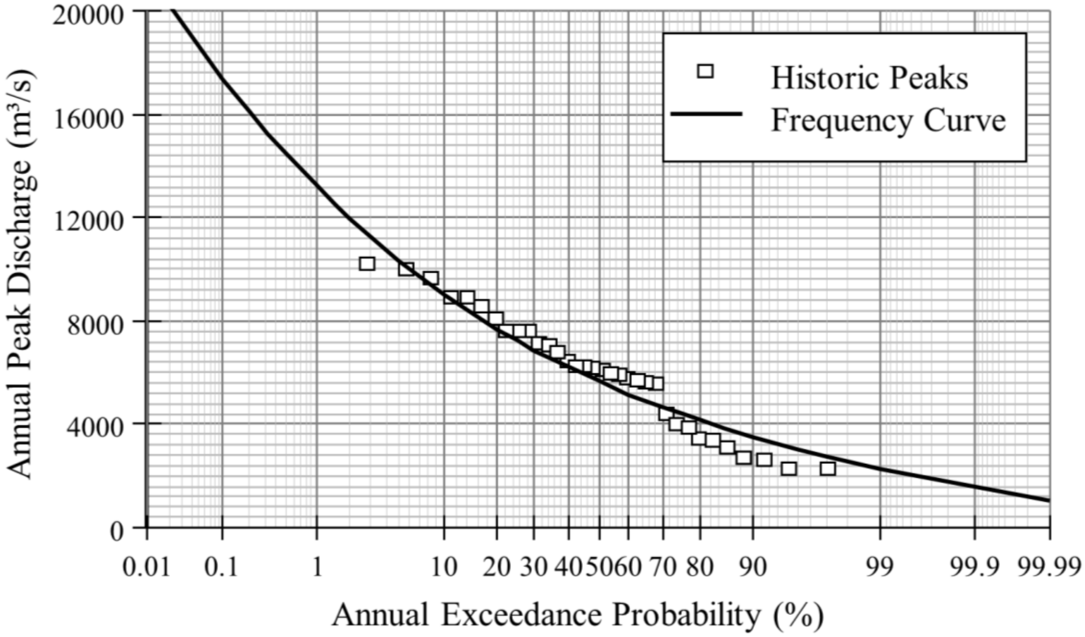


Figure 2.3. Flood frequency curve for the ‘Nong Son’ station; historic peaks refer to annual maximum peak flows from 1978 to 2016; the flood frequency curve is fitted using the Pearson Type III distribution.

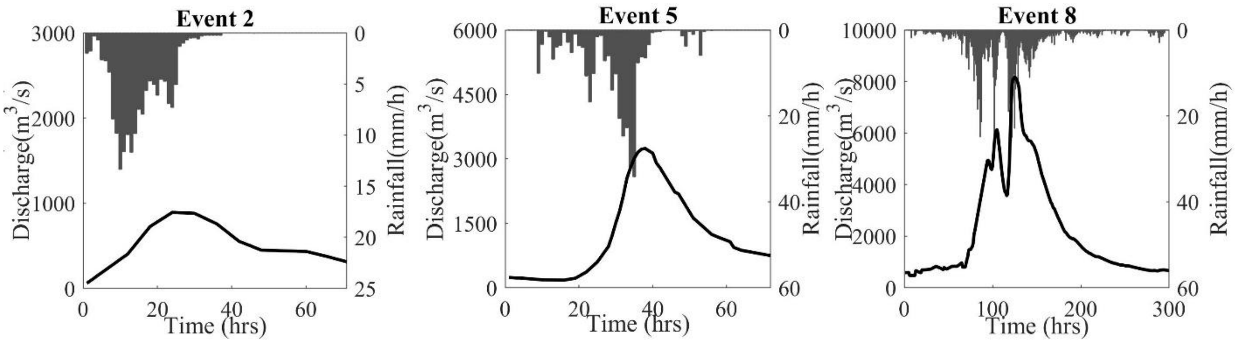


Figure 2.4. Observation of rainfall (histograms using the right axis) and discharge (lines using the left axis) for 3 flooding events corresponding to small, medium, and large return periods, respectively.

2.2.3 Results

2.2.3.1 Behavioral acceptance thresholds and the number of behavioral sets of GLUE

To determine the cutoff threshold values, we designed two indices (AI and EI) with opposite tendencies. Fig. 2.5 noticeably shows that the accuracy index increases while the efficiency index decreases as the conditions of each likelihood function are tightened. The *behavioral acceptance threshold* values for the 8 flooding events are computed from Fig. 2.5 (and the Appendix A) and are summarized in Table 2.6. From this, we can determine the behavioral acceptance threshold as the average over the results of 8 events, resulting in values of 0.82 for NSE, 4.05% for PE, and 4.35% for VE.

Table 2.6. Values of the acceptance behavioral threshold

Event	NSE [-]	PE [%]	VE [%]
1	0.79	4.87	5.00
2	0.85	4.48	4.81
3	0.80	4.91	5.12
4	0.81	5.41	5.17
5	0.85	4.47	4.83
6	0.83	2.50	3.81
7	0.83	2.85	3.61
8	0.83	2.93	2.48

Mean	0.82	4.05	4.35

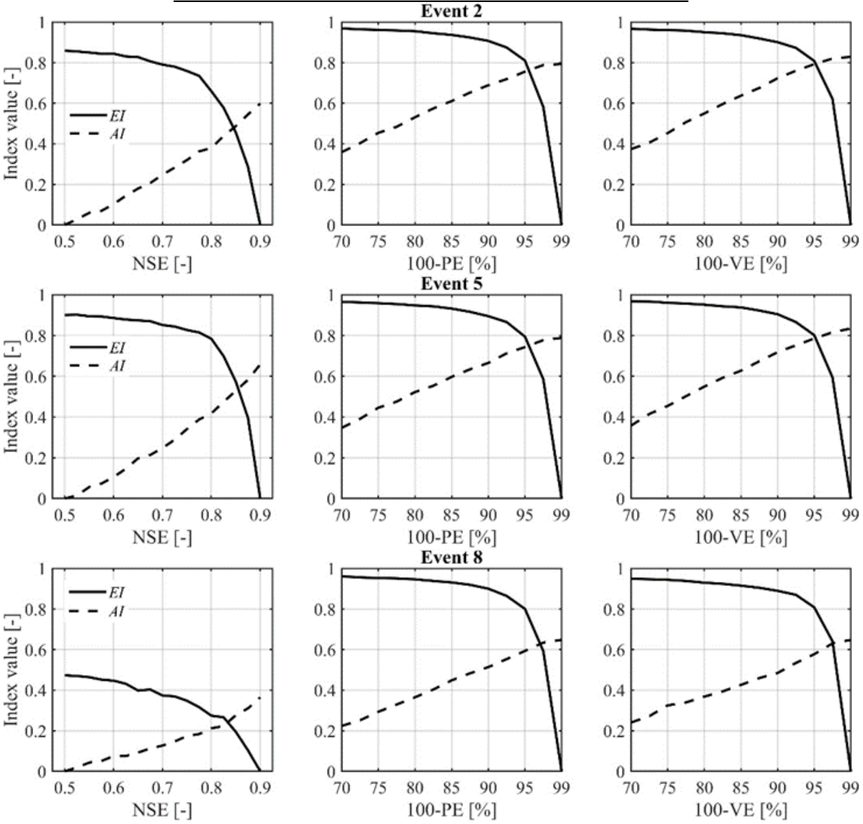


Figure 2.5. The acceptance threshold values of each likelihood function versus the accuracy (*AI*) and efficiency (*EI*) indices of NAM model for 3 flooding events.

The next question is how many behavioral sets are necessary to capture a sufficient uncertainty range. It is apparent that the more behavioral sets one has, the larger the uncertainty range becomes. However, we should determine the optimal number of ensembles for the behavioral sets to avoid excessive runtime. In previous studies, the ensemble size was chosen to be random or large enough to fully identify the confidence interval of the uncertainty [*Cameron et al.*, 2000; *Beven and Freer*, 2001; *Hossain and Anagnostou*, 2005; *Choi and Beven*, 2007; *Blasone et al.*, 2008b; *Jin et al.*, 2010; *Shen et al.*, 2012]. The optimal ensemble size can be determined based on a visual inspection of several key features of the hydrograph (i.e., flood peak and volume), rather than using a random selection. Another feature (i.e., time to flood peak) was not considered in our analysis because the differences between our ensemble results was not significant. Fig. 2.6 shows the evident dependency of the number of ensemble sizes on these key characteristics. As expected, increasing the ensemble size can amplify the chances of various outcomes, resulting in a greater uncertainty range (the latter is similarly computed as the difference between peak or volume values corresponding to 97.5% and 2.5% of ensemble members). Note that the range of uncertainty does not change significantly for an ensemble size of 500 for all the 8 flooding events (see Fig. 2.6 and the Appendix A). Thus, 500 behavioral sets are used as an optimal ensemble size in the study.

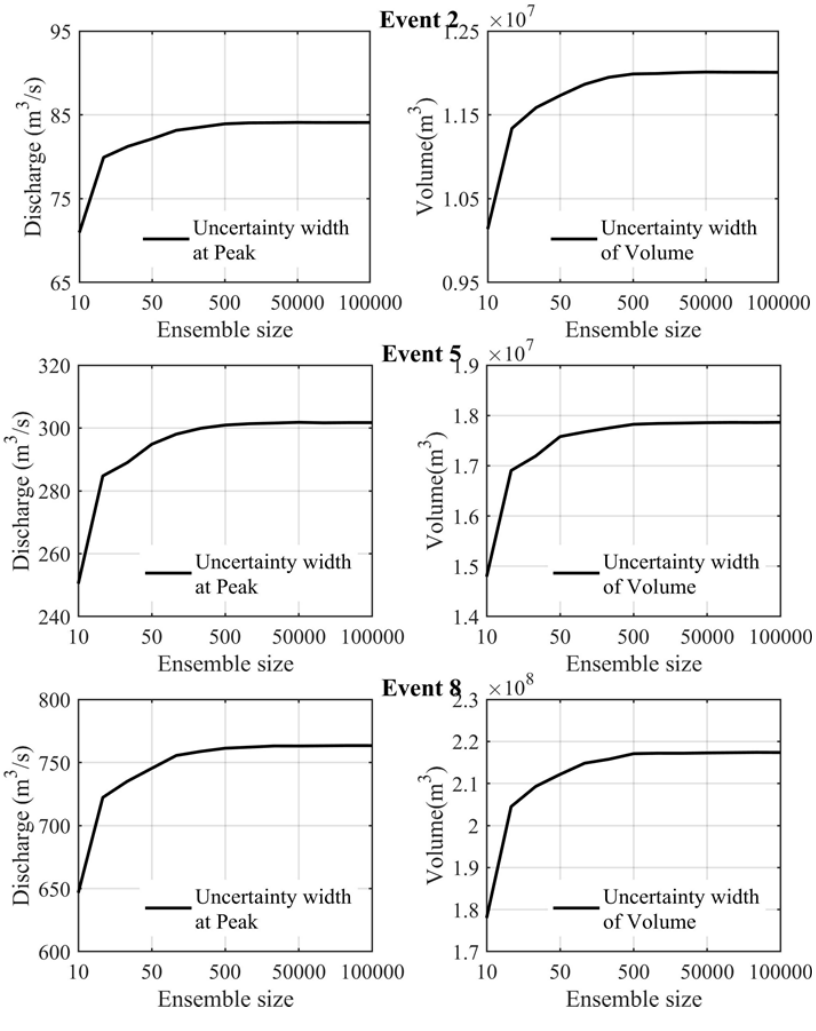


Figure 2.6. The number of ensemble size versus the uncertainty width at flood peak (left) and the uncertainty range of flood volume over entire period (right) for 3 flooding events.

2.2.3.2 Polynomial degree and the number of experiment design on building PCE

We investigated the effects of the experiment design, N and the polynomial degree, p on the PCE model results, thereby providing a guideline for selecting both parameters based on our cases. First, to assess the effect of N, we ran several simulations with N values varying between 10 and 1,500, while the value of p was controlled at 3 for every flood event. The results of 4 metrics of NSE, PE, VE, and R^2, which are compared with the observations show that the values of 4 metrics change significantly (Fig. 2.7) when N equals 50. Using Event 8 as an example, NSE increases from 0.83 to 0.96, PE decreases from 4.5 to 2.6%, and VE also reduces approximately from 17 to 4%when the N value increases from 10 to 50. For N values larger than 50, the model performances are generally comparable. This confirms that a PCE model constructed with an N of 50 has good simulation capacities that are similar to other models with a larger N.

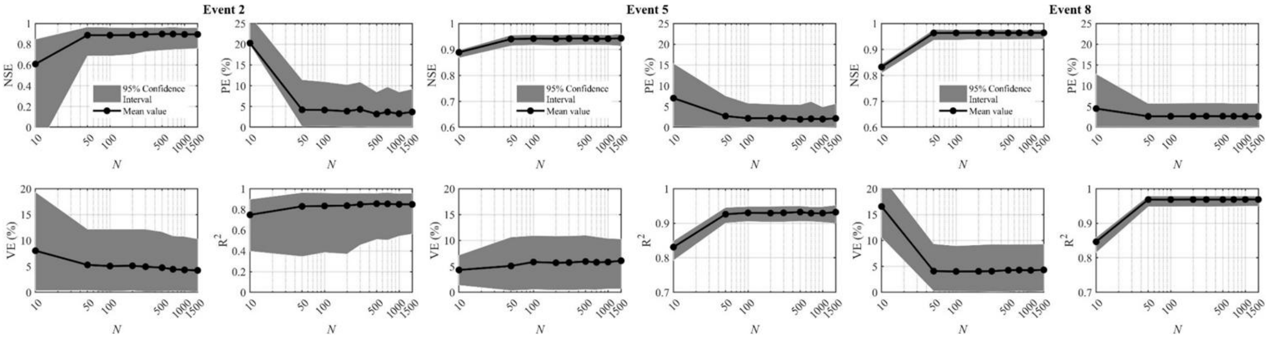


Figure 2.7. The effect of the number of experiment design (N) in the PCE model on four accuracy indices for 3 flooding events. The black line and the shaded region refer to the mean value and 95 % confidence interval of 500 behavioral PCE results, respectively.

Similar to N, we ran a number of simulations with p varying from 1 to 12 and with N controlled at 50. The results of Fig. 2.8 show that two tendencies can be generally observed, although they may vary depending on the events (see the Appendix A): (1) the values of the 4 metrics change considerably when p equals 3 or 4, and remain stable for large values of p especially in Event 2; (2) the values vary negligibly over the entire range of the p, especially in Event 8 and for NSE and R^2. For example, NSE values range from 0.935 to 0.946 and 0.963 to 0.965 for Events 5 and 8, respectively, and R^2 varies slightly as well. Since the p does not significantly affect the performance of the model, a low polynomial degree would be preferred in terms of reducing the computational time in constructing a PCE model. To sum up these two tendencies, a p of 3 or 4 would be a good choice. In the rest of this work, a p of 4 will be consistently employed to build the PCE model.

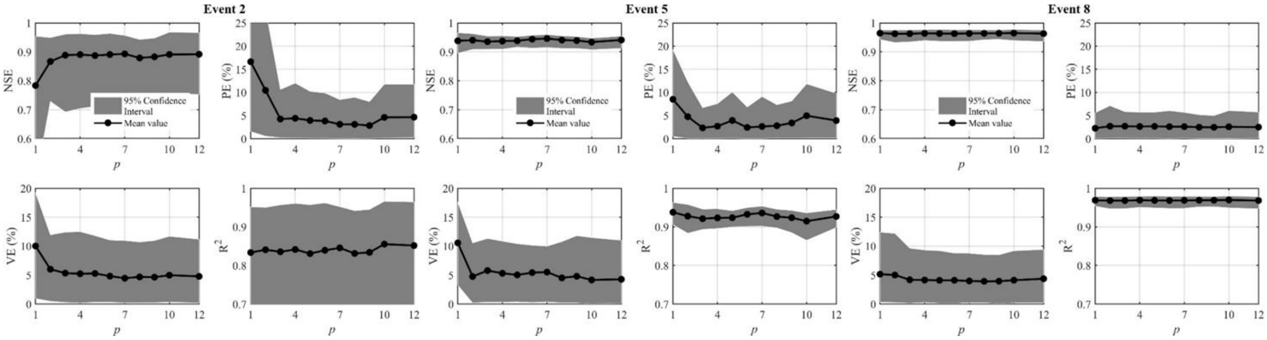


Figure 2.8. The effect of polynomial degree (p) in the PCE model on four accuracy indices for 3 flooding events. The black line and the shaded region refer to the mean value and 95 % confidence interval of 500 behavioral PCE results, respectively.

2.2.3.3 The accuracy and efficiency of PCE model

With the optimal coefficients of $N = 50$ and $p = 4$, a PCE model was built to quantify the uncertain range of flow predictions and to compare the degree of accuracy and efficiency with the results of the deterministic NAM. The uncertainty of both models is illustrated in Fig. 2.9 with the same 500 ensemble sets that were independently generated from the NAM and PCE. To be consistent, the uncertainty bound is also defined as a 95% confidence interval, which corresponds to 2.5% and 97.5% quantiles of the 500 ensemble distribution. Fig. 2.9 shows this uncertainty range of both NAM and PCE model in a hydrograph and their comparisons for a mean of a 500 ensemble results at each computation time. Since we used 500 behavioral sets from GLUE, which satisfies the behavioral acceptance thresholds, the overall comparison with observation is acceptable for both models and for all the flood events (see Table 2.7 and the Appendix A). Specifically, the results of PCE model show that the NSE of all the events is higher than 0.82, and the mean values of PE and VE of all events are approximately 3.9% and 2.6%, respectively; NSEs for Events 1, 2, or 3 (corresponding to a smaller frequency events) were greater than 0.9. Even for a flood hydrograph with a high peak and a complex shape (e.g., Event 8), the flood shape of the PCE model was very close to that of the observation, where the medians of 500 NSE, PE, and VE were 0.957, 2.415%, and 1.953%, respectively. Also, a comparison of the results of PCE models with observation and the NAM results is acceptable, resulting in only insignificant differences. The uncertainty bound of PCE model is slightly broader than that of the NAM for the relatively smaller Event 2. However, note the high R^2 values (greater than 0.99 for all 8 events, see Table 2.7) for a 1:1 comparison of ensemble mean values in time between the NAM and the PCE model results. One can therefore confirm that the PCE model provides a good simulation capability equivalent to the NAM in the Thu bon river watershed for diverse flooding events with different return periods.

Table 2.7. Metrics of NSE, PE, VE, and R^2 for (a) NAM and (b) PCE models compared with observation for all the events chosen in this study.

Event	**R^2**	**(a) NAM model**				**(b) PCE model**			
		NSE[-]	PE[%]	VE[%]	R^2[-]	NSE[-]	PE[%]	VE[%]	R^2[-]
1	0.996	0.966	2.343	2.756	0.965	0.962	3.546	2.718	0.960
2	0.980	0.928	2.166	2.160	0.892	0.930	3.727	3.437	0.908
3	0.998	0.937	1.885	2.261	0.937	0.937	2.511	2.183	0.935
4	0.989	0.822	3.440	1.024	0.860	0.820	5.701	3.574	0.848
5	0.995	0.955	2.498	2.998	0.944	0.944	4.995	1.610	0.926
6	0.990	0.825	3.225	2.115	0.857	0.824	5.447	3.152	0.851
7	0.999	0.852	1.969	2.134	0.871	0.853	2.498	1.994	0.872
8	0.999	0.957	2.031	2.009	0.964	0.957	2.415	1.953	0.964

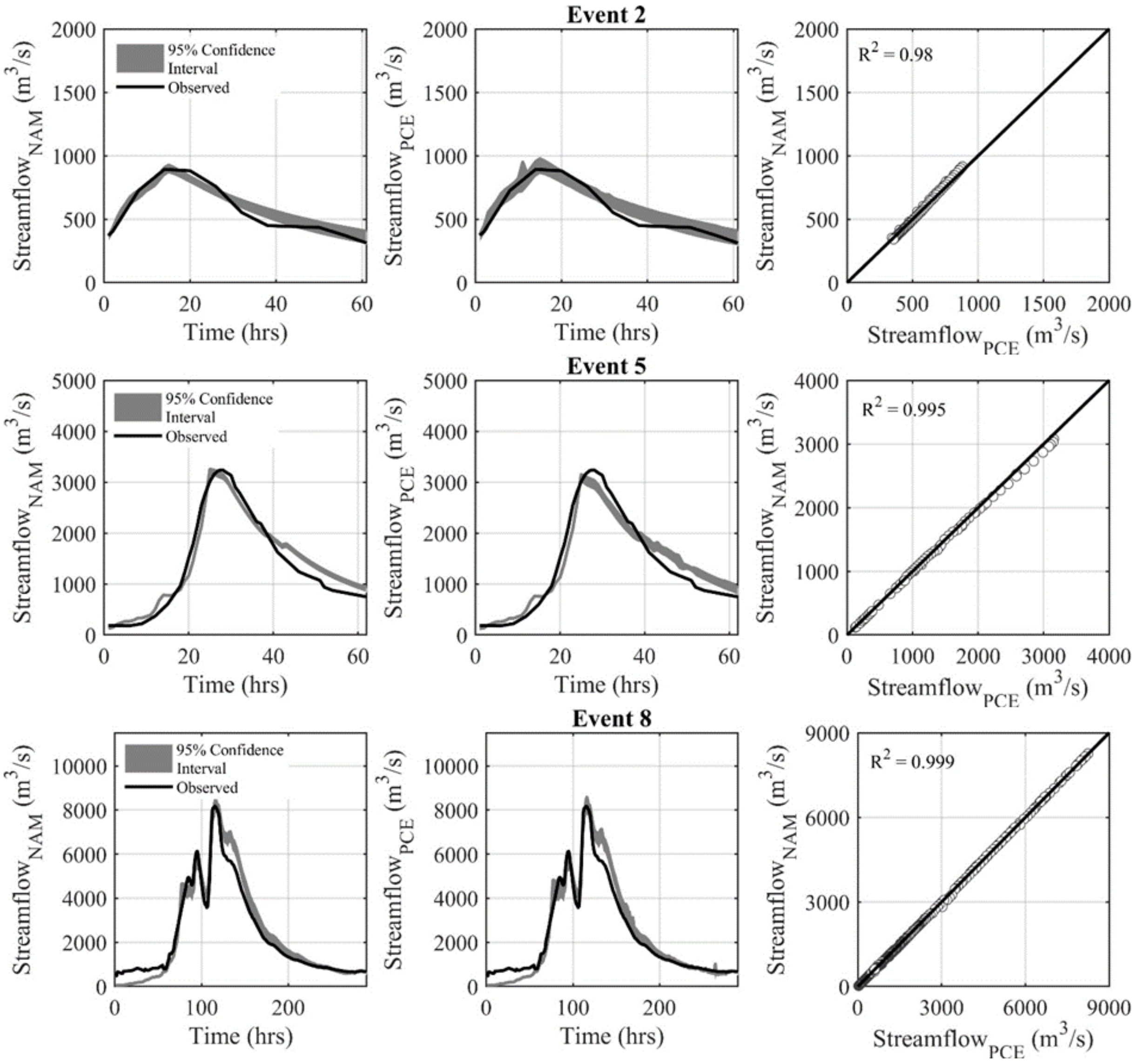


Figure 2.9. Comparison of 95% confidence interval of (left) 500 behavioral NAM runs estimated from GLUE and (middle) 500 PCE runs computed from the PCE model; (right) 1:1 comparison of the temporal values averaged over 500 ensemble for both models, for 3 flooding events.

The times required to obtain the above uncertainty range are then evaluated and compared to determine a relative efficiency. First, the total runtime needed to implement 500 NAM runs are 72, 86.4, and 92.6 secs for Event 2, 5, and 8, respectively, while PCE model are 15.1, 15.5, and 33.4 secs. On the other hand, the total runtime in the PCE model consists of 3 times: (i) the time required for the 500 PCE behavioral runs. Those are 5.9, 4.9, and 11.9 secs for Events 2, 5, and 8, respectively (see Table 2.8). The other 2 times are associated with building the PCE model: (ii) the time to secure the number of experimental designs (i.e., 50 runs of N) and (iii) the time to compute the PCE coefficients. As expected, the efficiency of the PCE model is not superior to the NAM when the ensemble size is small. For example, for a 500 ensemble size, there is only an efficiency improvement of approximately three times in Event 8 and five times in Event 2. However, its efficiency improved significantly for larger ensemble sizes, e.g., approximately eight to seventeen times faster for an ensemble size of 100,000 (Table 2.8).

Table 2.8. Comparison of the times required for constructing the ensemble of NAM and PCE models from 500 to 100,000 for 3 flooding events (in seconds). The times for PCE model consist of (i) the time to generate 500 model runs, (ii) the time to secure the number of the experiment design (i.e., 50 runs of *N*), and (iii) the time to compute the PCE coefficients.

Event	**Model**		**Ensemble size**			
			500	1000	10,000	100,000
2	NAM		72	141.3	1398.6	14292.2
	PCE	(i)	5.9	11.5	111.8	1116.1
		(ii)	7.1			
		(iii)	2.1			
		Total	15.1	20.7	121	1125.3
5	NAM		86.4	176.3	1675.9	18452.2
	PCE	(i)	4.9	9.9	99.8	1066.8
		(ii)	8.8			
		(iii)	1.8			
		Total	15.5	20.5	110.4	1077.4
8	NAM		92.6	184.1	1819	24017.1
	PCE	(i)	11.9	24.1	236.4	3033.1
		(ii)	9.2			
		(iii)	12.3			
		Total	33.4	45.6	257.9	3054.6

2.2.3.4 Toward a more efficient simulation with sensitivity analysis

The sensitivity of the 9 parameters adopted in the NAM and PCE model was investigated by both visual inspection of the posterior distribution and mathematical computations of the Sobol’ and Morris indices. First, one can graphically inspect the posterior distribution of each parameter that was derived from the GLUE based on the behavioral acceptance threshold of likelihood functions. Looking at Fig. 2.10, the shape of the posterior distribution varies depending on the parameter and the event. Note that some distributions have a pointed shape, indicating that those parameters are sensitive and well identifiable, while other flat distributions indicate that their parameters are relatively insensitive and more uncertain. For most events, Fig. 2.10 shows that the parameters of "CQOF" and "CK12" have quite narrow (pointed) distributions, representing a low uncertainty, while the remaining 7 parameters are almost equally distributed over the entire parameter range and have wider distributions, indicating a higher uncertainty. The narrowed range of the sensitive parameters CQOF and CK12 depends on the event. For Event 2, the CQOF and CK12 ranged from 0.25 to 0.44 and from 40 to 66, respectively, while for Event 5, these narrowed from 0.4 to 0.46 and from 40 to 45 (Fig. 2.10).

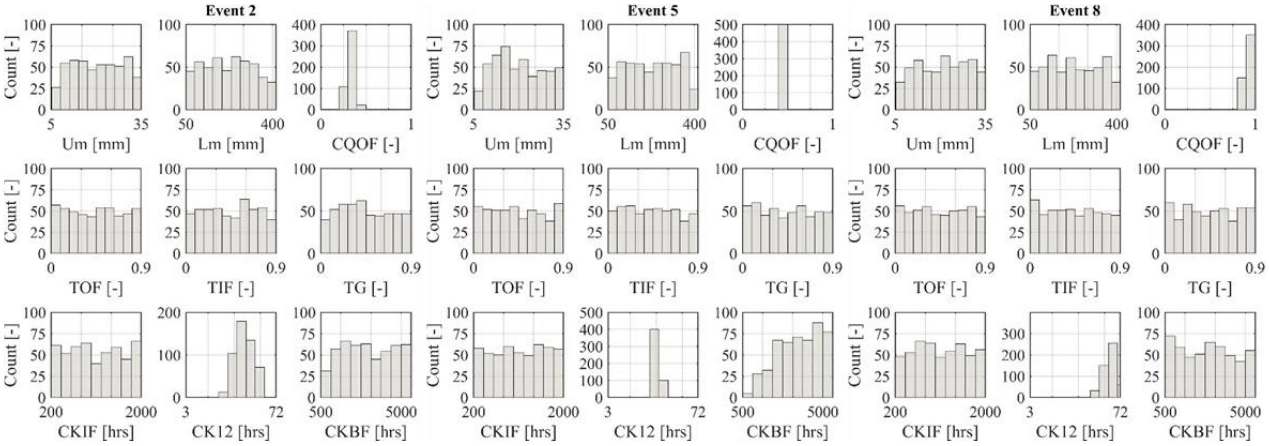


Figure 2.10. The posterior histograms of nine parameters of the NAM model inferred by GLUE for 3 flooding events.

Second, sensitivity analysis using Morris screening and the Sobol' indices showed that both results are comparable and consistent with the result inferred from the previously mentioned posterior distribution. The results for Events 2, 5, and 8 shown in Fig. 2.11 (and the Appendix A for other events) also confirmed that CQOF and CK 12 are the most sensitive parameters to all 3 likelihood functions, while other parameters have very low relative sensitivity. Specifically, the model outputs in terms of the metric of PE are sensitive to both CQOF and CK12, while those for VE are sensitive to COQF. NSE showed more mixed results for various events. In Event 2, the NSE outputs were also sensitive to CQOF and CK12, while in Event 8, those are most sensitive to only CQOF. Summing up all the events, the sensitivity of CQOF is the largest for most events, and CK12 is the next sensitive parameter, especially for events with smaller return periods. Therefore, it can be concluded that CQOF and CK12 are the most important parameters, and these should be preferentially considered for calibration.

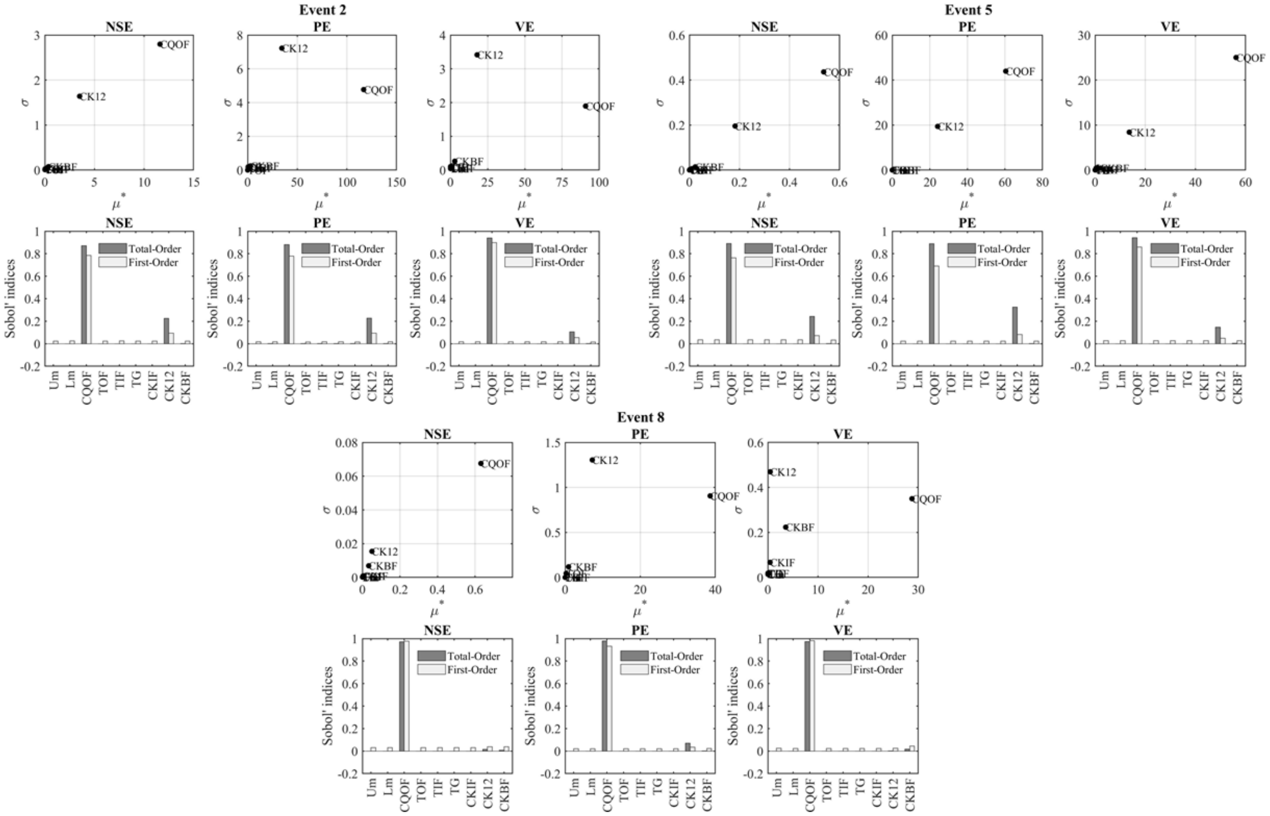


Figure 2.11. Sensitivity results of nine parameters based on different likelihood functions for (left) NSE, (middle) PE, and (right) VE for 3 flooding events. Morris screening index is in the first row of six subplots corresponding to each event, while Sobol' indices is in the second row.

Based on the sensitivity information obtained from the above inspections, one might want to save computational resources and increase efficiency in making any ensemble necessary for capturing uncertainty. We simply counted the number of repeated MC runs required for producing 500 behavioral sets for 3 cases below, which employ (Case 1) the uniform distribution for the reduced range (see Table 2.9), (Case 2) the posterior distribution of GLUE (Fig. 2.10), and (Case 3) a fixed value averaged over the posterior distribution (see Table 2.9). The number of MC runs to get 500 behavioral parameter sets decreased significantly as compared to the case when the original (wider) range of parameters were used (Table 2.10). This becomes more pronounced for Case 3 where the values of the 2 sensitive parameters were fixed as the mean value of the posterior distribution. The assessment results are as good as those given by using the original (wider) parameter range without deteriorating the model accuracy (Table 2.11). Therefore, this is one of the most efficient ways of indirectly reducing runtime by helping to find the behavioral set of parameters even faster by better identifying sensitive parameters from a narrowed parameter space during the calibration process.

Table 2.9. The original (prior) range, reduced (posterior) range from GLUE, and the average of the posterior distribution for CQOF and CK12 for all events

Event	**CQOF [-]**			**CK12 [hrs]**		
	original	reduced	mean	original	reduced	mean
1	0-1	0.35 – 0.52	0.435	3-72	40 – 58	44.873
2		0.25 – 0.44	0.347		40 – 66	49.550
3		0.44 – 0.61	0.539		40 – 70	53.445
4		0.42 – 0.46	0.449		40 – 44	41.080
5		0.40 – 0.46	0.431		40 – 45	41.560
6		0.41 – 0.60	0.519		46 – 70	60.568
7		0.78 – 0.95	0.881		43 – 70	58.804
8		0.85 – 0.98	0.916		52 - 70	63.685

Table 2.10 The number of model runs needed for obtaining 500 behavioral sets by using parameters sampled from the (a) prior and (b) posterior distributions for all events

Event	**(a) from the prior parameters range**	**(b) from the posterior parameter distributions**		
		Case 1	Case 2	Case 3
1	115,736	11,009	3,625	735
2	46,275	7,411	3,822	683
3	20,446	3,642	2,729	618
4	2,310,013	8,997	3,866	2,306
5	564,818	4,408	2,132	821
6	1,266,347	210,208	123,551	8,070
7	17,725	2,428	1,492	500
8	21,157	1,523	931	500

Table 2.11. The accuracy of the model results when using parameters sampled from the posterior distributions for all events, compared with observation

Event	**Case 1**			**Case 2**			**Case 3**		
	NSE[-]	PE[%]	VE[%]	NSE[-]	PE[%]	VE[%]	NSE[-]	PE[%]	VE[%]
1	0.966	2.473	2.585	0.963	2.052	2.452	0.967	2.629	2.209
2	0.931	2.021	2.125	0.936	1.969	1.962	0.937	1.779	1.539
3	0.938	1.933	2.118	0.935	1.839	2.137	0.948	2.017	1.415
4	0.822	3.458	1.051	0.822	3.350	1.019	0.823	3.566	1.108
5	0.955	2.429	2.997	0.957	2.200	2.821	0.955	2.595	2.772
6	0.826	3.312	2.123	0.826	3.276	2.240	0.825	3.150	1.807
7	0.852	1.943	2.196	0.853	1.994	1.894	0.858	0.335	0.472
8	0.956	2.041	1.991	0.956	1.934	1.856	0.959	0.597	1.373

Consequently, the behavioral set search becomes more efficient for a greater number of identifiable parameters. Since these results come from limited cases for flooding samples, investigating the sensitivity for low/medium flow data is meaningful. Thus, we additionally conducted experiments for three lower flow events (their peaks are 20 – 40 times smaller than that of the Event 8), but we could not find any more identifiable parameters except for CQOF and CK12. The results of posterior parameter distributions as well as sensitivity analysis still show that the most sensitive parameters are CQOF and CK12, while the remaining parameters are less sensitive, and their posteriors are almost uniformly distributed over their prior parameter ranges (not shown). This implies that only two parameters within the model structure play a decisive role when simulating events at small temporal scales. But, if one simulates events on a longer time scale where other land surface and subsurface dynamics become significant, other parameters may become important.

2.2.4 Discussions

One might contend that the advantage of PCE does not offset the extra effort required compared to using a simple model. Indeed, such an improvement in computational speed might look insignificant (3 to 17 times; see Table 2.8). However, this tendency will be noticeable if we use more complex models and consider the real-time flood forecasting problem. This is because the additional PCE efforts above can be made in advance before flooding occurs. Then, one just compares the times required to run any deterministic model versus its surrogate model. The time required to run the PCE (consisting of polynomials) is always similar and shorter, while the time required to run the original model increases significantly as the complexity of the latter model increases. Therefore, the effectiveness of the computation speed improves as the deterministic model becomes more complex.

Another concern about whether PCE can capture the dynamics of more complex models for extremes is the universal problems that can occur with any data-driven model (e.g., surrogate modeling, deep learning, genetic algorithms, artificial neural networks, or fuzzy systems). Although studies have not yet been conducted on whether PCE can be applied to more complex flooding phenomena like shock, backwater, and hydraulic jump, many hydrological studies show that its utility as a surrogate model has been successfully proved even for relatively simple models as well as complex distributed models including: lumped, rainfall-runoff models [*Fan et al.*, 2014; *Wang et al.*, 2017], subsurface flow models [*Sochala and Le Maître*, 2013; *Elsheikh et al.*, 2014], a complex numerical model of gas injection into porous media [*Baştuğ et al.*, 2013], and a complex, process-rich ecohydrologic model [*Dwelle et al.*, 2019]. Those studies have demonstrated that PCE both increases computational performance and mimics the deterministic models.

Re-building the PCE model is necessary whenever a new deterministic model is developed. However, as mentioned above, reconstructing the PCE can be pre-performed during off-peak periods irrespective of the flood, and can save much time in making flow ensembles during on-peak periods. Note the execution time of ~18 minutes versus ~5 hours in Event 5. This time reduction highlights that the efficiency of the PCE model is worthy of attention from a practical point of view because flood forecasting results should be provided early for flood warning or risk management.

We also pondered how these GLUE results would change if one had less restrictive threshold criteria (or if we used fewer likelihood functions). (i) The uncertainty band would be larger, the predictive power would be poorer, and computational power would improve. (ii) The GLUE's posterior parameter distribution would follow a more uniform distribution rather than a "real" posterior distribution. (iii) The curve in Fig. 2.6 would be shifted upward, and the inflection point of this curve (i.e., optimal ensemble number) would be larger. In contrast, if one uses a "much" more restrictive set of criteria, the prediction would be close to the observation itself, so that the curve in Fig. 2.6 would be shifted downward and the inflection point would be diminished (i.e., located near zero). (iv) The posterior distribution of all parameters would converge equally over the 500 optimal ensemble size. This means that an ensemble size of 500 can fully demonstrate the "real" posterior distributions as well as the uncertainty range of model results. But, the posterior distributions differ for an ensemble size smaller than 500 depending on the parameters, and those distributions have not converged except for the two sensitive parameters (Fig. 2.12).

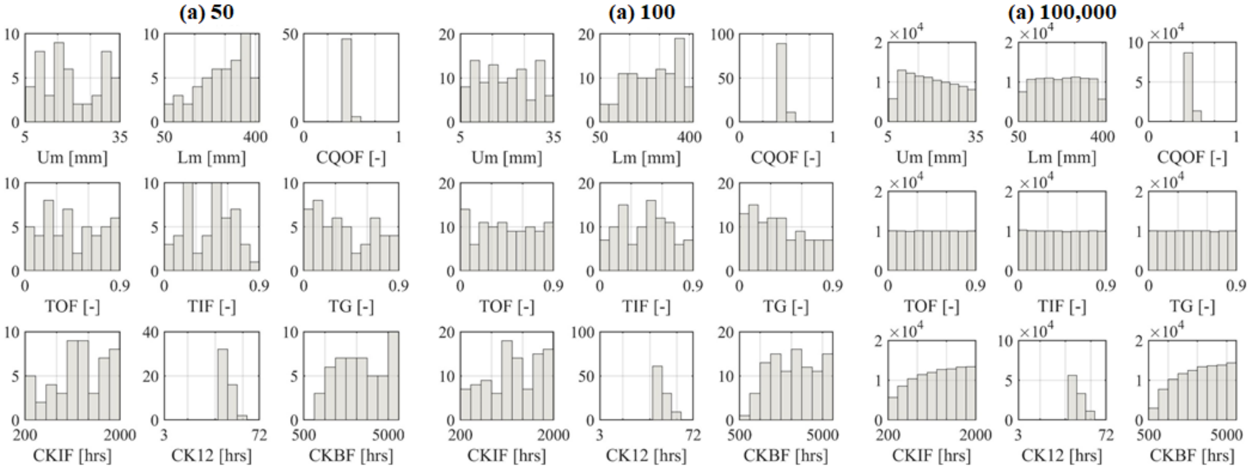


Figure 2.12. The posterior histograms of nine parameters of the NAM model inferred by GLUE for Event 1 with the ensemble size of (a) 50, (b) 100, and (c) 100,000.

2.3 Towards efficient surrogate model using sparse polynomial chaos expansion

2.3.1 Challenges in the use of the surrogate model

As originally proposed by *Wiener* [1938], PCE was based on normally-distributed random variables and a Hermite polynomial, and was later extended to be applied to any statistical distribution by *Xiu and Karniadakis* [2002]. The key to PCE effectiveness is how to estimate PCE coefficients from the response of an original model at design points in the input space [*Sudret*, 2008]. Two widely used methods for optimizing the PCE coefficients are the ‘projection’ method, which can be cast as a numerical integration problem using quadrature or sparse-grid methods, and the ‘regression’ method, which uses least square regression to minimize the mean square error between the surrogate model outputs and original model outputs [*Sudret*, 2008]. However, both methods is incompetent in optimizing a great number of PCE coefficients because of the large number of model evaluations entailed [*Blatman and Sudret*, 2010; *Razavi et al.*, 2012b; *Liu et al.*, 2013]. The number of PCE coefficients increases dramatically with the number of uncertain inputs and the polynomial order. This ‘full’ PCE requires an incredibly large number of model evaluations that severely restrict the engineering applications [*Sargsyan et al.*, 2014].

To circumvent this problem, methods of downsizing the PCE coefficients have been proposed such as sparse collocation [*Shi et al.*, 2009], Bayesian compressive sensing [*Sargsyan et al.*, 2014], and least angle regression (LAR) [*Blatman and Sudret*, 2008]. Among them, LAR has received attention recently because it has been proven to provide significant computational gains over original PCE. The purpose of LAR is generally to estimate only the coefficients for the important PCE basis terms and assign zero to the coefficients for the non-essential terms. LAR enables high orders of polynomials to be fit to nonlinear complex models without substantially increasing the computational cost during the construction of a surrogate model [*Zhang et al.*, 2020]. Although the effectiveness of LAR has been demonstrated, few studies have coupled LAR with PCE in order to quantify the uncertainty of a hydrologic model.

The aims of this work are to examine whether sparse PCE (SPCE) captures the behavior of a hydrological model well, quantifies the uncertainty of parameters of the hydrological model, and analyzes the sensitivity of parameters to hydrologic predictions. To highlight the effectiveness and robustness of LAR, a well-known method, the ordinary least square regression (OLS), is used and compared.

2.3.2 Methods

2.3.2.1 Sparse polynomial chaos expansion using least angle regression

The least angle regression (LAR) method is an advanced regression method in solving Eq. (2.5) where a modification for the penalty term $\lambda\|\varepsilon\|_1$ is added:

$$\varepsilon \ = \text{argmin}_{\varepsilon\in\mathbb{R}^{N_\Psi}}\mathbb{E}\left[\left(y - \sum_{\alpha=0}^{N_\Psi-1}\varepsilon_\alpha\Psi_\alpha(\boldsymbol{\theta})\right)^2\right] + \lambda\|\varepsilon\|_1 \tag{2.19}$$

where λ is a non-negative constant; $\|\varepsilon\|_1$ is a regularization term that forces a minimization to favor the sparse solution, computed as $\|\varepsilon\|_1 = \sum_{\alpha\in N_\Psi}|\varepsilon_\alpha|$.

The main difference between LAR and OLS lies in the number of PCE coefficients, which is smaller in LAR than in OLS. Specifically, in OLS, the number of PCE coefficients that need to be estimated is N_Ψ, which can be computed from Eq. (2.3). The surrogate model constructed by OLS is hereafter called full PCE (FPCE). On the other hand, LAR determines only the multivariate polynomials $\Psi_\alpha(\boldsymbol{\theta})$ that have the most impact on the model response, while discarding polynomial terms that do not. The chosen weighty PCE coefficients are estimated, while other insignificant coefficients are set to be zero. A surrogate model is then achieved based on the sparse set of PCE terms and can be delineated as Eq. (2.20). This surrogate model is hereafter called sparse PCE (SPCE). For a detailed description of SPCE, readers can refer to *Blatman and Sudret* [2011]. To verify the accuracy of constructed surrogate models, the leave-one-out cross-validation error (LOO) is commonly used.

$$y = \boldsymbol{M}(\boldsymbol{\theta}) \approx \boldsymbol{M}^{PCE}(\boldsymbol{\theta}) = \sum_{\alpha=0}^{S_\Psi-1}\varepsilon_\alpha^s\Psi_\alpha^s(\boldsymbol{\theta}) \tag{2.20}$$

$$LOO = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{\boldsymbol{\mathcal{M}}(\mathcal{X}^{(k)}) - \boldsymbol{\mathcal{M}}^{PCE}(\mathcal{X}^{(k)})}{1-\hbar_k}\right)^2 \tag{2.21}$$

where $\Psi_{\alpha}^{s}(\boldsymbol{\theta}) = \left\{\Psi_0^s(\boldsymbol{\theta}), \dots, \Psi_{S_\Psi - 1}^s(\boldsymbol{\theta})\right\}$ are the set of significant polynomials; $\varepsilon_{\alpha}^{s} = \left\{\varepsilon_0^s, \dots, \varepsilon_{S_\Psi - 1}^s\right\}$ are the corresponding coefficients; S_Ψ is the number of PCE terms that are retained; $\hbar_k$ is the k-th diagonal term of the matrix $\boldsymbol{F}(\boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T$ and the information matrix $\boldsymbol{F}$ is defined in Eq. 2.8.

2.3.2.2 Hydrological model: Storage function model

A conceptual, lumped, storage function-based hydrological model is employed, which has been adopted for flood prediction practice at the Han River Flood Control Office under the Ministry of Environment of Korea [*Bae and Lee*, 2011; *Office*, 2012; *Park et al.*, 2014; *Kim et al.*, 2019a]. The storage function model (SFM) [*Kimura*, 1961] is an event-based, lumped model that characterizes the relations of rainfall, runoff, and storage in watersheds and channels by solving the flow continuity equation. Rather than solving the full dynamic momentum equations, the SFM employs a nonlinear relation between storage and discharge for a given watershed and channel as:

$$S_{bas}(t) = K_{bas} \times Q_{bas}^{P_{bas}}(t) \tag{2.22}$$

$$S_{chn}(t) = K_{chn} \times Q_{chn}^{P_{chn}}(t) \tag{2.23}$$

where $S_{bas}(t)$ and $S_{chn}(t)$ are the storage amounts of the basin and channel at time t, respectively; $Q_{bas}(t)$ and $Q_{chn}(t)$ are the direct runoffs (flow rates) of the basin and channel at time t, respectively; K_{bas} and P_{bas} are the storage coefficient and exponent of the basin, while K_{chn} and P_{chn} are the storage coefficient and exponent of the channel.

The spatially-lumped continuity equation for a given basin and channel is expressed as:

$$\frac{dS_{bas}(t)}{dt} = R_e(t - Tl_{bas}) - Q_{bas}(t) \tag{2.24}$$

$$\frac{dS_{chn}(t)}{dt} = R_e(t - Tl_{chn}) - Q_{chn}(t) \tag{2.25}$$

where R_e is the effective rainfall, and Tl_{bas} and Tl_{chn} are time delays between the effective rainfall and the outflow of the basin and channel, respectively.

In SFM, $R_e(t)$ is estimated based on the saturated rainfall approach of *Sukegawa and Kitagawa* [1992]. Specifically, before the accumulated rainfall depth $\sum R(t)$ reaches the saturated rainfall R_{sa}, $R_e(t)$ is computed based on the primary runoff ratio (f_1); after $\sum R(t)$ exceeds R_{sa}, $R_e(t)$ is a function of the saturated runoff ratio (f_{sa}):

$$R_e(t) = \begin{cases} f_1 \times R(t) & \sum R(t) < R_{sa} \\ f_{sa} \times R(t) & \sum R(t) \geq R_{sa} \end{cases} \tag{2.26}$$

The lumped rainfall depth of the basin and channel ($R(t)$) is corrected based on observed rainfall depth ($R_{obs}(t)$) and rainfall multiplication factor (α): $R(t) = \alpha \times R_{obs}(t)$. From the brief description above, one can see that a total of 10 parameters are required to control the outflow of the watershed and implement the SFM (Table 2.12). For more detail, readers can refer to *Park et al.* [2014].

Table 2.12. Description of the SFM parameters

Parameter	Unit	Description	Lower bound	Upper bound
α	[-]	Rainfall magnification coefficient	0	1.3
f_1	[-]	Primary runoff ratio	0	1
f_{sa}	[-]	Saturated runoff ratio	0	1
R_{sa}	mm	Saturated rainfall	0	300
K_{bas}	[-]	Basin storage-discharge coefficient	1	100
P_{bas}	[-]	Basin storage-discharge exponent	0	1
Tl_{bas}	[hrs]	Time delay in watershed	0	1
K_{chn}	[-]	Channel storage-discharge coefficient	1	100
P_{chn}	[-]	Channel storage-discharge exponent	0	1
Tl_{chn}	[hrs]	Time delay in channel	0	1

2.3.3 Case study

The ‘Hongcheon’ watershed, which belongs to the Han river basin located in the central part of the Korean Peninsula, is chosen for this research (Fig. 2.13). The area of the basin is 883 km^2, its mainstream length is about 60 km, and its altitude ranges from 75 to 1180 m. This study collects data for the rainy season (June to September), focusing on the uncertainty of flood predictions. Rainfall data are observed at 15 weather stations near the study area, and streamflows are observed at the outlet of the watershed, ‘Hongcheon’ gauge station (Korea station ID = 2014650). Hourly observations of rainfall and streamflow data were downloaded from the Han River Flood Control Office (*http://www.hrfco.go.kr/main.do*). After inspecting the data quality and availability, nine streamflow events (Table 2.13) were chosen, corresponding to various (low, middle, and high) return periods based on frequency analysis (Fig. 2.14).

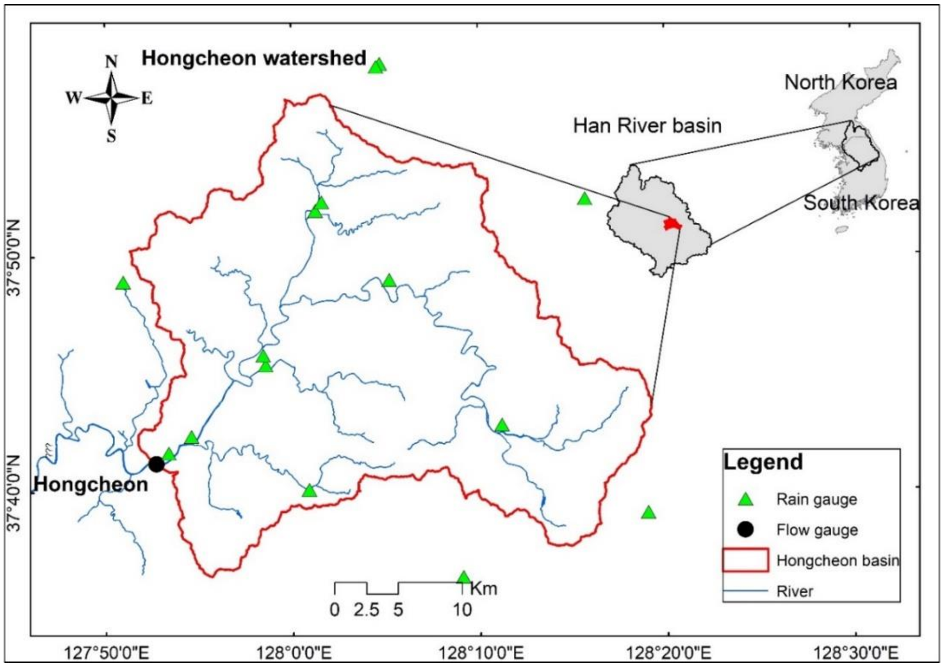


Figure 2.13. The ‘Hongcheon’ watershed belonging to Han river basin, and the locations of observed rainfall and flow gauges.

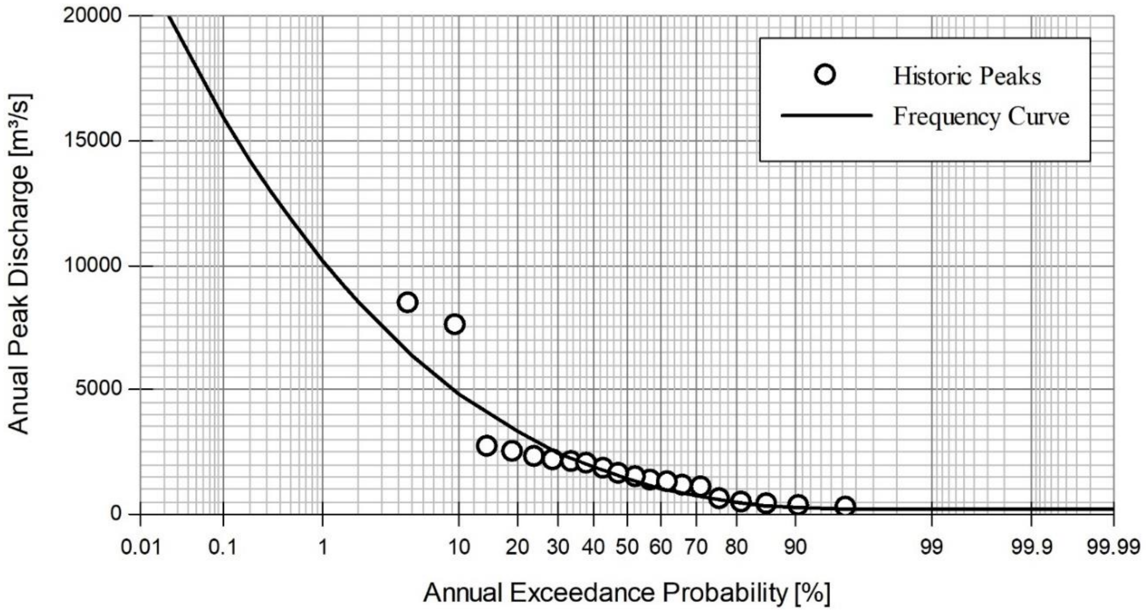


Figure 2.14. Flow frequency curve for the 'Hongcheon' station; historic peaks refer to annual maximum peak flows from 2000 to 2019; the flood frequency curve is fitted using the Gamma distribution.

Table 2.13. Characteristics of selected streamflow events in Hongcheon watershed

Event	Time (MM/DD/YYYY)	Flood peak (m^3/s)	Flood frequency (%)	Duration (hrs)
1	7/7/2009-7/17/2009	2485.33	19	241
2	7/10/2012-7/20/2012	416.61	86	241
3	7/10/2013-7/17/2013	2264.07	28	169
4	7/21/2013-7/27/2013	477.59	81	145
5	7/23/2015-7/27/2015	477.60	81	97
6	6/29/2016-7/9/2016	1460.90	52	241
7	6/30/2017-7/5/2017	1616.14	47	121
8	7/9/2017-7/13/2017	1337.97	57	97
9	8/27/2018-8/31/2018	689.41	76	97

2.3.4 Experimental configurations

The SPCE and FPCE models are compared by investigating the ability to construct a satisfactory surrogate model with a limited training dataset, the degree of accuracy reflecting uncertainty in streamflow prediction, and the degree of improvement in the efficiency of two surrogate models compared to the original model. The following three experiments were conducted.

The first experiment is designed to demonstrate the effectiveness of SPCE in a smaller experimental design. In the literature, the size ranges from 50 to $\mathcal{O}(10^4)$, based on the complexity of the original model [*Hampton and Doostan*, 2015; *Dwelle et al.*, 2019; *Torre et al.*, 2019; *Tran and Kim*, 2019]. In this experiment, a total of ten different sizes, N, from 10 to 5000 are used to build the surrogate model. A polynomial degree of 3 is used, as in previous studies [*Fan et al.*, 2016; *Wang et al.*, 2017; *Hu et al.*, 2019a; *Tran and Kim*, 2019; *Tran et al.*, 2020].

Given surrogate models constructed for the optimum value of N determined in the first experiment, the second experiment is conducted to quantify the uncertainty of streamflow prediction for nine rainfall events using GLUE. Prior distributions for the uncertain parameters are assumed to follow the uniform distribution over a given (prior) range [*Beven*, 2006; *Vrugt et al.*, 2008c; *Tran and Kim*, 2019]. Latin hypercube sampling (LHS) is used due to its efficiency [*Hu et al.*, 2019a]. Regarding the cutoff threshold, we employ the ratio of the total number of simulations based on the likelihood function value to differentiate between the behavior and non-behavior runs. Specifically, the cutoff threshold is designated as the highest 1% of Nash-Sutcliffe efficiency coefficient (NSE) values computed using 100,000 random parameters sampled from the prior distributions [*Beven*, 2012; *Tran and Kim*, 2019]. The uncertainty of streamflow is then represented by calculating the ensemble interval for the NSE and Peak Error (PE) metrics, which can indicate important features of a streamflow event.

Sensitivity analysis (SA) is implemented as the third experiment to recognize the critical parameters governing model behavior and to evaluate the influence of model parameters on model outputs [*Saltelli*, 2002a; *Tran and Kim*, 2019]. These key parameters can be identified qualitatively based on the shape of the posterior distributions obtained from GLUE, or quantitatively based on the global sensitivity analysis. The latter produces the sensitivity indices for both parameters and their interactions. The total-order Sobol' indices [*Sobol'*, 2001] are employed in this experiment.

2.3.5 Results and discussions

2.3.5.1 The construction of surrogate models

We investigate the effects of the size of the experimental design on the accuracy of surrogate models constructed by FPCE and SPCE, thereby (i) providing a guideline for choosing the appropriate size of experimental design and (ii) demonstrating the superiority of SPCE to FPCE. As described in Section 2.3.4, we built several surrogate models with N varying between 10 and 5,000 for both FPCE and SPCE. Looking at Fig. 2.15, one can see that the LOO values for all nine events decrease as the value of N increases and are almost indistinguishable when N reaches a certain value (about 2,000 and 500 for FPCE and SPCE, respectively). In other words, if one uses a larger experimental design (greater number of samples) for training, the overall accuracy increases, but at some point the accuracy stabilizes. Visual inspection from Fig. 2.15 confirmed that FPCE and SPCE developed with N of 2,000 and 500, respectively, are suitable to represent SFM. Fig. 2.15 also reveals that SPCE outperforms FPCE in providing lesser LOO values for all events. Specifically, if N is less than 200, the LOO values obtained using SPCE are smaller than 1, while these values for FPCE are larger, ranging from about 5 to 100. If N is greater than 200, the difference of LOO between two surrogates decreases by about 10%. For all events, the LOO values of SPCE constructed with N of 500 are equal to or even smaller than those of FPCE with N of 2,000. SPCE can build an efficient surrogate model with an accurate degree even if it utilizes an experimental design size that is four times smaller than that of FPCE.

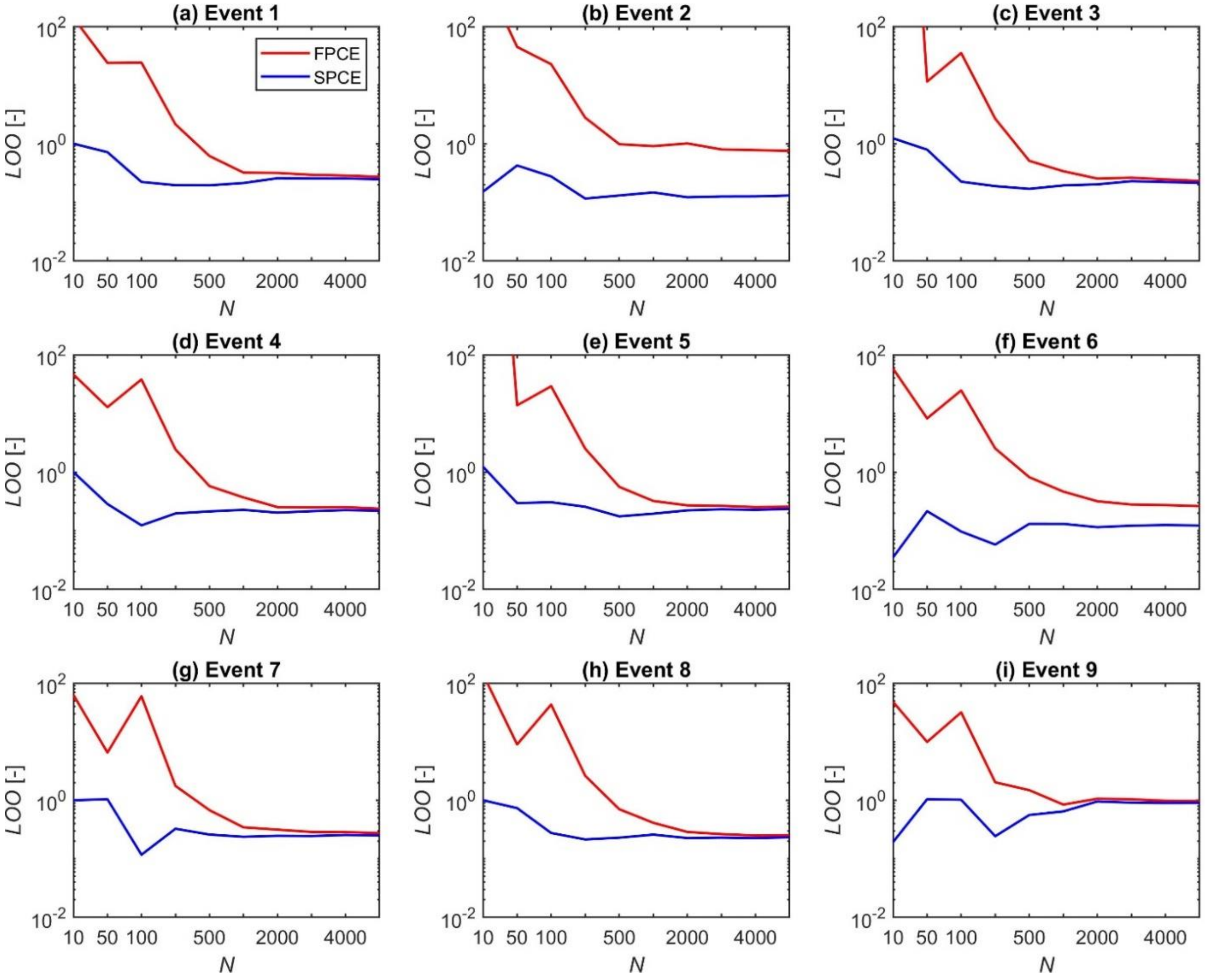


Figure 2.15. The effects of the size of experimental design (N) on the leave-one-out cross-validation error (LOO) in constructing surrogate models using FPCE and SPCE for 9 streamflow events.

As a follow-up discussion based on the benefits of SPCE above, it can be expected that the use of this sparse approach would be more effective, especially for high-dimensional models where heavy computation is required. Since these high-dimensional models contain a large number of uncertain parameters (often greater than $\mathcal{O}(10^2)$), the number of PCE coefficients (N_Ψ) that need to be estimated from Eq. (2.3) are also quite large. This requires a substantial number of model evaluations, up to $N = (p+1)^{N_P}$ [*Sudret*, 2008]. This computational burden emphasizes the need for a more efficient surrogate such as SPCE to reduce the number of PCE coefficients and save computational resources. For example, for FPCE in this work, a total of 286 PCE coefficients are required for all events (computed via Eq. (2.3) for 10 uncertain parameters and the polynomial degree of 3). For SPCE, the number of PCE coefficients (S_{Ψ}) used varies depending on events from 25 (Event 7) to 34 (Event 2) given N of 500 (Fig. 2.16a) and depending on N (Fig. 2.16b). S_{Ψ} increases with N until about 200, while it does not change much for N greater than 200. S_{Ψ} is always less than 50 for all events. With an appropriate value of N (e.g., 500), the significant multivariate polynomials $\Psi_{\boldsymbol{\alpha}}(\boldsymbol{\theta})$ can be fully detected and it is not necessary to use a larger N. Therefore, the number of PCE coefficients for SPCE is about 8 to 11 times smaller than that of FPCE for nine events.

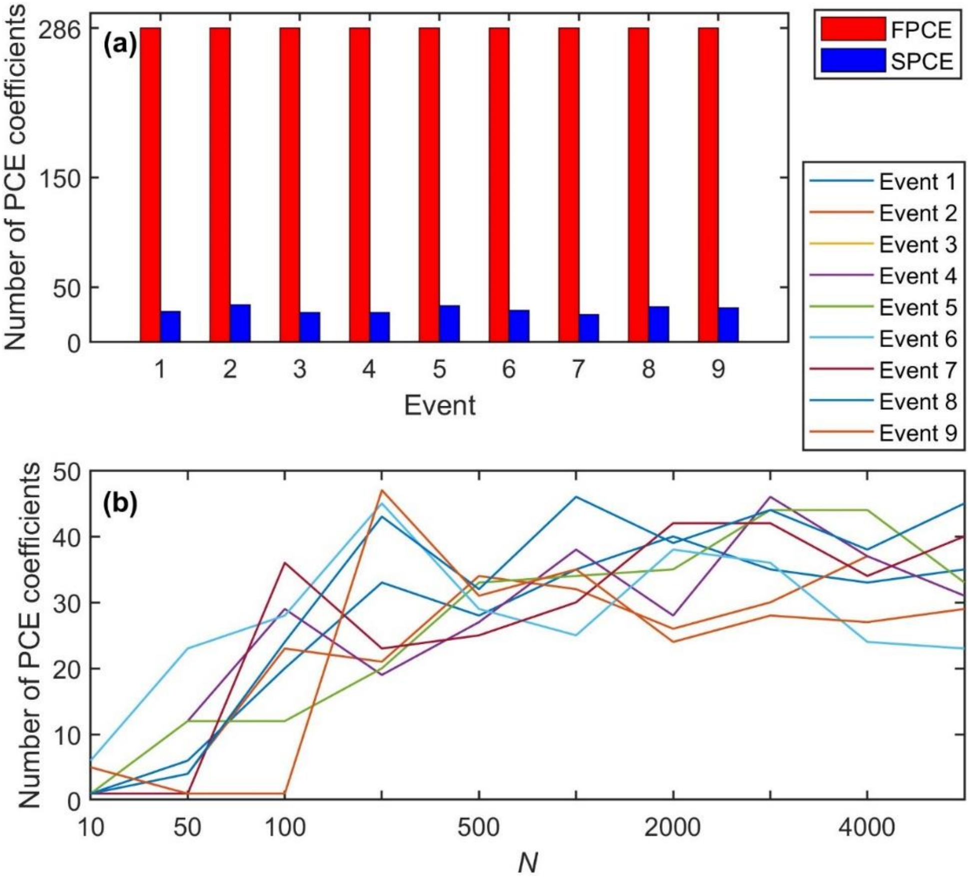


Figure 2.16. (a) The number of non-zero PCE coefficients in constructing FPCE (with N of 2000) and SPCE (with N of 500) for 9 streamflow events. (b) The effects of N on the number of PCE coefficients in SPCE for 9 events.

2.3.5.2 The accuracy of surrogate models

Based on the results from Section 2.3.5.2, optimum sizes of 2,000 and 500 are selected for N when building surrogate models for FPCE and SPCE, respectively. These surrogate models are then employed to quantify the uncertainty of hydrologic predictions through GLUE. The hydrographs of SFM, FPCE, and SPCE are presented with a 90% confidence range of 1,000 behavioral (posterior) hydrographs in Fig. 2.17. The posterior results of all three models are highly satisfactory for all nine events – their uncertainty ranges are very narrow and cover observations. The R^2 values for 1:1 comparisons between the ensemble mean results and observations are mostly higher than 0.8, and the R^2 values of two surrogate models and SFM are similar. The accuracy indices NSE and PE also confirm that both FPCE and SPCE provide a good simulation capability equivalent to SFM for diverse streamflow events with different return periods (Fig. 2.18 and Table 2.14). Additional comparisons between the surrogate models show that SPCE outperforms FPCE. Ensemble mean values for NSE and PE are as high as about 38% and 34% at the maximum, respectively (see Table 2.14 for Event 4). Additionally, the uncertainty ranges of NSE and PE for both surrogate models have smaller standard deviations (Std) than those for SFM. For example, in Event 1, the Std values of ensemble NSE for FPCE, SPCE, and SFM are 0.03, 0.03, and 0.06, respectively, while those of PE are 10.92, 10.46, and 14.05%, respectively (Table 2.14).

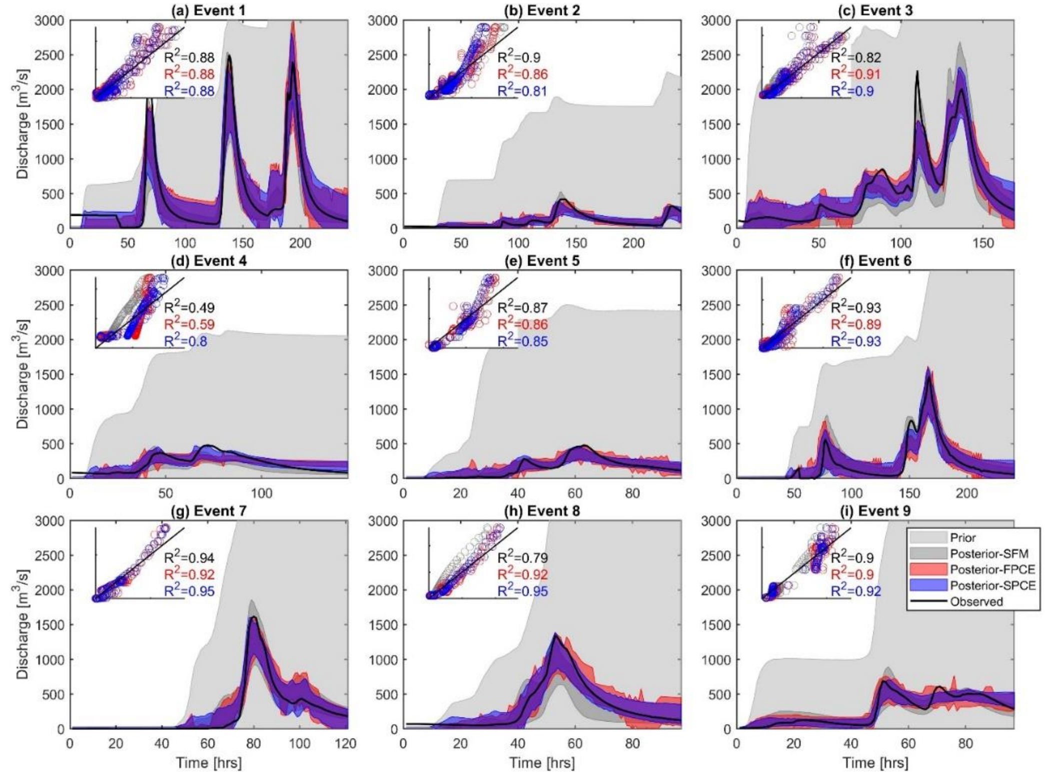


Figure 2.17. Streamflow predicted by SFM, FPCE, and SPCE for 9 streamflow events. The 90% confidence bands are drawn using 1,000 ensemble posterior members identified through GLUE. The scatter plots (and R^2 values) represent 1:1 comparisons between the ensemble mean predictions (y-axis) and the observations (x-axis).

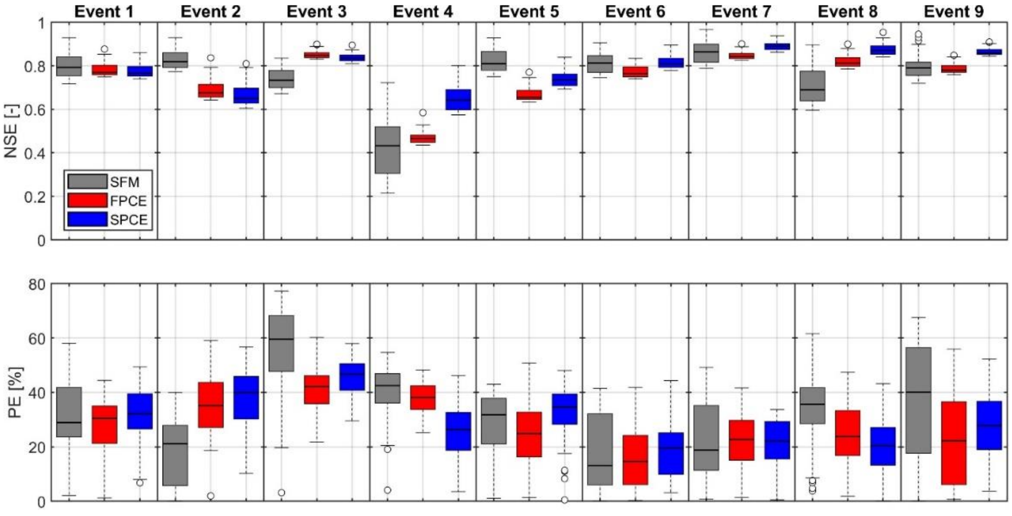


Figure 2.18. Comparisons of accuracy metrics, NSE and PE for 3 models (SFM, FPCE, and SPCE) for 9 streamflow events. The boxplots demonstrate the median (central mark), the 25th and 75th percentiles (the edges of the box), and the maximum and minimum (the upper and lower whiskers) except for outliers (circle symbols).

Table 2.14. Mean and standard deviation (Std) for 1,000 values of NSE and PE for SFM, FPCE, and SPCE for 9 streamflow events.

	NSE [-]						PE [%]					
Event	Mean			Std			Mean			Std		
	SFM	FPCE	SPCE	SFM	FPCE	SPCE	SFM	FPCE	SPCE	SFM	FPCE	SPCE
1	0.80	0.78	0.78	0.06	0.03	0.03	30.09	27.70	31.51	14.05	10.92	10.46
2	0.83	0.69	0.67	0.04	0.04	0.05	18.90	36.19	37.61	11.84	11.66	11.65
3	0.74	0.85	0.84	0.05	0.02	0.02	56.69	42.00	45.55	15.50	7.76	6.62
4	0.43	0.47	0.65	0.13	0.03	0.06	39.52	38.17	25.00	10.07	5.75	10.59
5	0.82	0.67	0.74	0.05	0.03	0.04	28.11	25.29	32.27	11.73	12.42	10.23
6	0.81	0.77	0.81	0.05	0.03	0.03	17.26	16.10	19.29	13.81	11.82	10.41
7	0.86	0.85	0.89	0.05	0.02	0.02	21.86	22.23	21.19	13.80	10.62	9.25
8	0.72	0.82	0.87	0.09	0.03	0.02	34.54	23.79	20.39	13.91	11.13	10.88
9	0.79	0.79	0.86	0.05	0.02	0.02	37.09	22.94	27.75	20.88	16.84	12.07

These ensemble results imply that the likelihood function and cutoff threshold must be carefully selected, which directly affect the prediction accuracy [*Beven*, 2006; *Tran and Kim*, 2019]. For example, since we chose NSE in this study to represent the goodness-of-fit between simulation and observation, the ensemble of NSE has a satisfactory value higher than 0.7 for most events (Figure 2.18). However, the peak error (PE) is relatively large, ranging from 40% (Event 6) to 78% (Event 3). That is, depending on the likelihood function preferred, one can control an outcome in flood prediction. If using a likelihood function that can represent the accuracy of the overall shape, peak size, time of arrival, and total flood volume of a streamflow event, it will make more informed decisions that better reflect each flood characteristic.

Second, to obtain more accurate ensemble results, more likelihood functions with tighter cutoff thresholds can be used. However, instead of attaining higher-accuracy ensemble results, there is a sacrifice of significantly increasing the number of random runs. For example, instead of using the 1% cutoff threshold used in this work, if we apply a cutoff threshold of 0.8 for the NSE likelihood function (this value is often considered as satisfactory [*Moriasi et al.*, 2007b]), the number of ensemble behavior sets decreases sharply (see the number of behavior runs for 100,000 and 10,000,000 prior runs in Table 2.15). The finding that there are only a very small number of ensembles signifies that random searches must be enhanced to obtain results that meet this level of accuracy. This is particularly noticeable for Event 4. With 10,000,000 random runs, SPCE could get only 88 behavior runs while FPCE could not attain even one behavior run. The fact that a large number of iterations are required to achieve the desired accuracy justifies the use of the surrogate model. Even in a simple model like SFM, the CPU runtime required to perform 100,000 random runs was about a month, so applying the model to practical problems is unreasonable. However, for SPCE, even 10,000,000 random simulations take only a few hours to run. The surrogate model consisting of the summations of polynomials has a great advantage for Monte-Carlo type repeated simulations. We will cover the computation time of each model in more detail in Section 2.3.5.4

Table 2.15. The number of behavior runs obtained through GLUE for 3 model (SFM, FPCE, and SPCE), based on the likelihood function of NSE with its acceptance threshold of 0.8. Column (a) and column (b) present results obtained from 100,000 and 10,000,000 random runs, respectively.

Event	**(a)**			**(b)**	
	SFM	**FPCE**	**SPCE**	**FPCE**	**SPCE**
1	104	14	11	1196	1456
2	142	1	1	45	103
3	20	104	62	9520	7503
4	0	0	1	0	88
5	104	0	5	15	521
6	111	8	32	679	3306
7	181	115	219	11147	21771
8	36	35	84	2783	8251
9	91	13	217	1710	21708

2.3.5.3 The sensitivity of uncertain parameters

The sensitivity of each of the 10 parameters of SFM, FPCE, and SPCE is analyzed from the posterior (behavior) parameter distributions obtained by GLUE, as depicted in Fig. 2.19. In general, parameters that have pointed distributions are relatively sensitive and identifiable, while parameters with flat-shaped distributions are insensitive and more uncertain. From a visual inspection of Fig. 2.19, it can be seen that the parameters α, K_{bas}, P_{bas}, and P_{chn} are highly sensitive to the value of the objective function, NSE, because their distributions are relatively narrow. The remaining parameters have broader distributions, so they cannot be specified by any certain value. Additionally, marginal differences can be observed in the posterior distributions between the three models. The sensitivity results of SPCE are more analogous to those of SFM than for FPCE, especially for insensitive parameters (Fig. 2.19). Several posterior parameter distributions obtained from FPCE have a narrower shape than those obtained from both SFM and SPCE – see K_{bas}, P_{bas}, and Tl_{bas} for Event 1; α, f_1, R_{sa}, K_{bas}, P_{bas}, and Tl_{bas} for Event 4; and P_{bas}, and Tl_{bas} for Event 7.

Similar interpretations can be drawn with quantitative sensitivity analysis using the Sobol' index (Fig. 2.20). It can be confirmed that the four parameters $\boldsymbol{\alpha}$, $\boldsymbol{K_{bas}}$, $\boldsymbol{P_{bas}}$, and $\boldsymbol{P_{chn}}$ are the most sensitive parameters to the likelihood function, NSE, in all events. Specifically, the sensitivities of $\boldsymbol{K_{bas}}$ and $\boldsymbol{P_{bas}}$ are the largest for most events, and $\boldsymbol{\alpha}$ and $\boldsymbol{P_{chn}}$ are the next most sensitive parameters. For events with smaller return periods (e.g., Events 4 and 5), $\boldsymbol{P_{chn}}$ becomes more sensitive than the severe flood events. For medium to large streamflow (e.g., Events 1, 3, 7, and 9), the Sobol' index values of the four above-mentioned sensitive parameters in SPCE are more similar to those in SFM than in FPCE (Fig. 2.20).

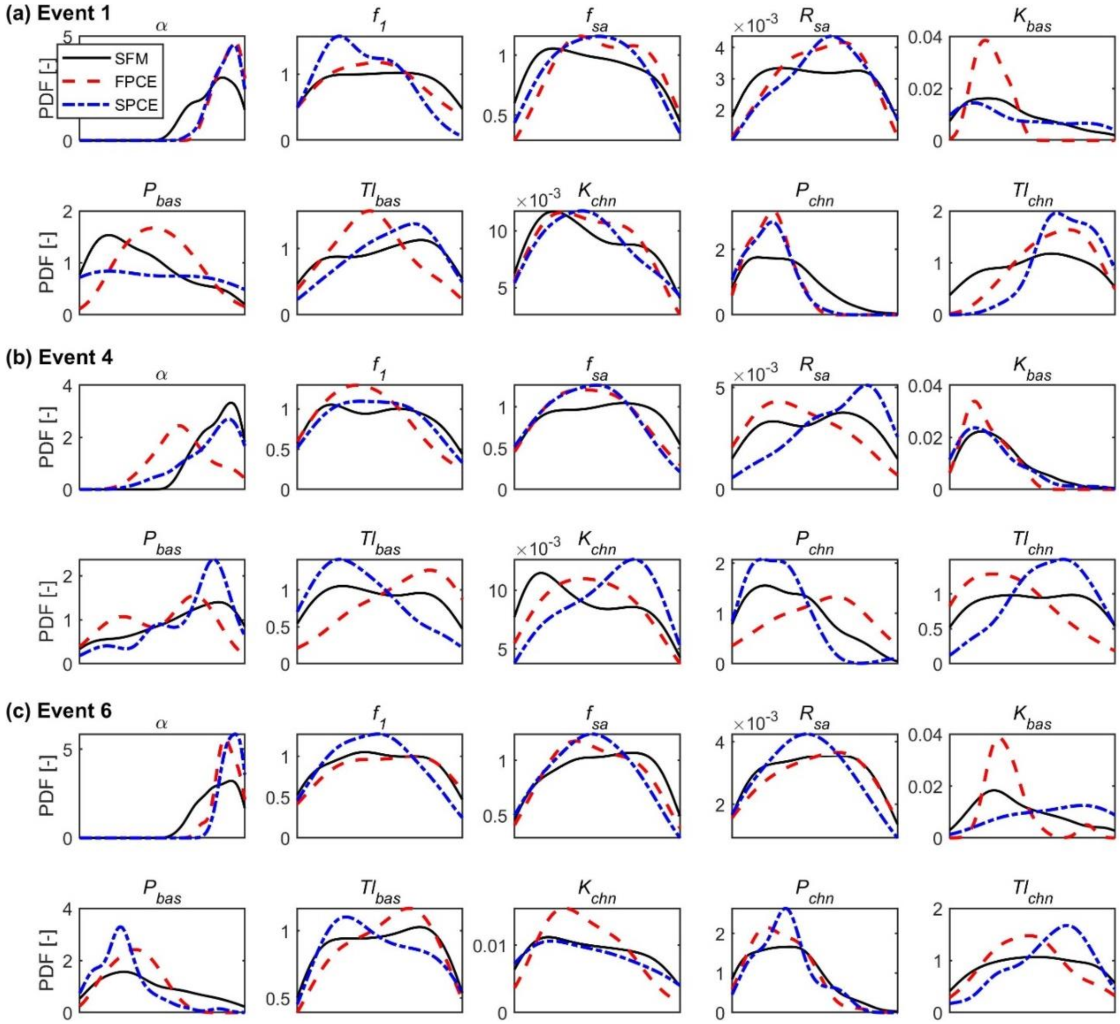


Figure 2.19. Posterior distributions of 10 model parameters for 3 streamflow events. In each subplot, probability density functions (PDFs) are drawn by using the kernel density estimation for the 1,000 behavior parameters obtained through GLUE. The range on the x-axis matches the original range values for each parameter presented in Table 2.12. Results for high, medium, and low return periods are only demonstrated for simplicity.

Identification of principal parameters through SA can improve efficiency in the process of optimizing parameters [*Zhang et al.*, 2013]; through the analysis of the interactions, influences, and correlations among parameters, we can support a better understanding of the process mechanisms of hydrological systems [*Ricciuto et al.*, 2018; *Dwelle et al.*, 2019; *Tran and Kim*, 2019; *Wang et al.*, 2020]. Besides these benefits, SA helps to construct a more efficient surrogate model embracing only a subset of principal parameters. Thus, the number of PCE coefficients and the size of the experimental design could be reduced, minimizing the complexity of the model.

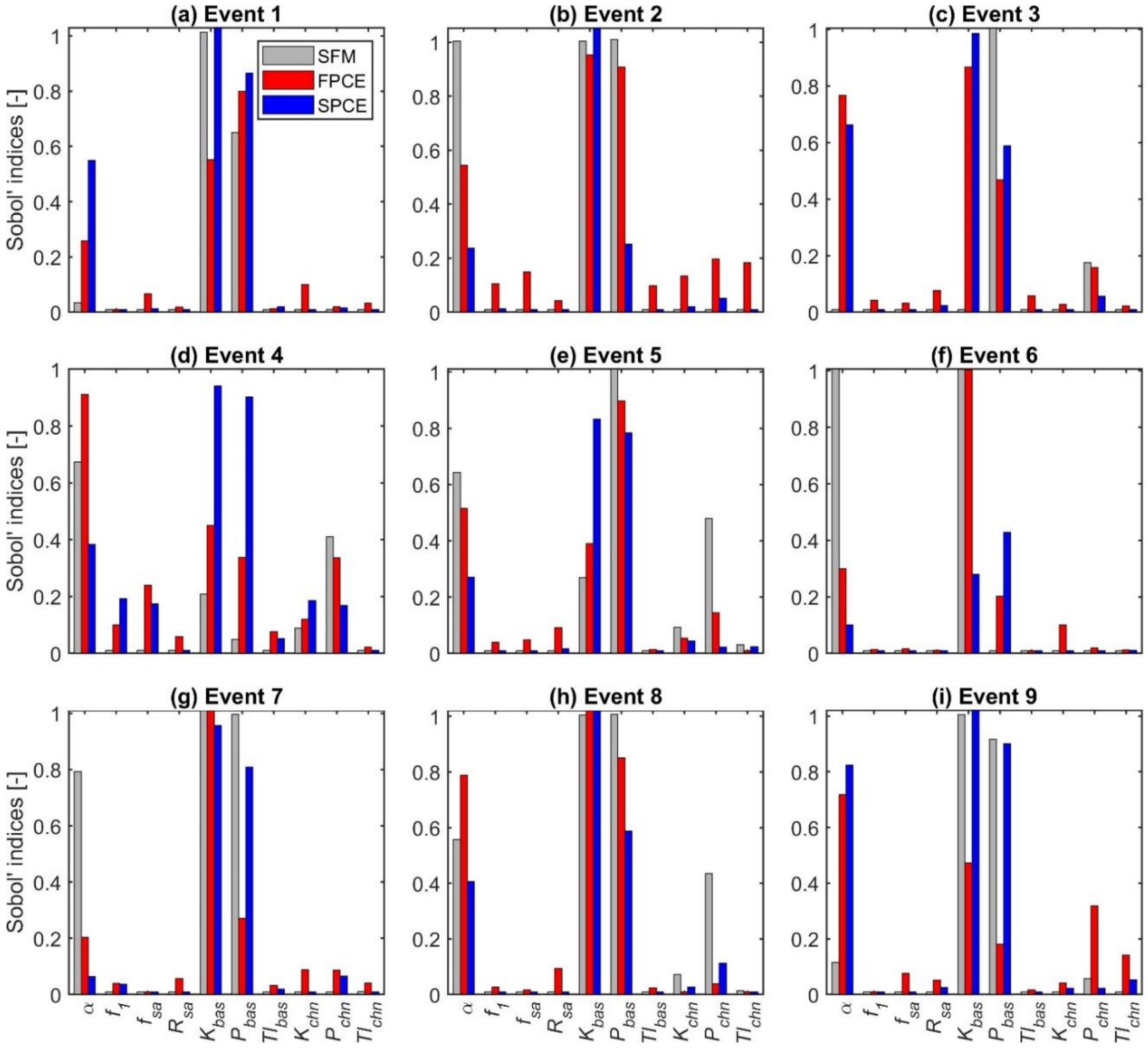


Figure 2.20. Sobol' sensitivity analysis for the ten parameters of SFM (grey), FPCE (red), and SPCE (blue), computed for the objective function of NSE over nine streamflow events.

2.3.5.4 The efficiency of surrogate models

To investigate efficiency performance, all simulations were implemented under the same computer configuration (CPU Intel(R) Xeon(R) CPU E5-4660 v4 @ 2.20GHz). The total time required for executing the (surrogate) models includes building time and runtime. The building time consists of the time for evaluating the experimental design and the time for estimating the PCE coefficients; the runtime refers to the time for performing ensemble simulations (Fig. 2.21 and Table 2.16). Note that the total runtime of SFM includes only the runtime, that is, the building time is zero. Table 2.16 shows the comparisons of the total runtime to obtain 100,000 ensemble runs among models for nine streamflow events. Although the total runtime may vary depending on the duration of the event, SFM took 12 to 30 days to perform 100,000 ensemble runs, while it took 6.1 to 14.3 hours for FPCE and only 1.5 to 3.6 hours for SPCE to produce the same number of ensembles. In other words, the degree of efficiency improvement can be up to about 50 times for FPCE relative to SFM and up to about 200 times for SPCE. The efficiency increases for greater than 100,000 ensemble runs (Fig. 2.21b). For example, SPCE can complete even 10,000,000 ensemble runs within 2 to 4 hours, whereas SFM can take up to several years. When comparing the total runtime between the surrogate models, SPCE is about four times faster than FPCE. The main reason for such a difference in efficiency is that the size of the experimental design required in SPCE is smaller (N = 500 vs. 2,000 in FPCE). Thus, the time to secure the experimental design is about four times shorter than that of FPCE (Table 2.16 and Figure 2.21a). Also, with fewer polynomial terms used, the runtime of SPCE is faster than that of FPCE about 12-14 times (see (ii) in Table 2.16). The ability of SPCE to perform thousands of model runs in a very short wall time enables computational problems that require a significant number of iterative calls, such as local or global optimization, data assimilation, and sensitivity analysis, to be solved efficiently [*Tran et al.*, 2020].

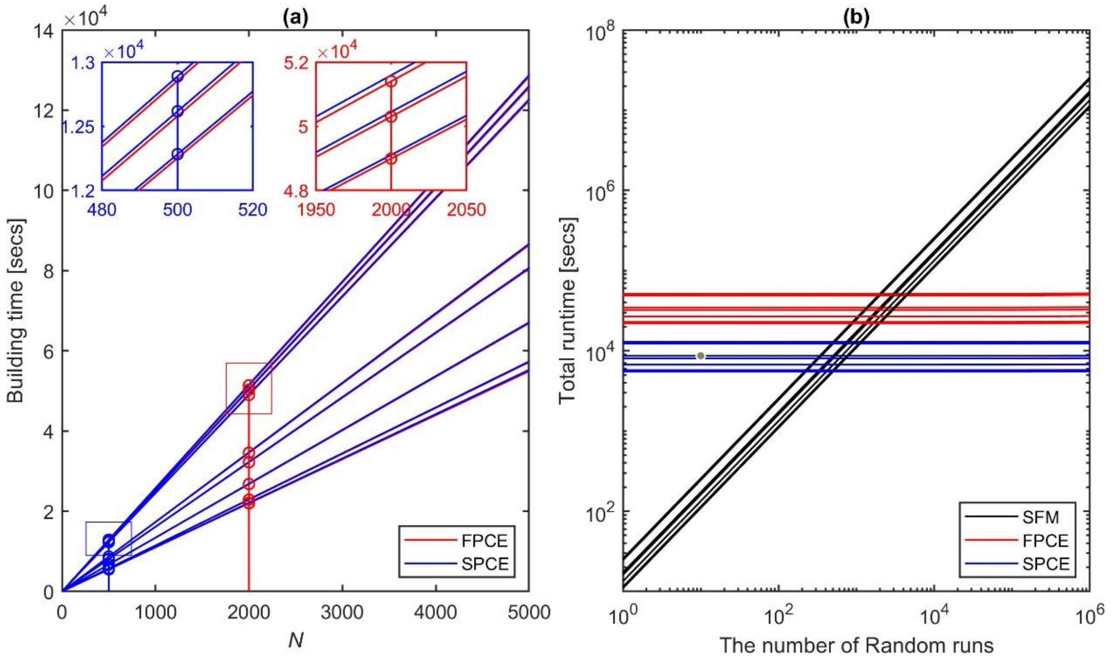


Figure 2.21. (a) Building time of FPCE and SPCE versus the size of experimental design (N) for 9 streamflow events. The building times at the optimal N = 2,000 for FPCE and at N = 500 for SPCE are used for (b) (see the stem plots and zoom-in sub-boxes in (a)). (b) Total runtime needed for carrying out the number of model (SFM, FPCE, and SPCE) runs (from 1 to 1,000,000 on x-axis) for the 9 events. Note that the intercepts of FPCE and SPCE in (b) are equal to the building times computed in (a); and the intercepts of SFM are zero.

Table 2.16. Comparisons of the total runtime for 9 streamflow events. The total runtime consists of (i) the building time and (ii) the running time. In surrogate models (FPCE and SPCE), the additional building time consists of (i-1) the time to secure the experiment design (i.e., the optimal 2,000 runs for FPCE and 500 for SPCE) and (i-2) the time to compute the PCE coefficients. (ii) the latter runtime refers to the time for performing 100,000 ensemble model (SFM, FPCE, and SPCE) simulations. The unit of values is in seconds.

Event	SFM	FPCE				SPCE			
	Total	Total	(i)		(ii)	Total	(i)		(ii)
	(ii)	(i)+(ii)	(i-1)	(i-2)		(i)+(ii)	(i-1)	(i-2)	
1	2569231	51473	51385	27	61	12894	12846	43	5
2	2448715	49066	48974	29	63	12287	12244	39	4
3	1728473	34634	34569	21	44	8679	8642	33	4
4	1610260	32260	32205	17	38	8078	8051	24	3
5	1144300	22923	22886	11	26	5745	5722	21	2
6	2514114	50373	50282	29	62	12622	12571	46	5
7	1338584	26818	26772	14	32	6716	6693	20	3
8	1098940	22016	21979	11	26	5514	5495	17	2
9	1102654	22090	22053	11	26	5531	5513	16	2

2.4 Conclusions

A unified framework for quantitatively performing uncertainty analysis in hydrologic flow prediction is presented in this Chapter. In the framework, the posterior probabilities of uncertain parameters were inferred using the GLUE method, and their uncertainty propagation was quantified efficiently using PCE theory. The applicability of this framework is confirmed through two case studies, including the Vu Gia river watershed in Vietnam and the Hongcheon river watershed in South Korea, and the corresponding two models used that are NAM and SFM, respectively.

For the first case study, overall results demonstrate that both GLUE and PCE methods have produced reasonable uncertainty when compared with observed data. The principal outcomes are summarized as follows:

- First, the results of the PCE model confirmed its ability to mimic a deterministic model and quantify its uncertainty for hydrologic prediction. The PCE model results are comparable to NAM results as well as observation – this can be validated from the high degree of matching between NSE, PE or VE values. We also investigated the performance of a PCE model for reducing computation time. Particularly, the efficiency of the PCE model increases significantly as the ensemble size for uncertainty quantification is increased. This increase is up to seventeen times faster (~18 minutes versus ~5 hours in Event 5) for an ensemble size of 100,000. Therefore, the efficiency of the PCE model is worthy of attention from a practical point of view and that flood forecasting results should be provided early for flood warning or risk management.

- Second, despite the significant advantages of GLUE, its inherent difficulty lies in the subjectivity related to the selections of the likelihood function and cutoff threshold. To determine the cutoff threshold values, we designed two indices (AI and EI) such that their tendencies are opposite. The intersection point of these two index curves can be accounted for as an optimal value that meets some degree of accuracy and efficiency at the same time, referred to as *behavioral acceptance threshold*. In this research, we determined the behavioral acceptance threshold as the average over the results of 8 events, resulting in a value of 0.82 for NSE, 4.05% for PE, and 4.35% for VE. Furthermore, a way of resolving the number of ensemble behavioral sets was presented to maintain a sufficient range of uncertainty and to avoid any unnecessary computation. An ensemble size of 500 was determined based on a visual inspection of Fig. 2.6 for the uncertainty characteristics of hydrographs (i.e., flood peak and flood volume).
- Third, the total time needed for estimating the uncertainty of a PCE model consists of (i) the time required to carry out 500 PCE behavioral runs, (ii) the time to determine the number of experimental designs and (iii) the time to compute the PCE coefficients (Table 8). The least square regression method was employed to efficiently estimate the PCE coefficients. This procedure is affected by the number of experimental designs (N) and degree of polynomial (p). Figs. 7 and 8 illustrate the effects of N and p on the accuracy of the PCE model results (i.e., NSE, PE, VE, and R^2), and we concluded that a N of 50 and a p of 4 would be preferred.
- Last, both the visual inspection of the posterior distribution and the mathematical computations of the Sobol' and Morris indices were investigated to see if constraining the range of dominant parameters could enhance efficiency. Two parameters ("CQOF"

and “CK12”) turned out to be more sensitive than the other seven parameters. Excluding these dominant parameters from the calibration process (by using Case 1 to 3) helped to determine the behavioral sets even faster, and thus it can be one of the most efficient ways to decrease runtime.

Regarding the second study, an approach to building an efficient surrogate model is focused. Specifically, SPCE and LAR were combined to allow for efficient construction of a surrogate model and fast quantification of its uncertainty for hydrological predictions. The essence of LAR is to learn and retain only the most significant polynomial basis terms, resulting in a sparse set of PCE coefficients that could be estimated more straightforwardly. The advantages of SPCE were investigated in comparison to the performance of a surrogate model (FPCE) constructed using ordinary least square regression (OLS), as follows:

- The performance of SPCE is superior to FPCE because SPCE can build a more accurate surrogate model (i.e., smaller LOO) with an experimental design of one-quarter the size (i.e., 500 versus 2,000).
- Streamflow results obtained through GLUE demonstrated that SPCE could sufficiently capture the uncertainty of the streamflow, which is comparable to that of SFM (see high degree of agreement for NSE and PE).
- Sensitivity analysis attained through visual inspection of the posterior parameter distributions and mathematical computation of the Sobol’ index has been of great success for SPCE to capture the parameter sensitivity of SFM in middle to high flow predictions. In all models and in all events, the four parameters α, K_{bas}, P_{bas}, and P_{chn} were most sensitive to the likelihood function, NSE.

- The computational power of SPCE is about 200 times faster than SFM and about four times faster than FPCE when executing 100,000 ensemble runs. This efficiency enhancement of SPCE is particularly important when larger ensemble runs are needed.

Overall, this Chapter has provided an efficient framework to quantify the uncertainty in hydrological model simulations without sacrificing accuracy as compared to typical deterministic model results. Although we applied the methodology to two watersheds and two lumped models, it is expected to work well when estimating streamflow for other regions and other distributed models as well. One can maximize efficiency, especially when using complex models. Therefore, the proposed framework will ultimately be helpful for providing flood warnings and mitigating flood risk in a timely manner.

CHAPTER III

A novel modeling framework for real-time ensemble flood forecasting with uncertainty quantification

"There's a way to do it better, find it"

- *(Edison, TA)*

3.1 Introduction

Real-time forecasting is an important component of flood risk management and mitigation but is subject to multiple uncertainties caused by meteorological inputs, initial states, model structures, and model parameters [*Beven*, 1989; *Ajami et al.*, 2007; *Moradkhani and Sorooshian*, 2008; *Mockler et al.*, 2016]. Due to the complexities of natural phenomena represented by equifinality [*Beven and Freer*, 2001; *Beven*, 2006], hysteresis [*Wei and Dewoolkar*, 2006; *Ivanov et al.*, 2010; *Fatichi et al.*, 2015; *Fatichi et al.*, 2016b], non-uniqueness [*Beven*, 2000; *McKenna et al.*, 2003; *Kim and Ivanov*, 2014; *Kim et al.*, 2016a], non-linearity [*Kitanidis and Bras*, 1980; *Xie and Zhang*, 2010; *Kim and Ivanov*, 2015], and internal variability [*Nikiema and Laprise*, 2011; *Mondal and Mujumdar*, 2012; *Lafaysse et al.*, 2014; *Fatichi et al.*, 2016a; *Kim et al.*, 2016c; *Kim et al.*, 2016b; *Kim et al.*, 2018], perfect predictions using numerical models are infeasible. The problem exacerbates, if one attempts to simulate constitutive models derived from empirical or phenomenological observations rather than basic conservation laws of physics that would also require embracing a large number of parameters. Forecasting systems must therefore rely on approaches with intrinsic tools to quantify and reduce associated uncertainties and allow end-users to make informed decisions [*Todini*, 1999; 2004].

Forecasts made with sufficient lead time can mitigate flood damages considerably. Results should therefore be provided within a predetermined time horizon and accurate enough to promote community confidence in actions taken for emergency preparedness [*Todini*, 2004; *APFM*, 2013]. Extensive efforts have been devoted to enhance forecast accuracy, predictability, and efficiency in real time with uncertainty quantification (Table 3.1). However, simultaneous improvement of predictive accuracy and efficiency, while evaluating effectiveness, remains a major challenge [*Liu et al.*, 2012; *Cintra and Velho*, 2018].

For the purpose of enhancing model accuracy in real-time flood forecasting where no information of model states and parameters is available, data assimilation (DA) has been proven useful. Due to the nature of forecasting, the effect of future unknowns (model parameters and states) on flood prediction will change over time. In addition, uncertainty can be amplified not only by the features of the model itself, but also by errors in forcing data and observations. Therefore, model adjustment for the forecasting period may be necessary [*Young*, 2002; *Moradkhani et al.*, 2005c]. Several assimilation methods have been developed using Kalman or particle filters and optimization or inference techniques such as the back-fitting algorithm [*Zhang et al.*, 2018b], shuffled complex evolution algorithm [*Li et al.*, 2014], shuffled complex evolution metropolis [*Vrugt et al.*, 2005], generalized likelihood uncertainty estimation (GLUE) [*Beven and Freer*, 2001], and sequential Bayesian combination [*DeChant and Moradkhani*, 2014]. Due to the higher computational requirements of the latter techniques, filter-type approaches have attracted attention as assimilation tools [*Moradkhani and Sorooshian*, 2008; *Gharamti et al.*, 2013].

Table 3.1. Literature review of applications involving real-time, ensemble streamflow forecasting. The last column corresponds to the taxonomy of the predictive approach (**A**) that we define in Table 3.2. "Warm-up" methods have a warm-up period, while "Arbitrary" do not.

Study	**Deterministic model**	**Surrogate model**	**Parameter specification**	**State initialization**	**DA**	**A**
Zhang et al. [2018b]	Xinanjiang	-	Optimization	Warm-up	Dual	A12
Abbaszadeh et al. [2018]	SAC-SMA	-	Random	Arbitrary	Dual	A12
Wang et al. [2018]	HyMOD	PCE	NA	Warm-up	Dual	A6
Davison et al. [2017]	MESH	-	Random	NA	Dual	A3
Thiboult et al. [2016]	Multimodels	-	NA	Warm-up	Single	A2
Fan et al. [2016]	HyMOD	PCE	Random	NA	Dual	A6
Zahmatkesh et al. [2015]	HyMOD, HBV, SWMM	-	Bayesian inference	Warm-up	None	A10
Li et al. [2014]	GR4H	-	Optimization	NA	Dual	A12
DeChant and Moradkhani [2014]	VIC	-	NA	Warm-up	Dual	A3
Xie and Zhang [2013]	SWAT	-	Random	Warm-up	Dual	A3
Chen et al. [2013]	HyMOD	-	Bayesian inference	NA	Single	A11
Moradkhani et al. [2012]	HyMOD	-	Random	Warm-up	Dual	A3
He et al. [2012]	SNOW17+ SAC-SMA	-	Bayesian inference	Warm-up	Single	A11
Mendoza et al. [2012]	TopNet	-	Manual calibration	Warm-up	Single	A11
Clark et al. [2008]	TopNet	-	Bayesian inference	Warm-up	Single	A11
Ajami et al. [2007]	HyMOD, SWB	-	Bayesian inference	Warm-up	None	A10
Weerts and El Serafy [2006]	HBV-96	-	NA	NA	Single	A2
Vrugt et al. [2005]	HyMOD	-	Random	Arbitrary	Dual	A3
Moradkhani et al. [2005b]	HyMOD	-	Random	Arbitrary	Dual	A3
Madsen and Skotner [2005]	Mike 11	-	Optimization	Warm-up	Single	A11
Beven and Freer [2001]	TOPMODEL	-	Bayesian inference	Warm-up	Dual	A12

NA: Not Available

Currently, the ensemble Kalman filter (EnKF) [*Evensen*, 1994] and its modifications (e.g., ensemble Kalman smoothers, ensemble square-root filters, and gain function) are the most commonly used techniques in the hydrology community (Table 3.1), despite the issue of slow convergence caused by intrinsic assumptions, especially for domains with complexities [*Moradkhani et al.*, 2005b; *Weerts and El Serafy*, 2006; *Moradkhani et al.*, 2012; *Wang et al.*, 2017]. Recent studies have suggested that particle filtering (PF) [*Arulampalam et al.*, 2002] is an alternative method to resolve the inclusion of unrealistic Gaussian assumptions in the EnKF. The PF method has more advantages than EnKF in reducing numerical instability by providing particle weights and using non-Gaussian state-space models [*Liu et al.*, 2012]. However, this method is computationally more expensive as it generally requires more ensemble members [*Moradkhani et al.*, 2005b; *Liu et al.*, 2012].

When assimilating data, model parameter specification and state initialization may play a crucial role, especially for short-range forecasting [*Houtekamer and Zhang*, 2016]. Generally, ensemble initialization of model states and parameters for the forecasting period can be generated approximately, e.g., using a random selection from uniform distributions for parameters and setting up the initial state values as an arbitrary number (e.g., zero) at the beginning of the forecasting period [*Moradkhani et al.*, 2005b; *Vrugt et al.*, 2005; *Moradkhani et al.*, 2012; *Xie and Zhang*, 2013; *DeChant and Moradkhani*, 2014; *Davison et al.*, 2017; *Abbaszadeh et al.*, 2018]. Alternatively, the ensemble can be generated more carefully, e.g., specifying parameters from relevant distributions [*Beven and Freer*, 2001; *Madsen and Skotner*, 2005; *Ajami et al.*, 2007; *Clark et al.*, 2008; *He et al.*, 2012; *Mendoza et al.*, 2012; *Chen et al.*, 2013; *Zahmatkesh et al.*, 2015] and using a warm-up technique for states [*Ajami et al.*, 2007; *He et al.*, 2012; *Mendoza et al.*, 2012; *DeChant and Moradkhani*, 2014; *Wang et al.*, 2018], as summarized in Table 3.1.

The assimilation techniques described above generally require a large number of model evaluations to update parameter and state values and present predictive uncertainties, leading to computational challenges [*Vrugt et al.*, 2008c; *Vrugt*, 2016; *Zhang et al.*, 2017], even with the benefit of parallel computation with multiple processors [*Cintra and Velho*, 2018]. Because keeping calculation time to a minimum is a key element for timely flood warnings and responding to emergency situations [*Ballio and Guadagnini*, 2004; *Sene*, 2008], it is necessary to find alternatives that significantly increase forecast lead time. Surrogate modeling can address this challenge by substituting computationally intensive models with computationally efficient metamodels, such as the polynomial chaos expansion (PCE). Through the expansion of orthogonal polynomials, approximate functions can be constructed and applied to hydrologic models. Recent studies have used PCE to perform robust uncertainty assessment of diverse hydrologic problems [*Sochala and Le Maître*, 2013; *Fan et al.*, 2014; *Wu et al.*, 2014; *Wang et al.*, 2015; *Fan et al.*, 2016; *Wang et al.*, 2017; *Wang et al.*, 2018; *Dwelle et al.*, 2019] rather than running deterministic models. However, few studies have tested its effectiveness in a setting of real-time flood forecasting [*Wang et al.*, 2015; *Fan et al.*, 2016; *Wang et al.*, 2017; *Wang et al.*, 2018].

To fill the above gaps, we propose a novel integrated modeling framework that improves accuracy, predictability, and efficiency of real-time flood forecasting. Eighteen approaches to the framework are presented, combining ways of constructing the surrogate models, specifying model parameters and states, and assimilating newly observed data. This Chapter investigates (i) the effects of building methods of the PCE model and its capacity for real-time flood forecasting; (ii) the effects of specifying methods on predictive performance; (iii) the effects of single- and dual-assimilation techniques; and (iv) the computational time of the proposed approaches.

3.2 Methods

3.2.1 New invariant surrogate model: polynomial chaos expansion

Polynomial chaos expansion (PCE) [*Wiener*, 1938; *Ghanem and Spanos*, 1991] can build a surrogate model ($\boldsymbol{\mathcal{M}}^{PCE}$) for any (deterministic rainfall-runoff) model ($\boldsymbol{\mathcal{M}}$) through the expansions of orthogonal polynomials. This enables a polynomial approximation of the model through its deterministic input/output relationship. The form of a PCE model approximating a model output (e.g., streamflow y_t) as a function of model parameters $\boldsymbol{\theta}_t$ is given as:

$$y_t = \boldsymbol{\mathcal{M}}(\boldsymbol{\theta}_t) \approx \boldsymbol{\mathcal{M}}^{PCE_t}(\boldsymbol{\theta}_t) \tag{3.1}$$

Note that the surrogate model ($\boldsymbol{\mathcal{M}}^{PCE}$) in Eq. 3.1 has the subscript of t, indicating that the surrogate model is a collection of PCEs constructed at each time step of interest. Also, only the parameter $\boldsymbol{\theta}_t$ (this includes a subscript of t as well) is chosen as an input variable during PCE construction, and other forcing or state inputs required to simulate hydrologic models are held constant [*Sochala and Le Maître*, 2013; *Fan et al.*, 2016; *Meng and Li*, 2018; *Wang et al.*, 2018; *Dwelle et al.*, 2019; *Tran and Kim*, 2019]. This mathematical formulation conveys that PCE should be built separately for each time step at which a meteorological condition or model state is updated.

Unlike previous studies based on Eq. 3.1, this work constructs the surrogate PCE model with Eq. 3.2, which has three characteristics: (i) the model input consists of meteorological data, model states, and model parameters; (ii) model parameters do not change over time, which is different from Eq. 3.1; and (iii) there is no need to constantly create the PCE model over time (which is the most important practical feature). The single PCE model represents streamflow phenomena over the entire calibration period during which the PCE model was generated. Specifically, ensemble model output ($\boldsymbol{Y_t}$) at each time step, including streamflow (y_t) and states ($\boldsymbol{x}_t$), can be written as a function of model inputs ($\boldsymbol{X}_t$**),** including states ($\boldsymbol{x}_{t-1}$), climate data ($\boldsymbol{u}_t$), and time-invariant parameters ($\boldsymbol{\theta}$):

$$\boldsymbol{Y_t} = \boldsymbol{\mathcal{M}}(\boldsymbol{X}_t) \approx \boldsymbol{\mathcal{M}}^{PCE}(\boldsymbol{X}_t) = \sum_{\alpha=0}^{N_\Psi - 1} \varepsilon_\alpha \Psi_\alpha(\boldsymbol{X}_t) \tag{3.2}$$

$$\boldsymbol{Y_t} = [y_t\ \boldsymbol{x}_t],\ \boldsymbol{X}_t = [\boldsymbol{x}_{t-1}\ \boldsymbol{\theta}\ \boldsymbol{u}_t] \tag{3.3}$$

where $\Psi_{\alpha}(\boldsymbol{X}_t)$ represents the multivariate polynomials corresponding to the given input $\boldsymbol{X}_t$. The polynomials are constructed as the product of univariate orthonormal polynomials:

$$\Psi_{\alpha}(\boldsymbol{X}_t) = \prod_{j=1}^{N_X} \Psi_{\alpha_j}^{(j)}\left(\boldsymbol{X}_t^j\right) \tag{3.4}$$

where the size of $\boldsymbol{X}_t$, N_X, is equal to the summation of the number of parameters, states, and forcing inputs of the deterministic model (i.e., $N_X = N_P + N_S + N_I$). Note that the NAM model is selected to implement in this Chapter, that is, $N_P = 9$, $N_S = 5$, and $N_I = 1$ are used. More information about the NAM model can be found in Section 2.2.1.2. Thus, the number of PCE coefficient can be determined as:

$$N_{\Psi} = \frac{(N_X+p)!}{N_X!p!} \tag{3.5}$$

Given the set of multivariate orthonormal polynomials $(\Psi_{\alpha}(\boldsymbol{X}_t))$, the next step is to compute the PCE coefficients (ε_{α}), which are influenced by the number of experimental designs (N) and the polynomial degree, p [*Blatman and Sudret*, 2010; *Blatman and Sudret*, 2011]. For this work, the least-squares regression (OLS) is adopted. The detail of this method can be found in Section 2.2.1.1. According to the approach by *Blatman and Sudret* [2010], a metric of the leave-one-out (LOO) cross-validation error in Eq. 2.21 can illustrate the performance of the PCE model.

3.2.2 Parameter inference: GLUE

GLUE [*Beven and Binley*, 1992] refers to a series of procedures for inferring parameter posterior distributions and quantifying the associated uncertainties. The objective of GLUE is to select "behavioral" model runs based on the threshold values of likelihood functions with observations, among a large number of runs simulated with random combinations of parameter values. The latter parameter's values can be sampled randomly from the prior distributions of each parameter (constrained in this study with upper and lower bounds of Table 2.3) using Monte Carlo or Latin hypercube sampling (LHS). For more efficient performance, LHS was used [*Helton and Davis*, 2003]. The likelihood functions proposed are three metrics of Nash–Sutcliffe efficiency (NSE), peak error (PE), and volume error (VE) defined in Eqs. (2.9-2.11), representing the model performance with respect to the shape, peak, and volume of hydrograph, respectively. Acceptance threshold values are determined according to an approach presented in Section 2.2.3.1 in which relationships between accuracy and efficiency indices are identified for their determinations. Specifically, cutoff threshold values for the likelihood functions of NSE, PE, and VE are suggested as 0.8, 5%, and 5%, respectively. The model runs (or parameters) that satisfy the modelling error within the above thresholds for all the likelihood functions are defined here as "behavioral" runs (or parameters).

3.2.3 Data assimilation: Ensemble Kalman filter

Among many reported techniques, the single ensemble Karman filter (EnKF) and the dual-ensemble Karman filter (dual EnKF) are often chosen to optimally update the ensemble of model states (and parameters) of forecasting systems with real-time observations, which can be coupled with any models [*Evensen*, 1994; *Burgers et al.*, 1998; *Moradkhani et al.*, 2005c; *Whitaker*, 2012].

3.2.3.1 States updated

An ensemble of state vector, $\boldsymbol{x}$ consisting of n ensemble members by N_S is propagated through $\boldsymbol{Model}$ of both deterministic model and PCE models, such that each state vector represents one realization of the model states. Then, the state forecast is made for each ensemble member as follows (forecast step):

$$\boldsymbol{x}_t^{i-} = f\left(\boldsymbol{x}_{t-1}^{i+}, \boldsymbol{\theta}^i, \boldsymbol{u}_t\right) + w_t^i, \qquad i = 1, \dots, n \tag{3.6}$$

where $\boldsymbol{x}_t^{i-}$ is the i-th forecasted states vector at time t, $\boldsymbol{x}_{t-1}^{i+}$ is the i-th updated states vector at time $t-1$, N_S is the number of model states $\boldsymbol{x} = \{x_j,\ j = 1, \dots, N_s\}$, and n is the number of ensemble members. The nonlinear propagator $f(\cdot)$ contains N_I model input vector $\boldsymbol{u}_t, \{u_{1,t}, \dots, u_{N_I,t}\}$ and the i-th model parameter vector $\boldsymbol{\theta}^i$ corresponding to the model state $\boldsymbol{x}_{t-1}^{i+}$. The term w_t^i is the i-th model error and presents all uncertainty related to model structure, forcing data and model parameter [*Moradkhani et al.*, 2005c]. In this work, the model error is represented by the uncertainty of model parameters.

Suppose that the actual observation (y_{t+1}^{obs}) is taken at time $t+1$ and that we intend to assimilate the vector of observations into the model. The predicted output of model, $y_{t\ +1}^i$ at time $t+1$ is computed with the propagator $h(\cdot)$ as a function of $\boldsymbol{\theta}^i$, $\boldsymbol{u}_{t+1}$, and $\boldsymbol{x}_t^{i-}$, which can be written as:

$$y_{t+1}^i = h\left(\boldsymbol{x}_t^{i-}, \boldsymbol{\theta}^i, \boldsymbol{u}_{t+1}\right) \tag{3.7}$$

To represent the error statistics in the forecast step, we assume that at time $t+1$, we have an ensemble of n forecasted states, $\boldsymbol{x}_t^- \triangleq (\boldsymbol{x}_t^{1-}, \dots, \boldsymbol{x}_t^{n-})$ and an ensemble of n forecasted outputs, $y_{t+1} \triangleq (y_{t+1}^1, \dots, y_{t+1}^n)$. Then the ensemble means of forecasted state ($\overline{\boldsymbol{x}}_t^-$) and the ensemble mean of forecasted output ($\bar{y}_{t+1}$) are estimated by:

$$\overline{\boldsymbol{x}}_t^- \triangleq \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_t^{i-} \tag{3.8}$$

$$\bar{y}_{t+1} \triangleq \frac{1}{n}\sum_{i=1}^n y_{t+1}^i \tag{3.9}$$

Then, we define the ensemble error matrix of forecasted state, E_{t+1}^{-} around the ensemble mean by:

$$E_{t+1}^{-} \triangleq [\boldsymbol{x}_t^{1-} - \overline{\boldsymbol{x}}_t^{-} \dots \boldsymbol{x}_t^{n-} - \overline{\boldsymbol{x}}_t^{-}] \tag{3.10}$$

and the ensemble of output error matrix, E_{t+1}^{y} is:

$$E_{t+1}^{y} \triangleq [y_{t+1}^{1} - \bar{y}_{t+1} \dots y_{t+1}^{n} - \bar{y}_{t+1}] \tag{3.11}$$

The error covariance matrix is calculated including:

- The error covariance matrix of ensemble forecast state:

$$\boldsymbol{Q}_{t+1}^{x} = \frac{1}{n-1} E_{t+1}^{-} (E_{t+1}^{-})^{\mathrm{T}} \tag{3.12}$$

- The error covariance matrix of model output:

$$\boldsymbol{Q}_{t+1}^{y} = \frac{1}{n-1} E_{t+1}^{y} (E_{t+1}^{y})^{\mathrm{T}} \tag{3.13}$$

- The forecast cross-covariance of the states and output:

$$\boldsymbol{Q}_{t+1}^{xy} = \frac{1}{n-1} E_{t+1}^{-} (E_{t+1}^{y})^{\mathrm{T}} \tag{3.14}$$

In order for the EnKF to maintain sufficient spreads in ensemble and to prevent from filter divergence [*Whitaker and Hamill*, 2002], observations should be treated as random variables. At each time, an observation is perturbed by adding noise drawn from a Gaussian distribution of mean zero and predefined covariance [*Burgers et al.*, 1998]. Thus, in the updated step, the forecasted state set $\boldsymbol{x}_{t+1}^{i-}$ is updated using the Kalman gain K_{t+1}^{x} as follow:

$$\boldsymbol{x}_t^{i+} = \boldsymbol{x}_t^{i-} + K_{t+1}^{x} \left(y_{t+1}^{obs,i} - y_{t+1}^{i} \right) \tag{3.15}$$

where $y_{t+1}^{obs,i}$ is the i-th trajectory of the observation replicates generated by adding to the actual observation (y_{t+1}^{obs}) error, η (i.e., a perturbation to observation) that has zero mean and the covariance, $E_{t+1}^{y^{obs}}$, which is determined as follow:

$$y_{t+1}^{obs,i} = y_{t+1}^{obs} + \eta_{t+1}^{i},\ \eta_{t+1}^{i} \sim N\left(0, E_{t+1}^{y^{obs}}\right) \tag{3.16}$$

The Kalman gain matrix can be calculated by:

$$K_{t+1}^{x} = \boldsymbol{Q}_{t+1}^{xy}\left[\boldsymbol{Q}_{t+1}^{y} + \boldsymbol{Q}_{t+1}^{obs}\right]^{-1} \tag{3.17}$$

where $\boldsymbol{Q}_{t+1}^{obs}$ is the covariance matrix of the observation, $y_{t+1}^{obs,i}$, which is defined similar to $\boldsymbol{Q}_{t+1}^{y}$.

$$\boldsymbol{Q}_{t+1}^{obs} = \frac{1}{n-1} E_{t+1}^{obs}(E_{t+1}^{obs})^{\mathrm{T}} \tag{3.18}$$

$$E_{t+1}^{obs} \triangleq \left[y_{t+1}^{obs,1} - y_{t+1}^{obs} \ldots\ y_{t+1}^{obs,n} - y_{t+1}^{obs}\right] \tag{3.19}$$

3.2.3.2 Dual parameters-states updated

The dual EnKF requires two interactive and parallel filters for the states and parameters estimation [*Moradkhani et al.*, 2005c]. The parameters are first updated and then the states. In order to extend the applicability of the single EnKF to the simultaneous parameters–states EnKF, one needs to treat the ensemble size of parameter sets similar to the model state. However, the parameter values are not changed after the forecast step:

$$\boldsymbol{\theta}_{t+1}^{i-} = \boldsymbol{\theta}_{t}^{i+} \tag{3.20}$$

Using the parameters forecasted and the replicates of forcing data, states of the ensemble model and model prediction are computed as follows:

$$\boldsymbol{x}_t^{i-} = f\left(\boldsymbol{x}_{t-1}^{i+}, \boldsymbol{\theta}_{t+1}^{i-}, \boldsymbol{u}_t\right) + w_t^i, \qquad i = 1, \ldots n \tag{3.21}$$

$$y_{t+1}^i = h\left(\boldsymbol{x}_t^{i-}, \boldsymbol{\theta}_{t+1}^{i-}, \boldsymbol{u}_{t+1}\right) \tag{3.22}$$

Updating the ensemble parameter member is made:

$$\boldsymbol{\theta}_{t+1}^{i+} = \boldsymbol{\theta}_{t+1}^{i-} + K_{t+1}^{\boldsymbol{\theta}}\left(y_{t+1}^{obs,i} - y_{t+1}^i\right) \tag{3.23}$$

where $K_{t+1}^{\boldsymbol{\theta}}$ is the Kalman gain for correcting the parameter trajectories obtained with:

$$K_{t+1}^{\boldsymbol{\theta}} = \boldsymbol{Q}_{t+1}^{\boldsymbol{\theta}y}\left[\boldsymbol{Q}_{t+1}^{y} + \boldsymbol{Q}_{t+1}^{obs}\right]^{-1} \tag{3.24}$$

where $\boldsymbol{Q}_{t+1}^{\boldsymbol{\theta}y}$ is the cross-covariance matrix of model parameters and model output. Now use the updated parameter $\boldsymbol{\theta}_{t+1}^{i+}$ to the step given in Section 3.2.3.1 to update the ensemble model states simultaneously.

3.3 New modeling framework

3.3.1 Obtaining prior and posterior parameter distributions of a deterministic model

The first preparation step of the modeling framework is to obtain the *prior* and *posterior* parameter distributions for a deterministic model. There could be various ways to handle this, but in this study the following assumptions and methodologies are specifically applied. We first assume that each of the parameters follows a uniform distribution within specified bounds – the prior parameter distributions are simply attained by utilizing prior-known information for the bounds in Table 2.3. In contrast, the posterior parameter distributions are fitted to the 500 behavior parameters of GLUE – the 500 NAM behavior samples are identified as an optimal number from Chapter II which has confirmed that more than the 500 parameter sets does not change the shape of the posterior distributions (see Section 2.2.3.1). For consistency, this number will be also used for making the posterior distributions of PCE-I and PCE-II in Sec. 3.3.2.

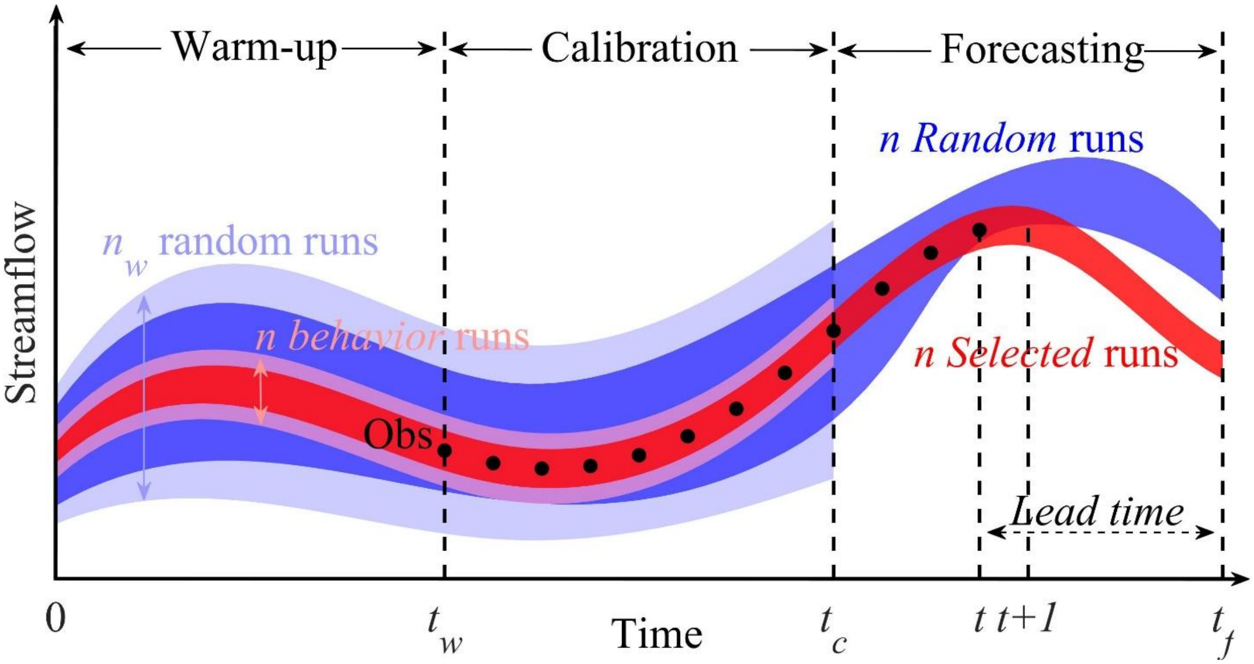


Figure 3.1. A schematic illustration for real-time ensemble flood forecasting, which consists of 3 intervals: warm-up, calibration, and forecasting periods. The light red shaded region in the warm-up and calibration periods refers to the n behavior results of GLUE that are employed to estimate posterior parameter distributions, while the light blue region refers to the n_w random results obtained from parameter sets sampled from prior (uniform) distributions to attain the n behavior runs. The construction of PCE models is carried out over the calibration period: PCE-I model is built from the N_{I} training samples extracted from the light blue region, while PCE-II is from the N_{II} samples from the light red region. The (dense) blue and red shaded regions correspond to the approaches using the "*Random*" (A1 to A9) and "*Selected*" (A10 to A18) parameter specifications with the same n ensemble runs, respectively.

The mathematical expression of this step is as follows. For the warm-up and calibration periods, a model $\boldsymbol{\mathcal{M}}$ (NAM) can be simulated to attain behavioral runs with GLUE, i.e.,

$$\left[y_t^{ii}\ \ \boldsymbol{x}_t^{ii}\right] = \boldsymbol{\mathcal{M}}\left(\boldsymbol{x}_{t-1}^{ii}, \boldsymbol{\theta}^{ii}, \boldsymbol{u}_t\right),\ ii = 1, \dots, n_w;\ t = 1, \dots, t_c \tag{3.25}$$

where n_w is the number of model runs to obtain the n number of the behavioral set based on the likelihood scores estimated with the GLUE method. Among the n_w random runs (referring to the light blue shaded region in Fig. 3.1) that are simulated by using parameter sets ($\boldsymbol{\theta}^{ii}$) sampled randomly from the prior (uniform) distributions, the only n behavior runs (referring to the light red shaded region in Fig. 3.1) are employed for making the posterior distributions.

Reducing the effects of uncertainty by initial conditions ($\boldsymbol{x}_0^{ii}$) is necessary for modeling. In this framework, a "warm-up" technique was employed to calibrate the deterministic model. Generally, a sufficient period of time (called the 'warm-up' period) can be set such that the influence of the initial condition is dissipated, and the warm-up is performed before entering the calibration period. This technique produces behavioral parameter sets much faster in GLUE, compared with cases that do not use the warm-up technique.

3.3.2 Building PCE with two types of experimental design

We propose two types of approaches for constructing the PCE model, depending on how the sample collections of the experimental design ($\boldsymbol{\mathcal{X}_t}$) is composed. One approach is to build a PCE model ("PCE-I") by collecting the training samples that are generated from the *prior* parameter distributions. The other approach is ("PCE-II") uses samples that are formed by the *posterior* parameters distributions. The associated mathematical expression is

$$\left[y_t^{iii}\ \ \boldsymbol{x}_t^{iii}\right] = \boldsymbol{\mathcal{M}}\left(\boldsymbol{x}_{t-1}^{iii}, \boldsymbol{\theta}^{iii}, \boldsymbol{u}_t\right),\ iii = 1, \dots, N;\ t = 1, \dots, t_c \qquad (3.26)$$

where the N_{I} set of $\boldsymbol{\mathcal{X}_t}$ (i.e., $N = N_{\mathrm{I}}$ for PCE-I) consists of model $\boldsymbol{\mathcal{M}}$ simulation results calculated from parameters sampled from the prior distributions (correspond to N_{I} set sampled randomly from the results in the light blue shaded region over the calibration period in Fig. 3.1) [*Sudret*, 2008; *Blatman and Sudret*, 2010; *Blatman and Sudret*, 2011]. In contrast, the experimental design of the latter approach assumes that the N_{II} set of $\boldsymbol{\mathcal{X}_t}$ (i.e., $N = N_{\mathrm{II}}$ for PCE-II) are drawn from the more constrained, posterior parameter distributions (correspond to the light red shaded region over the calibration period in Fig. 3.1). All the samples were taken through LHS sampling [*McKay et al.*, 1979b].

The former approach can be implemented easily and therefore has been used more commonly in the literature [*Sudret*, 2008; *Blatman and Sudret*, 2010; *Blatman and Sudret*, 2011]. However, for past periods in which observations exist, the second approach using a well-calibrated set of parameters is beneficial in significantly reducing computational time [*Tran and Kim*, 2019]. It takes less time to build PCE in the second approach because less training samples (N_{I} is generally larger than N_{II}) are required when estimating coefficients. On the other hand, in the context of real-time forecasting when no observations have been attained, the latter approach might cause a problem. Specifically, PCE models built with a set of "good" posterior parameters sets obtained only for a certain historic period of time would not necessarily demonstrate validity for unknown prediction periods. Evaluation of the applicability of the two approaches to real-time flood forecasting will be addressed in Section 3.5.

Once the PCE models were constructed, the same GLUE procedure is made to obtain the posterior parameter distributions of both PCE models:

$$\left[y_t^{ii}\ \ \boldsymbol{x}_t^{ii}\right] = \boldsymbol{\mathcal{M}}^{PCE}\left(\boldsymbol{x}_{t-1}^{ii}, \boldsymbol{\theta}^{ii}, \boldsymbol{u}_t\right),\ ii = 1, \dots, n_w;\ t = 1, \dots, t_c \qquad (3.27)$$

Note that the number n_w is different depending on $\boldsymbol{Model} =$ {NAM, PCE-I, PCE-II}.

3.3.3 Specifying model parameters for data assimilation

Determining initial conditions and parameter values before assimilating real-time observations over the forecasting period is a necessary step. The mathematical expression for preparing data assimilation (forecasting) is written as:

$$[y_t^i \; \boldsymbol{x}_t^i] = \boldsymbol{Model}(\boldsymbol{x}_{t-1}^i, \boldsymbol{\theta}^i, \boldsymbol{u}_t), \; i = 1, \ldots, n; \; t = 1, \ldots, t_c \tag{3.28}$$

where the initial ensemble of states ($\boldsymbol{x}_0^i$) is set to an arbitrary number (e.g., zero) at the beginning of simulation (i.e., $t = 0$) (Fig. 3.1). In terms of specifying the model parameters, two types of approach are proposed. First, similarly to most previous studies of data assimilation [*Moradkhani et al.*, 2005c; *Vrugt et al.*, 2005; *Wang et al.*, 2009; *Gharamti et al.*, 2013; *Xie and Zhang*, 2013; *DeChant and Moradkhani*, 2014; *Davison et al.*, 2017], the ensemble of parameters over the periods ($0 \leq t \leq t_c$) is assumed to follow a prior distribution. That is, the n number of parameter sets are sampled from uniform distributions with predefined bounded ranges (i.e., from the results in the light blue shaded region in Fig. 3.1). The values of parameters remain unchanged, while those of state vectors are continuously updated until the beginning of the forecasting period (i.e., $t = t_c$). This is hereafter named "*Random*" set — referring to the use of random parameter sets for running ***Model*** of NAM, PCE-I, and PCE-II.

An alternative way to this *Random* specification method is enabled by taking the advantage of the ability to calibrate model parameters with observed data before the forecasting period. Specifically, this method uses the posterior results of GLUE behavioral runs (referring to the light red shaded region in Fig. 3.1), i.e., selected parameter sets for running ***Model*** — called "*Selected*" specification method. The selected parameter sets for ***Model*** remain unchanged over the warm-up and calibration periods as well. As with the former approach, the values of state vectors are initially set to be zero at $t = 0$ but are continuously updated until $t = t_c$. We expected to see the EnKF process converge much faster and the forecasting results improve.

3.3.4 Modeling approaches for forecasting

In total, 18 modeling approaches (see Fig. 3.2) were developed by combining the modeling options with various techniques (NAM + PCE + GLUE + EnKF) in Sections 3.3.2 and 3.3.3. The modeling techniques were coupled to successfully perform ensemble flood forecasting and to meet the need for accurate and efficient flood forecasting. The 18 approaches represent permutations of the 3 × 2 × 3 subcases (Table 3.2). First, they were divided into three subcases corresponding to $\boldsymbol{Model}$, depending on whether a deterministic model or a PCE model was used over the calibration period (see Section 3.3.2) and how the latter was developed. Second, these modeling sets were divided into two subcases corresponding to *Random* or *Selected* sets, depending on how the parameter sets before the forecasting period were specified (see Section 3.3.3). Lastly, they were divided into three subcases depending on the methodology of data assimilation. The first of the three subcases did not use any data assimilation, and the other two used single- and dual-ensemble Kalman filters (see Section 3.2.3). We evaluated the modeling performance of the coupling framework by assessing accuracy, efficiency, and predictability in Section 3.5.2. The performance comparisons of the 18 approaches are expected to be a guide to which approach demonstrates better skill and most appropriate and which should be avoided.

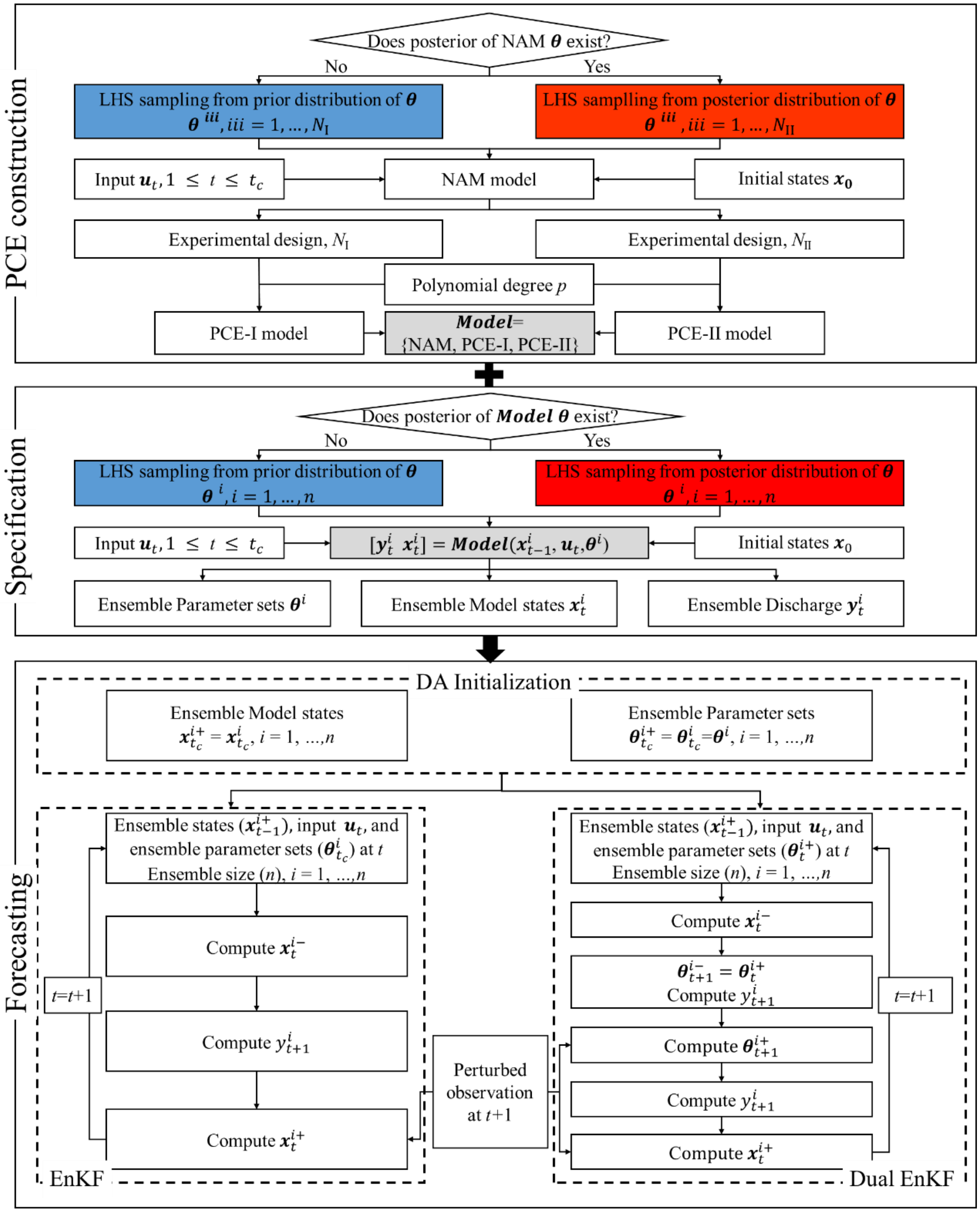

PCE construction
Does posterior of NAM $\boldsymbol{\theta}$ exist?
No
Yes
LHS sampling from prior distribution of $\boldsymbol{\theta}$
$\boldsymbol{\theta}^{iii}, iii = 1, \ldots, N_{\mathrm{I}}$
LHS samplling from posterior distribution of $\boldsymbol{\theta}$
$\boldsymbol{\theta}^{iii}, iii = 1, \ldots, N_{\mathrm{II}}$
Input $\boldsymbol{u}_t, 1 \leq t \leq t_c$
NAM model
Initial states $\boldsymbol{x}_0$
Experimental design, N_{I}
Experimental design, N_{II}
Polynomial degree p
PCE-I model
$\boldsymbol{Model}$= {NAM, PCE-I, PCE-II}
PCE-II model
Specification
Does posterior of $\boldsymbol{Model}$ $\boldsymbol{\theta}$ exist?
No
Yes
LHS sampling from prior distribution of $\boldsymbol{\theta}$
$\boldsymbol{\theta}^{i}, i = 1, \ldots, n$
LHS sampling from posterior distribution of $\boldsymbol{\theta}$
$\boldsymbol{\theta}^{i}, i = 1, \ldots, n$
Input $\boldsymbol{u}_t, 1 \leq t \leq t_c$
$[\boldsymbol{y}_t^i\ \boldsymbol{x}_t^i] = \boldsymbol{Model}(\boldsymbol{x}_{t-1}^i, \boldsymbol{u}_t, \boldsymbol{\theta}^i)$
Initial states $\boldsymbol{x}_0$
Ensemble Parameter sets $\boldsymbol{\theta}^i$
Ensemble Model states $\boldsymbol{x}_t^i$
Ensemble Discharge $\boldsymbol{y}_t^i$
Forecasting
DA Initialization
Ensemble Model states
$\boldsymbol{x}_{t_c}^{i+} = \boldsymbol{x}_{t_c}^i, i = 1, \ldots, n$
Ensemble Parameter sets
$\boldsymbol{\theta}_{t_c}^{i+} = \boldsymbol{\theta}_{t_c}^i = \boldsymbol{\theta}^i, i = 1, \ldots, n$
Ensemble states $(\boldsymbol{x}_{t-1}^{i+})$, input $\boldsymbol{u}_t$, and ensemble parameter sets $(\boldsymbol{\theta}_{t_c}^i)$ at t
Ensemble size (n), $i = 1, \ldots, n$
Compute $\boldsymbol{x}_t^{i-}$
Compute y_{t+1}^i
Compute $\boldsymbol{x}_t^{i+}$
t=t+1
EnKF
Perturbed observation at t+1
Ensemble states $(\boldsymbol{x}_{t-1}^{i+})$, input $\boldsymbol{u}_t$, and ensemble parameter sets $(\boldsymbol{\theta}_t^{i+})$ at t
Ensemble size (n), $i = 1, \ldots, n$
Compute $\boldsymbol{x}_t^{i-}$
$\boldsymbol{\theta}_{t+1}^{i-} = \boldsymbol{\theta}_t^{i+}$
Compute y_{t+1}^i
Compute $\boldsymbol{\theta}_{t+1}^{i+}$
Compute y_{t+1}^i
Compute $\boldsymbol{x}_t^{i+}$
t=t+1
Dual EnKF

Figure 3.2. The overview of an ensemble flood forecasting framework. The top box, "PCE construction" is for the process of building 2 PCE models (blue box for PCE-I and red box for PCE-II). The middle box, "Specification" describes 2 distinct approaches of specifying model parameters before forecasting including *Random* (blue box) and *Selected* (red box). The bottom box, "Forecasting" corresponds to data assimilation for flood forecasting in real-time (single and dual EnKFs). The top and middle blue boxes correspond to sampling N_{I} and n, independently from the same prior uniform distributions, respectively, while the red boxes sampling N_{II} and n from the same posterior distributions.

Table 3.2. Forecasting approaches employed in this Chapter

Approach	**Specification**	***Model***	**Data assimilation**
A1	*Random*	NAM	None
A2			EnKF
A3			Dual EnKF
A4		PCE-I	None
A5			EnKF
A6			Dual EnKF
A7		PCE-II	None
A8			EnKF
A9			Dual EnKF
A10	*Selected*	NAM	None
A11			EnKF
A12			Dual EnKF
A13		PCE-I	None
A14			EnKF
A15			Dual EnKF
A16		PCE-II	None
A17			EnKF
A18			Dual EnKF

3.4 Experimental setups

3.4.1 Case study

In this work, the unified framework is applied to predict hourly streamflow in the Vu Gia watershed as shown in Fig. 3.3. The watershed is one of the largest in central Vietnam, with a total area of 1,679.8 km^2 in the tropical region. It experiences a typical continental monsoon climate, with concentrated rainfall mainly from September to December. As the Vu Gia watershed is characterized by a large difference in elevation (slopes of approximately 30 %), floods occur rapidly and frequently. The region has experienced intense severe flooding and significant damage [*UNDP*, 1999; *Nga et al.*, 2015].

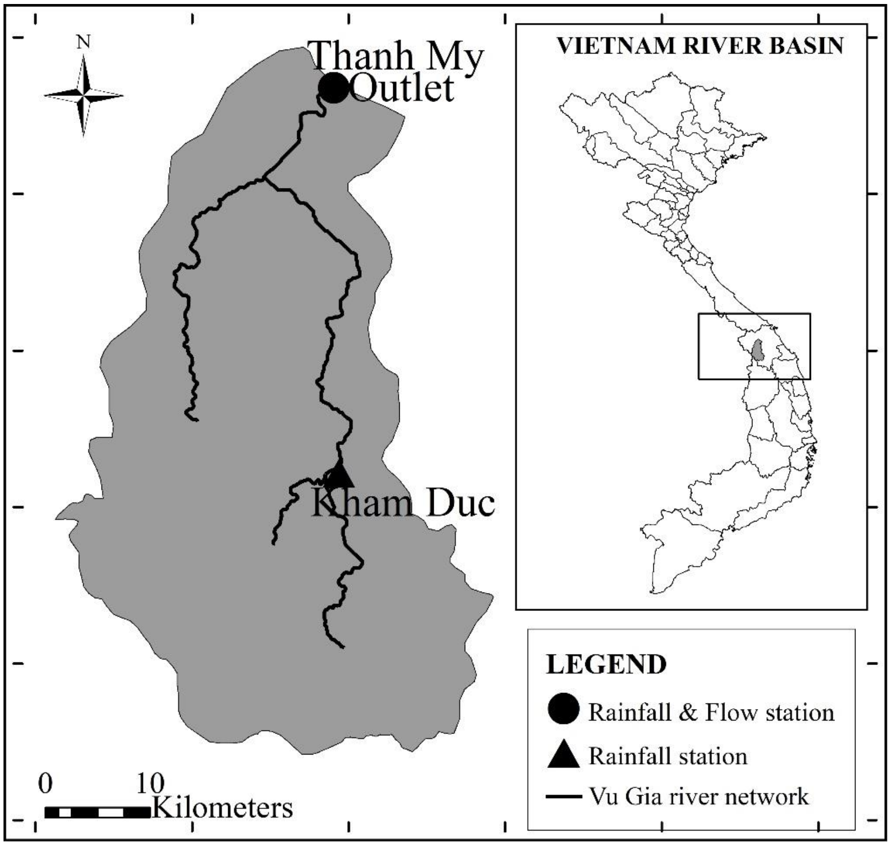


Figure 1.3. Study area: Vu Gia watershed

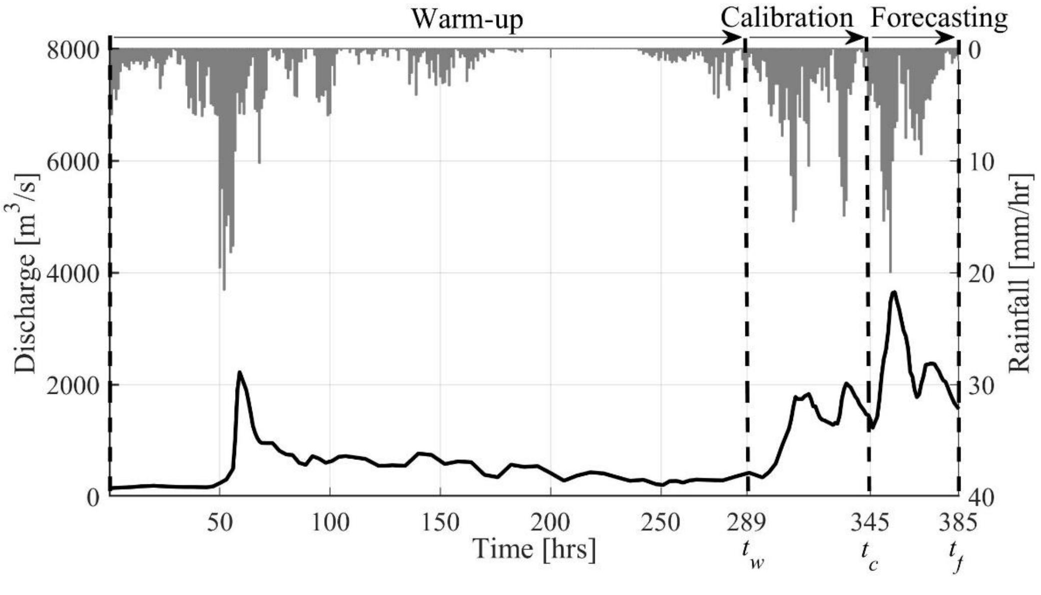


Figure 3.4. Flood event used in the study. The black line is the discharge of outlet and the gray hyetograph represents the average rainfall of Vu Gia watershed.

Streamflow data used for the outlet of the basin was collected hourly at Thanh My station – the only hydrometric station in the domain. Rainfall data was also observed hourly and obtained from two weather stations near the study area (Thanh My and Kham Duc stations). The average rainfall over the basin (Fig. 3.4) was calculated through the Thiessen polygon method. Observations from Dec. 1 to 17, 2016, are employed, in which the data from Dec. 1 to 13 was used for the warm-up period (i.e., from 0 to t_w), the data from Dec. 13 to 15 for the calibration period (i.e., from t_w to t_c), and the remaining data (assuming numerically that this data was newly provided at an hourly basis) corresponds to the forecasting period (i.e., from t_c to t_f) (Fig. 3.4). Note that rainfall forecasts has not been considered in this experimental design, what is done is hindcasting but one refers to the period between t_c and t_f as the "forecasting period", allowing for replicating real-life operational flood-forecasting process. Also note that a source of uncertainty for rainfall forecasts has not been presented, but it could have been addressed in Eq. 3.1 that has the flexibility to include *ensemble* precipitation inputs ($\boldsymbol{u}_t$).

To investigate the effects of different warm-up lengths, we initialize NAM a number of days before the calibration period. The warm-up periods chosen are 12 days (289 hours), 20, 30, and 60 days. We then compared the posterior probability densities estimated through GLUE for 9 model parameters and 5 model states at the beginning of the calibration period (at t = 290 hr). Fig. 3.5 shows that longer warm-up lengths do not change the posterior distributions of model parameters and states significantly, although have some effects on the parameter of Lm and the variable of L related to subsurface flow. However, we can confirm that the most sensitive two parameters of CQOF and CK12 on the flood hydrograph reach their equilibrium. Such an equilibrium can be obviously found in the hydrographs and the metrics (Fig. 3.6) over the calibration period. From above, we contend selecting 12 (289 hours) days as the length of the warm-up period is acceptable.

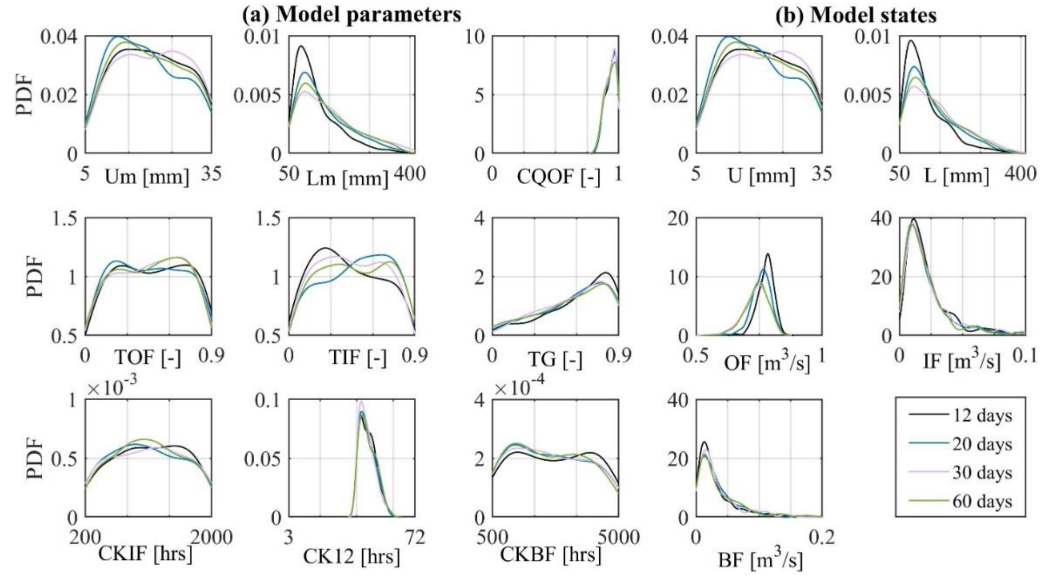


Figure 3.5. The posterior probability distributions of (left) 9 model parameters and (right) 5 model states at the beginning of the calibration period (at *t* = 290) for different warm-up lengths.

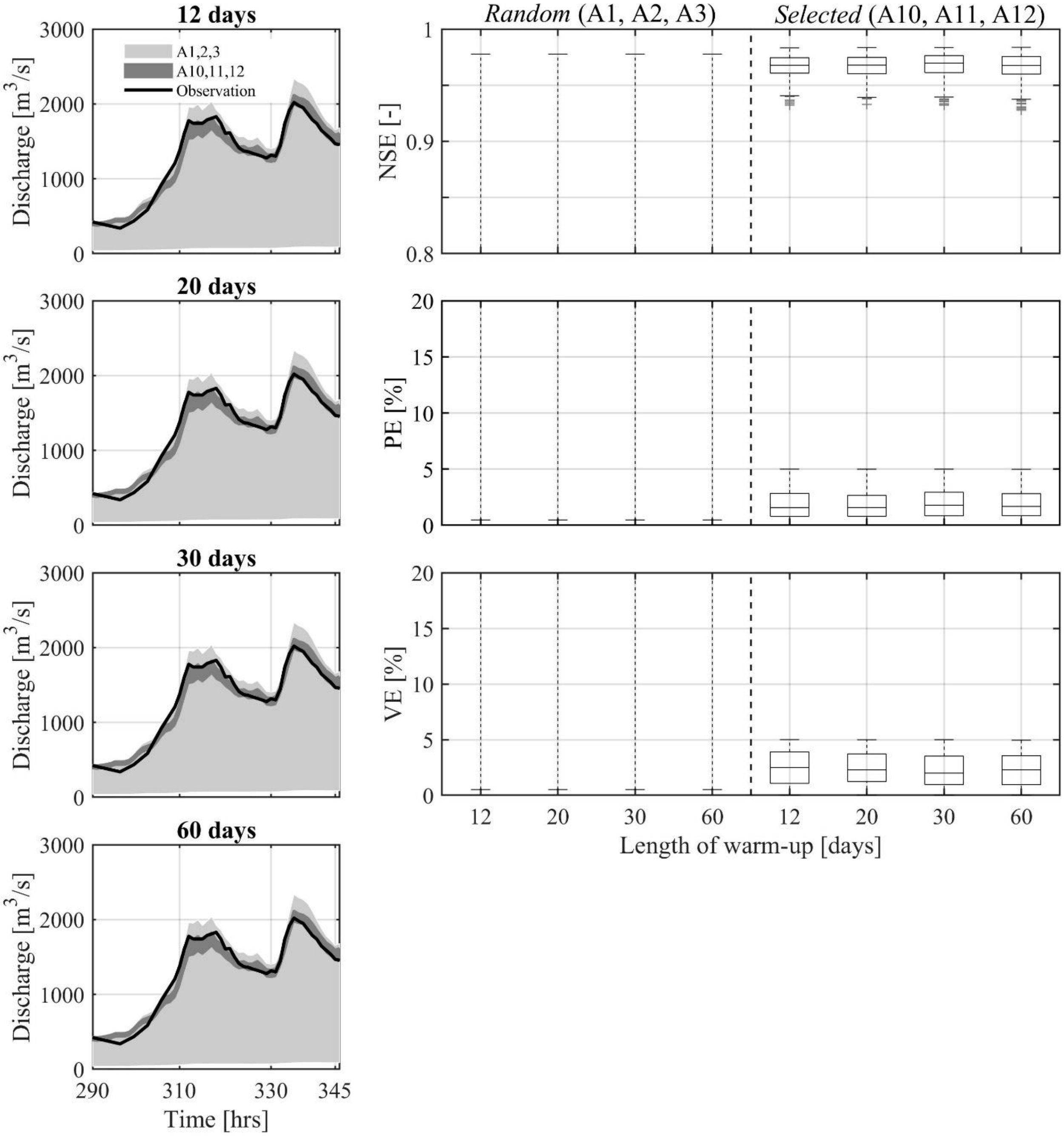


Figure 3.6. The effects of different warm-up lengths on (left) hydrographs and (right) NSE, PE, and VE metrics for A1, A2, A3 and A10, A11, A12 over the calibration period.

3.4.2 Data assimilation setups

The EnKF allows for the perturbation of observations to generate replicates of $\boldsymbol{x}_{t-1}$ and $\boldsymbol{\theta}_t$, and the correction of the ensemble forecast members through an update step [*Moradkhani et al.*, 2005c]. This prevents the EnKF from a collapse in which all ensemble forecast members are likely to have similar values [*Burgers et al.*, 1998]. As shown in Eq. 3.16, observations can be perturbed by adding stochastic noise to the observed value. This observed error in measurements is assumed to be independent and is set to be proportional to the observed values, following a Gaussian distribution with predetermined variance. In this work, to select an appropriate observational error, we attempt to conduct sensitivity analysis of the noises of observation and rainfall in the data assimilation. In order to evaluate the ensemble performance, the Normalized RMSE Ratio (NRR) discussed by *Anderson* [2001] and *Moradkhani et al.* [2005c] is used.

$$NRR = \frac{Ra}{E[Ra]} \tag{3.29}$$

$$Ra = \frac{R_1}{R_2} \tag{3.30}$$

$$R_1 = \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\sqrt{\left[\left(\frac{1}{n}\sum_{i=1}^{n} y_t^i\right) - y_t^{obs}\right]^2} \tag{3.31}$$

$$R_2 = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\left(y_t^i - y_t^{obs,i}\right)^2} \tag{3.32}$$

$$E[Ra] = \sqrt{\frac{(n+1)}{2n}} \tag{3.33}$$

where *Ra* is the ratio of the time-averaged RMSE of the ensemble mean (R_1) to the ensemble-averaged RMSE of the ensemble members (R_2); $E[Ra]$ is the expected value of the RMSE ratio Ra. An ideal ensemble generation should produce a NRR value close to unity value (i.e., 1).

$NRR > 1$ indicates that the ensemble has too little spread, while $NRR < 1$ is an indication of an ensemble with too much spread.

A sensitivity analysis is examined for the range of observation noise (1 - 30 %) and rainfall noise (1 - 30 %). Fig. 3.7 shows that NRR is more sensitive to the observation noise than the rainfall noise. It means that accurate prediction is highly dependent on the observation replication. As seen, the small values of observation noise (approximately 4 and 5 %) will keep the NRR within the acceptable range (0.99 - 1.01). Thus, 5 % was selected as the perturbation of streamflow observation in this study.

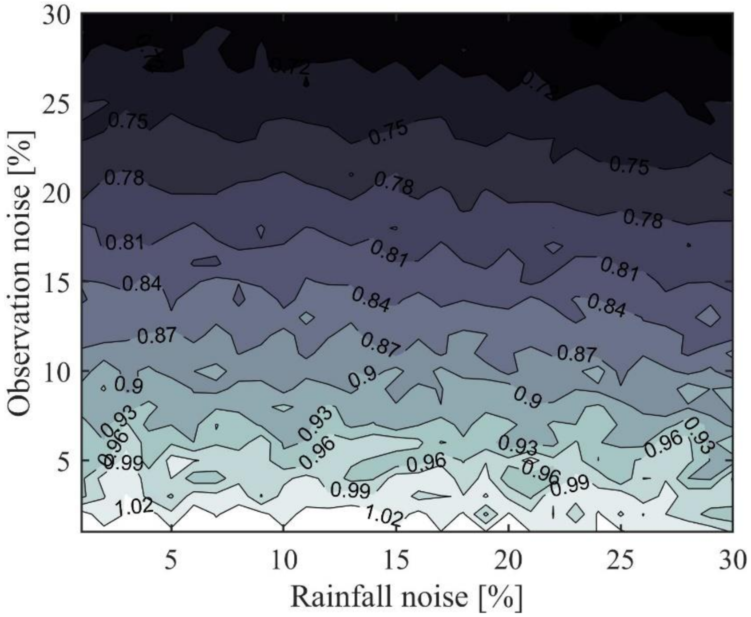


Figure 3.7. NRR space with respect to the varying noises of rainfall and observation. 500 ensemble members are used for the computation of NRR.

Furthermore, overshooting or filter divergence problem in data assimilation happens when the ensemble size is small or the initial values of ensemble members are quite different from the true. To resolve this issue, we used a sufficiently large ensemble size and the posterior information of parameters to initialize the ensemble of EnKF. Also, determining the size of ensemble for forecasting (n) is related to quantifying the uncertainty bounds and representing the EnKF. In previous studies, the ensemble size was selected randomly or large enough (at least 100 members) to fully identify the uncertainty confidence intervals [*Cameron et al.*, 2000; *Beven and Freer*, 2001; *Hossain and Anagnostou*, 2005; *Choi and Beven*, 2007; *Blasone et al.*, 2008b; *Jin et al.*, 2010; *Shen et al.*, 2012]. A sufficient number of ensemble parameter sets to achieve both goals of efficiency and uncertainty quantification should be determined. Following the results in Chapter II, we used an *n* of 500 as the optimal size of the ensemble.

3.4.3 Performance metrics

To assess the modeling performance of the 18 approaches, metrics representing accuracy, predictability, and efficiency were chosen, beginning with the accuracy metrics of Nash–Sutcliffe efficiency (NSE), absolute error (AE), and relative entropy (RE) [*Kullback and Leibler*, 1951; *Kullback*, 1997; *Kleeman*, 2002]. Second, Brier scores (BS) [*Brier*, 1950], and the range of uncertainty (UR) were used to assess the predictability of probabilistic forecasts. Lastly, a metric calculating total runtime (TRT) was evaluated to compare the computational efficiency of the tested approaches.

NSE, which is traditionally used to evaluate the accuracy power of deterministic models, is computed for each ensemble member (i) over the entire computation time (see Eq. (2.9))

Absolute error (AE) is differences between actual observations and predictions of each ensemble members at each time *t*. Thus, it varies with time and can be written as:

$$AE_t^i = \left|y_t^{obs} - y_t^i\right|,\ t = 1, \dots, \mathrm{T};\ i = 1, \dots, n \qquad (3.34)$$

Relative entropy (RE) is a measure of the statistical difference between probability distributions over the entire forecasting period of observations and model simulations [*Kleeman*, 2002; *Shukla et al.*, 2006; *Giannakis and Majda*, 2012]. Following *Kleeman* [2002] and *Heo et al.* [2014], RE can be defined as:

$$RE^{i} = \left[\log\frac{\sigma^{2}_{y^{obs}}}{\sigma^{2}_{y^{i}}} + \frac{\sigma^{2}_{y^{i}}}{\sigma^{2}_{y^{obs}}} - 1\right] + \left[\frac{(\mu_{y^{i}} - \mu_{y^{obs}})^{2}}{\sigma^{2}_{y^{obs}}}\right],\ i = 1, \dots, n \tag{3.35}$$

where $\mu_{y^{obs}}$ and $\mu_{y^{i}}$ are the mean, while $\sigma_{y^{obs}}$ and $\sigma_{y^{i}}$ are the variance of streamflow observation and the *i*-th model prediction over the entire computation time from t_c to t_f. Small values of relative entropy indicate that distribution of a given model is close to that of the observation. This is also called Kullback-Leibler divergence between the two distributions, model and data, assuming Gaussianity of both.

The Brier score (BS) is one of the most commonly used verification measures for assessing the predictability of probabilistic forecasts. The score is defined as the mean squared error of the probabilistic forecasts over the verification sample, expressed as:

$$BS = \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\left(p_t^{f} - o_t\right)^{2} \tag{3.36}$$

where p_t^{f} is the forecast probability for the *t*-th time, which refers to the ratio among ensemble reaching a predefined flow threshold; o_t is the observed probability, which is 1 if observation at *t*-th time, y_t^{obs} is larger than the threshold, and 0 if it is not. In this study, this threshold value was chosen as the proportional rate of 90% of the true discharge peak.

The uncertainty range (UR) is the range between the 5th and 95th percentiles of the ensemble outcomes (y). It is computed over each computational time t in hydrographs, expressed as:

$$UR_t = y_t^{95} - y_t^{5},\ t = \ 1, \dots, \mathrm{T} \tag{3.37}$$

Lastly, the total run time (TRT) for all of the approaches is defined as:

$$TRT = \left(RT_{w+c,\boldsymbol{Model}} \times fac_{\boldsymbol{Model}} + RT_{f,\boldsymbol{Model},DA}\right) \times n + RT_{build,\boldsymbol{Model}} \tag{3.38}$$

where $RT_{w+c,\boldsymbol{Model}}$ is the run time to compute one simulation of ***Model*** (NAM, PCE-I, and PCE-II) over the warm-up and calibration periods, i.e., from 0 to t_c; $RT_{f,\boldsymbol{Model},DA}$ is the run time to compute one simulation of ***Model*** with different DA methods over the forecasting period, i.e., from t_c to t_f; and $RT_{build,\boldsymbol{Model}}$ is the run time needed for building ***Model***. For example, because it is unnecessary for constructing the deterministic model, the time for NAM is zero. The building run times for PCE-I and PCE-II will be calculated in detail in Sec. 3.5.1.2. The factor $fac_{\boldsymbol{Model}}$ represents the number of ***Model*** runs to obtain a single behavior run in GLUE, and remains 1 in A1 to A9, while it depends on ***Model*** for the rest of approaches.

Eq. 3.38 is a linear function with respect to the number of ensembles run, in which $RT_{w+c,\boldsymbol{Model}} \times fac_{\boldsymbol{Model}} + RT_{f,\boldsymbol{Model},DA}$ serves as the slope of the linear function and $RT_{build,\boldsymbol{Model}}$ the intercept. The values of the slope and intercept and the executed times of the 18 approaches are addressed in Section 3.5.2.

3.5 Results

3.5.1 Preparation steps before forecasting

3.5.1.1 Attaining parameter posterior distributions

The posterior distributions of parameters can be generally attained by using Bayesian inference. As detailed in Section 3.2.2, we employed a relatively simple and robust method, GLUE [*Beven and Binley*, 1992], that does not require reformulation of the deterministic model code. Details on why we choose the likelihood functions of NSE, PE, and VE, how we determine the cutoff threshold values of each function, and which parameters are more sensitive, are described in Chapter II. We confirmed the benefits of a warm-up technique that significantly speeds up the GLUE process of finding the behavioral sets: without warm-up, no behavioral set was obtained from GLUE even after a sufficiently large number of NAM model runs, while with warm-up, a behavioral set was obtained after approximately 118.0 model runs for NAM (A10 to A12), 26.9 for PCE-I (A13 to A15), and 3.6 for PCE-II (A16 to A18), respectively. Therefore, the factors, $fac_{\boldsymbol{Model}}$ are 118.0, 26.9, and 3.6 for NAM, PCE-I, and PCE-II, respectively in A10 to A18.

3.5.1.2 Constructing the PCE models

Determining the coefficients of the PCE-I and PCE-II models depends on the number of the experimental design (N) and the polynomial degree (p) [*Blatman and Sudret*, 2010; *Blatman and Sudret*, 2011]. To discover appropriate values for N and p, the effect of experimental design N on PCE performance was first evaluated. Specifically, a number of simulations were repeated with the N value varied between 10 and 5,000 but the value of p was set as 3, and the performance results of LOO for streamflow (y) and the five model states computed. Fig. 3.8 shows that the LOO values for streamflow and five state variables become smaller as the value of N increases, and ceases to become smaller when N approaches a certain value. For N values larger than this threshold, the model performance was almost indistinguishable (the left column plots in Fig. 3.8). From a visual inspection of Fig. 3.8, the optimal N value for constructing the PCE-I and PCE-II models would be 1,000 and 100, respectively.

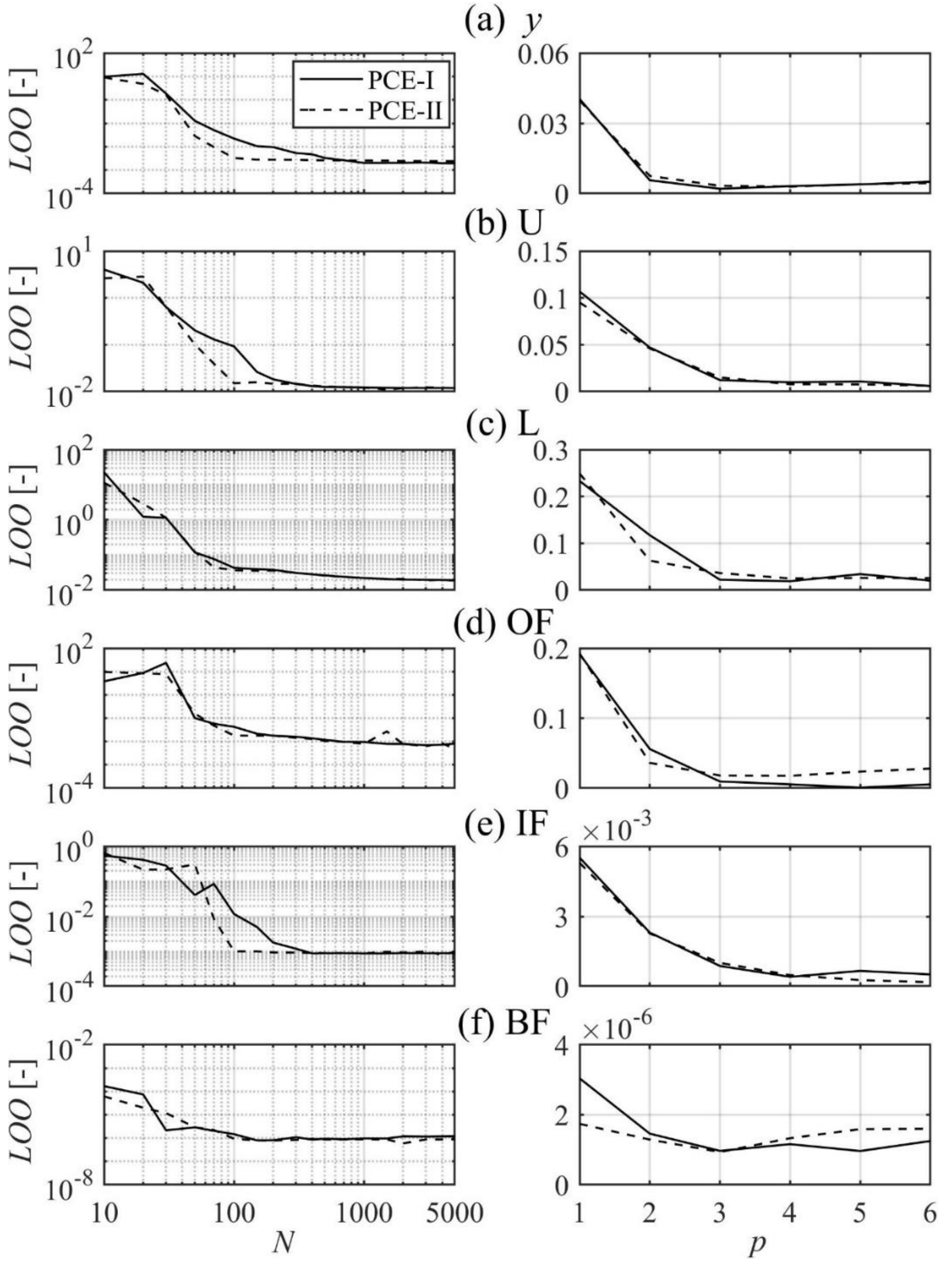


Figure 3.8. The effects of (left plots) the experimental design, N and (right plots) the polynomial degree, p on the leave-one-out cross-validation error (LOO) in constructing PCE-I and PCE-II models, for (a) streamflow and (b to f) 5 model states.

A selection of the polynomial degree p was made in a fashion similar to the aforementioned procedure. The value of p was varied from 1 to 6 and N was set as 1,000 (PCE-I) and 100 (PCE-II). From the results of Fig. 3.8 (the right column), the gradients of the LOO metrics assessed changed considerably when p was set to 3 and the values remained stable for large magnitudes of p. In terms of reducing the computational time to construct a PCE model, a low polynomial degree would be preferred. Thus, a p of 3 would be an appropriate value to use when building both PCE models. With optimal values of N of 1,000 and 100, and a p of 3, PCE-I and PCE-II models can be built to quantify the uncertainty range for flow prediction and to compare the degree of accuracy and efficiency with the results of the deterministic NAM.

The total time to establish both PCE models is described. Obviously, the larger the number of the experimental design set, the more time is needed for computing N ensemble runs. The time required to perform the N_{I} and N_{II} ensemble runs of NAM was 121.9 and 12.6 seconds for PCE-I and PCE-II, respectively. It also takes much more time to estimate PCE-I coefficients if one uses an ensemble set (N_{I}) generated from the prior distribution of the parameters than to compute PCE-II coefficients from parameter sets informed by the likelihood function. The time required to estimate PCE coefficients was 419.3 and 11.3 seconds, respectively. The summation of these two times was considered to be the total time required to build the PCE models before forecasting: approximately 541.2 and 23.9 seconds for PCE-I and PCE-II, respectively. The construction time of PCE-II is much (~22 times) faster than that of PCE-I.

3.5.1.3 Comparing the ensemble results of NAM and PCE models

Over the calibration period, ensemble results composed of 500 *Random* and *Selected* runs were compared for three different models. To make the 500 *Selected* behavioral results, 58,977, 13,444, and 1,822 (n_w) random runs were required for NAM, PCE-I, and PCE-II, respectively. Compared with the NAM itself, using PCE models can reduce the amount of computational runs by a factor of about 4.4 for PCE-I and 32.4 times for PCE-II model. The composing behavioral set for PCE-II was even faster (~7.4 times) than for PCE-I.

Fig. 3.9 shows hydrographs for the 500 *Random* (A1 to A9) and *Selected* (A10 to A18) simulations for the three models. Their uncertainties are illustrated with a 90% confidence interval, which corresponds to 5 and 95% quantiles of the 500 ensemble members. Because we controlled the conditions for the behavioral set of GLUE, the overall comparison with the observed values for the results of the *Selected* cases (A10 to A18) is very satisfactory. Specifically, the NSE value was always higher than 0.9 and both PE and VE values were less than 5% for all cases. However, streamflow curves for the *Random* simulations (A1 to A9) clearly show different patterns depending on the model. It can be anticipated that the results of these *Random* cases will not be encouraging and their uncertainties will be large. However, the results of some cases using PCE-II model were very satisfactory and their uncertainties small.

As mentioned above, when making using observations to constrain the parameter sets (A10 to A18), the results of both PCE models are similar to those of the NAM and no substantial differences were observed. This confirmed that both PCE models have an equivalent degree of accuracy as the NAM and can provide an excellent match to the deterministic model. In terms of efficiency, it is also advantageous to use the PCE model (discussed in Sections 3.5.2.1 and 3.6.1), and there is no reason to hesitate adopting the PCE model for streamflow prediction.

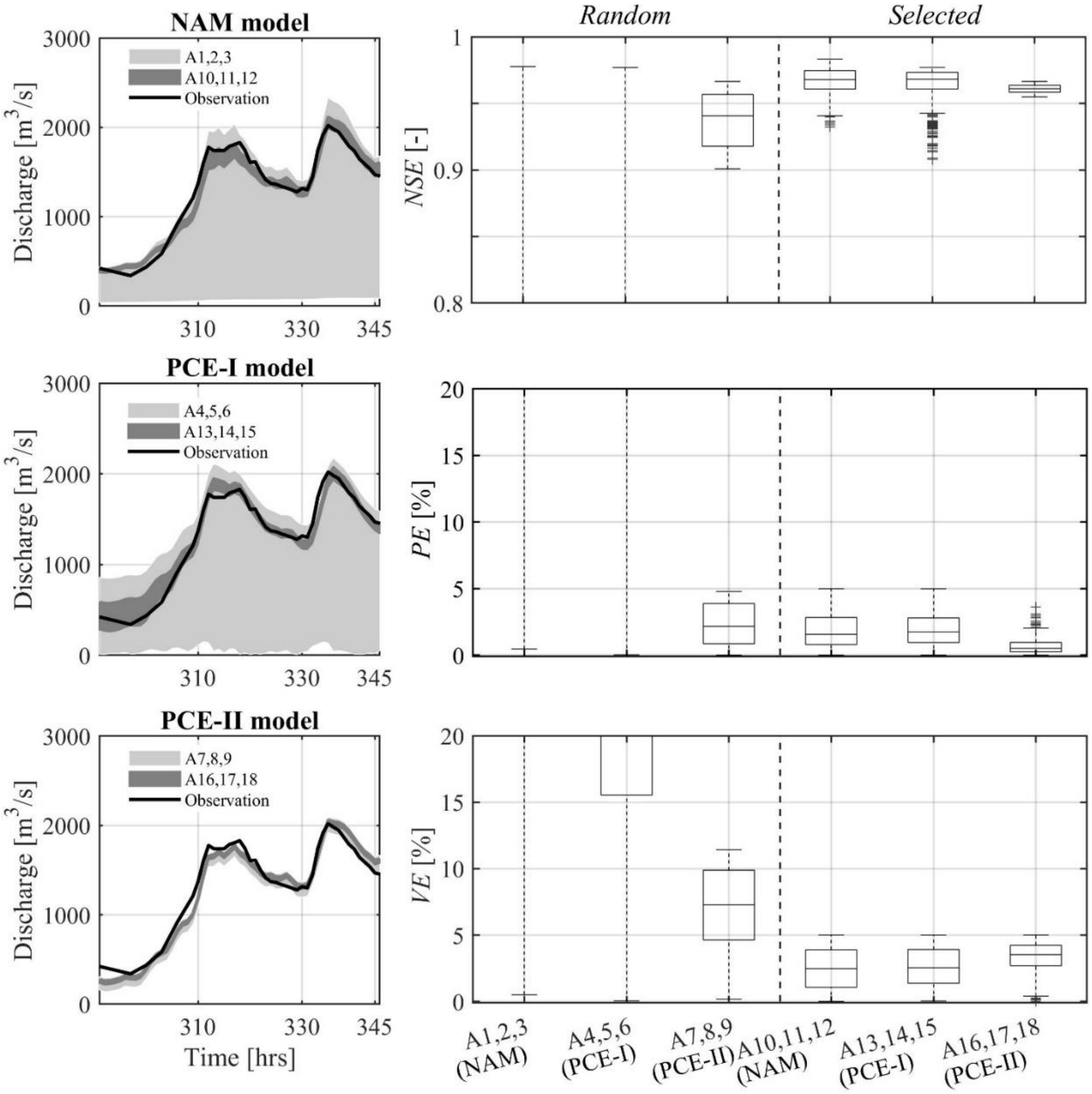


Figure 3.9. The left column plots show hydrographs for the calibration period. The shades in the plots correspond to 90% confidence interval for 500 *Random* model runs (A1 to A9, light gray shade) and 500 *Selected* model runs (A10 to 18, dark gray shade) for 18 approaches in Table 3.2. The boxplots in the right column demonstrate the verification metrics of NSE, PE, and VE for the 18 approaches used.

3.5.2 Flood forecasting with 18 approaches

3.5.2.1 PCE-I versus PCE-II model for real-time flood forecasting

Depending on the model used in forward simulations (NAM, PCE-I, and PCE-II), the results for the 18 approaches were divided into three groups. Almost all of the results of the six approaches using the PCE-II model were worse than those obtained with both NAM and PCE-I (Figs. 3.10 and 3.11). The only exception is for the A1 and A4, which did not have assimilation and whose parameter sets used were based on prior uniform distributions. No verification metrics computed using the results of forecasting based on PCE-II were satisfactory, except for the metric of $\overline{UR}$. However, if the accuracy is not ensured, the better performance in terms of $\overline{UR}$ is not meaningful. Specifically, NSE values were low, approximately 0.7; AE values at flood peak time (AE_{peak}) were larger than 750 m^3/s; RE was approximately 0.01; and BS was equal to 1 (Fig. 3.12). No metric improvements was found for the approaches based on PCE-II, even if combinations of assimilation and calibration techniques were applied. We concluded that the PCE-II model can reproduce streamflow characteristics well for the past period, but not for the future.

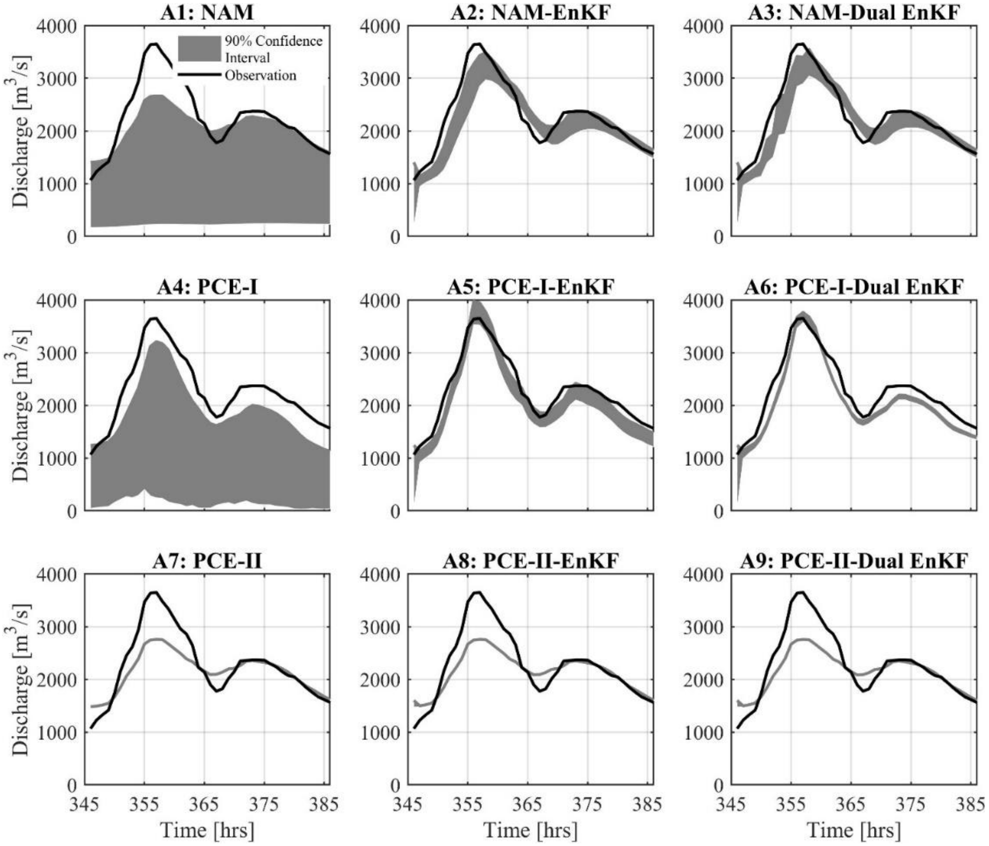


Figure 3.10. Hydrographs over the forecasting period, with a 90 % confidence interval of 500 *Random* model runs (A1 to A9).

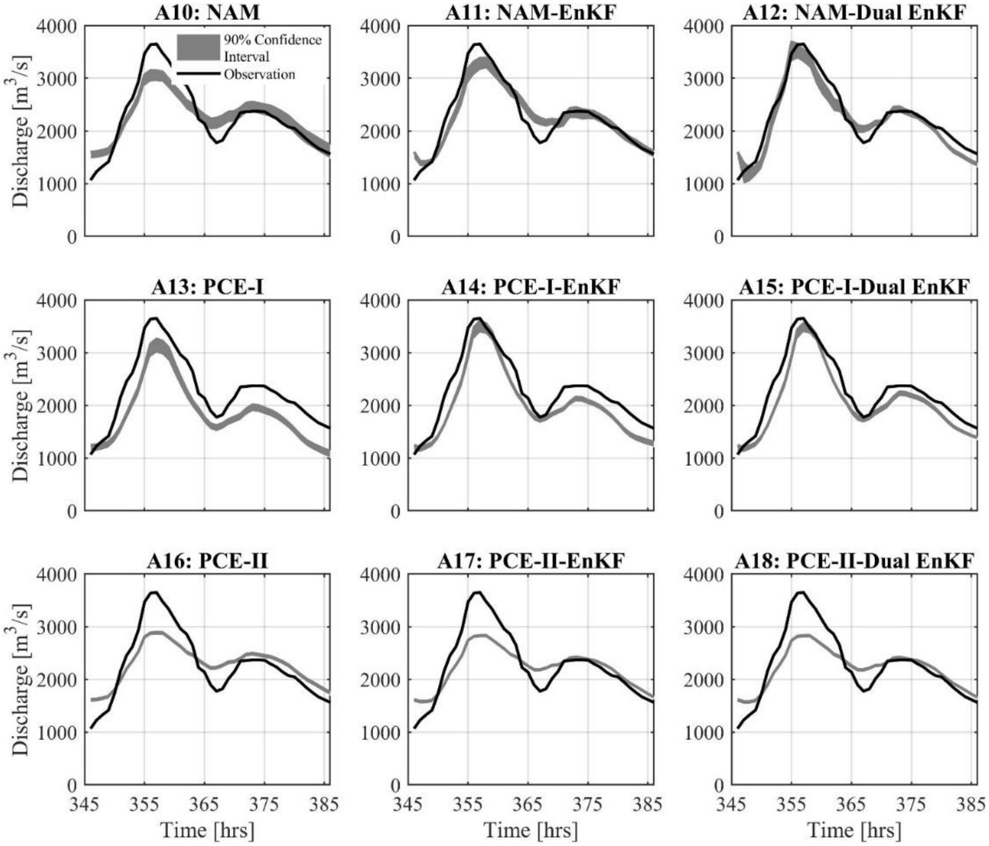


Figure 3.11. Hydrographs over the forecasting period, with a 90 % confidence interval of 500 *Selected* model runs (A10 to A18).

Conversely, the forecasting results of the approaches based on the PCE-I model are almost similar to those obtained with NAM, and in some cases even better. The latter can be seen in Fig. 3.12; the verification metrics of NSE, AE_{peak}, RE, and $\overline{UR}$ show better performance for PCE-I than for NAM results (e.g., A5 vs. A2, A6 vs. A3, A14 vs. A11, and A15 vs. A12) (see Table 3.3). In particular, the RE results in Fig. 3.12c illustrate that the PCE-I results are closer to the observed values than those obtained with NAM (A15 is the best result with the smallest value of RE). BS corresponding to PCE-I also has smaller values, close to zero, which indicates instances of when predictability of probabilistic forecasts matched predictability of observation (Fig. 3.12d).

Therefore, the PCE-I model can be adapted to substitute the NAM in performing real-time flood forecasting, as well as in capturing the uncertainty of calibration period.

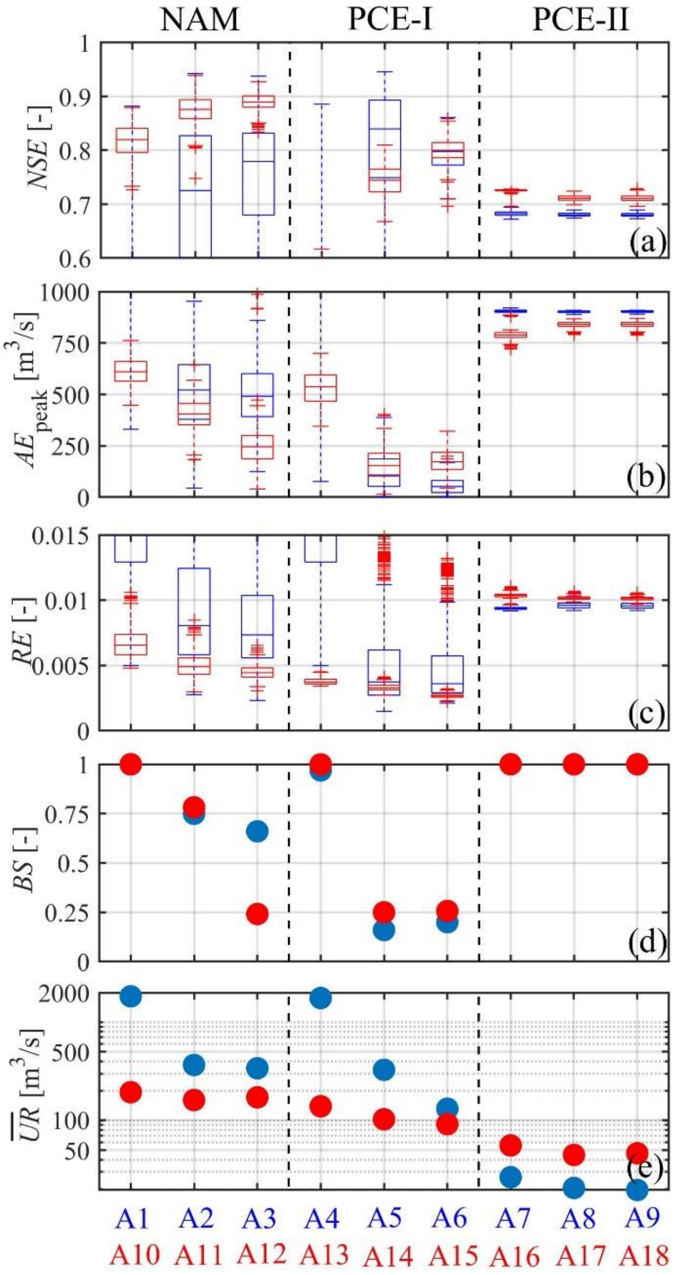


Figure 3.12. The performance metrics reflecting accuracy and predictability of the 18 approaches for the forecasting period. Boxplots of (a) NSE, (b) AE_{peak} (AE at flood peak time) and (c) RE show 500 ensemble values with the statistics of median (central mark), the 25th and 75th percentiles (edges of the box), and maximum and minimum except for outliers (whiskers). (e) $\overline{UR}$ is the mean of uncertainty range, UR_t over the entire forecasting period.

Table 3.3. Performance metric values of 18 approaches. The values in the first 3 columns are the medians of NSE, AE_{peak}, and RE, which match the values in Fig. 3.12.

Approach	**median of NSE [-]**	**median of AE_{peak} [m^3/s]**	**median of RE [-]**	**BS [-]**	**$\overline{UR}$ [m^3/s]**
A1	-3.62	2292.19	0.019	1.00	1834.18
A2	0.70	515.64	0.009	0.75	367.34
A3	0.75	500.45	0.009	0.66	340.12
A4	-3.84	1928.86	0.015	0.97	1760.11
A5	0.82	126.48	0.005	0.16	327.86
A6	0.79	55.84	0.005	0.20	131.92
A7	0.68	904.21	0.009	1.00	26.52
A8	0.68	901.91	0.010	1.00	20.59
A9	0.68	902.79	0.010	1.00	19.63
A10	0.82	612.60	0.007	1.00	193.95
A11	0.88	401.26	0.005	0.78	161.69
A12	0.89	242.62	0.005	0.24	172.18
A13	0.44	532.35	0.004	1.00	139.54
A14	0.74	157.07	0.003	0.25	102.98
A15	0.80	176.00	0.003	0.26	91.98
A16	0.73	787.18	0.010	1.00	55.75
A17	0.71	839.31	0.010	1.00	44.64
A18	0.71	840.19	0.010	1.00	46.37

Comparing the modeling results in terms of the computation speed, it is clear that simulating a surrogate model using the PCE theory is significantly faster than with a deterministic model such as NAM. The “slopes” of the runtime curves of Fig. 3.13 indicate both PCE approaches are approximately 20 times faster (A4 to A9) and ~80 times faster (A13 to A18) than the corresponding approaches using the NAM. Similarly, if we compare efficiency between PCE model approaches, using PCE-II may or may not offer much improvement in efficiency over PCE-I. There is only 10 % improvement when *Random* specification is applied (see the slope of A4, A5, A6 vs. A7, A8, A9 in Fig. 3.13), while there is about six times improvement when simulating *Selected* approaches (see the slope of A13, A14, A15 vs. A16, A17, A18). The use of surrogate models therefore did not sacrifice accuracy. The flood prediction accuracy of PCE-1 model presented here is similar to that of the original NAM, and computational efficiency has been found to be highly superior.

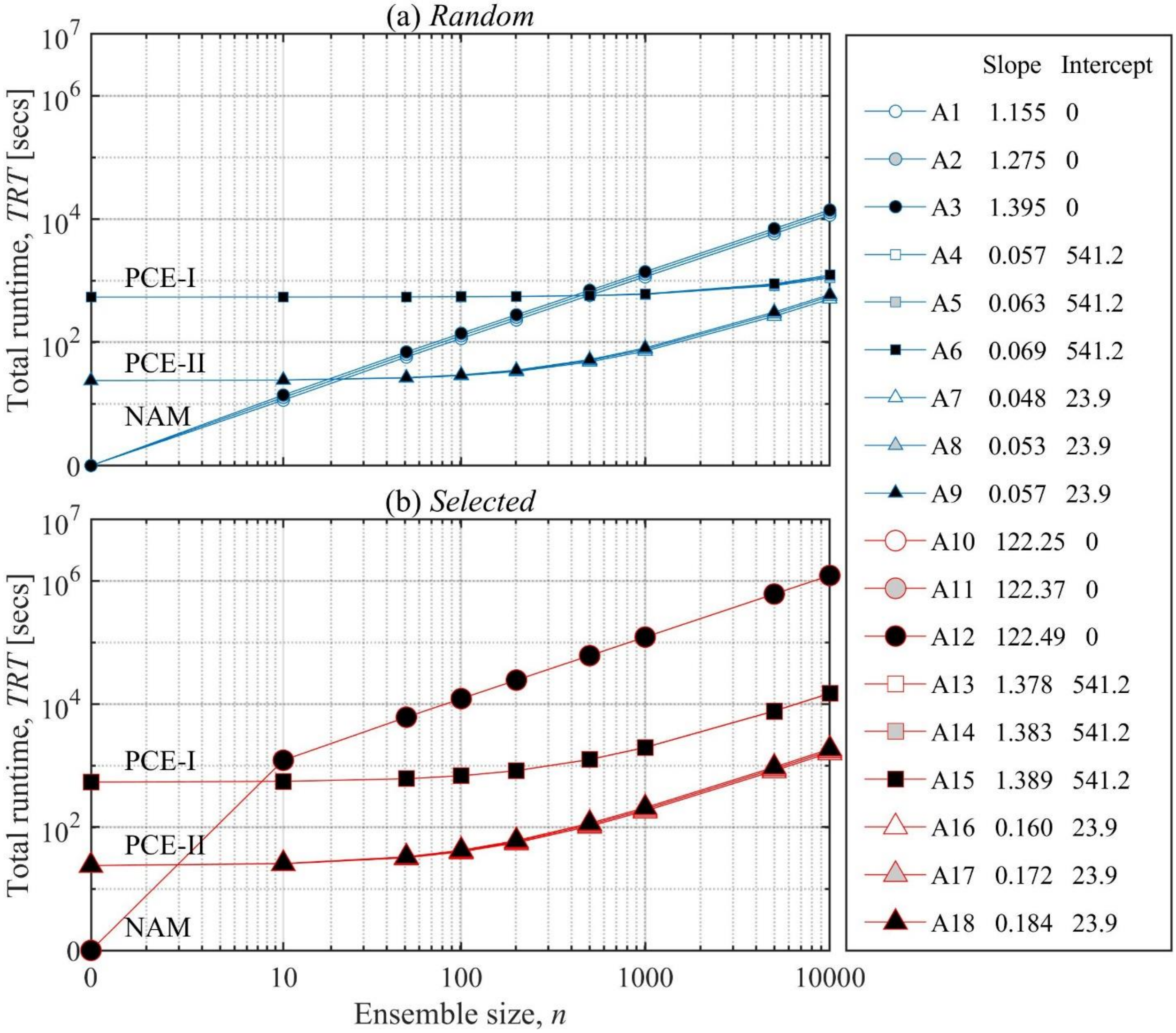


Figure 3.13. The total runtime (TRT) corresponding to 18 approaches in the forecasting period versus ensemble size (n). Note that although we plot on logarithmic axis, the actual total runtime has the form of a linear function with the ensemble size at linear scale; its slope and intercept values for all approaches are tabulated on the right.

3.5.2.2 Random versus Selected specification for forecasting

The approaches using the *Selected* specification generally show a better performance than those using the *Random* specification. This is especially noticeable in the NAM and PCE-I approaches, and rarely in PCE-II. First, in the approaches without data assimilation, their accuracy was significantly improved (compare A1 vs. A10 and A4 vs. A13). The performance of A10, represented by the NSE, AE_{peak}, RE, and $\overline{UR}$ metrics, was improved by about 95, 73, 61, and 89% compared with A1, and the performance of A13 about 86, 72, 79, and 92% over A5, respectively. Despite the noticeable improvement of A10 and A13, these results were still not ideal. The large AE error at the peak of A10 and A13 was approximately 450 m^3/s less than the observation, and the BS value was close to 1 (Fig. 3.12, Table 3.3). On the other hand, in the approaches in which data assimilation was used, the improvement effect for *Selected* specification was not greater than when it was not used. The increasing performance for the same metrics was about 55, 22, 36, and 56% (A2 vs. A11), and about 56, 52, 44, and 49% (A3 vs. A12). Here, the parameter specification effect was smaller because DA improves the absolute error magnitude.

Determination of states and parameters that can increase accuracy and predictability requires more computation time because a large number of model runs are carried out to make an inference for posterior distributions. For approaches using NAM (A1 vs. A10, A2 vs. A11, and A3 vs. A12), it took 56, 41, and 30 times longer; while for PCE-I (A4 vs. A13, A5 vs. A14, and A6 vs. A15), it took 13, 10, and 8 times, respectively (Fig. 3.13). Because of this computational burden, parameter inference can be a weakness for real-time flood forecasts where it is important to ensure sufficient time ahead. However, if the surrogate model is employed, the necessary repetition of estimating the posterior distribution can be performed quickly, and such a weakness can be overcome.

3.5.2.3 Single versus dual EnKF in real-time flood forecasting

Convincing evidence is presented that both single and dual EnKF can improve accuracy and predictability during real-time forecasting (with the exception of approaches using PCE-II). Both of these techniques perform well but the dual EnKF is the superior choice. As an example of the approaches using NAM, the three metrics of AE_{peak}, BS, and $\overline{UR}$ in the *Random* cases provided slightly better results: 515.64 vs. 500.45, 0.75 vs. 0.66, and 367.34 vs. 340.12, respectively (A2 vs. A3). But, in the *Selected* cases, there was a relatively large performance improvement for the two metrics of AE_{peak} and BS: 401.26 vs. 242.62 and 0.78 vs. 0.24 (A11 vs. A12). Similar trends were observed when using PCE-I, and the difference is remarkable, especially for the AE_{peak} metric (e.g., about 2.5 times for A5 vs. A6).

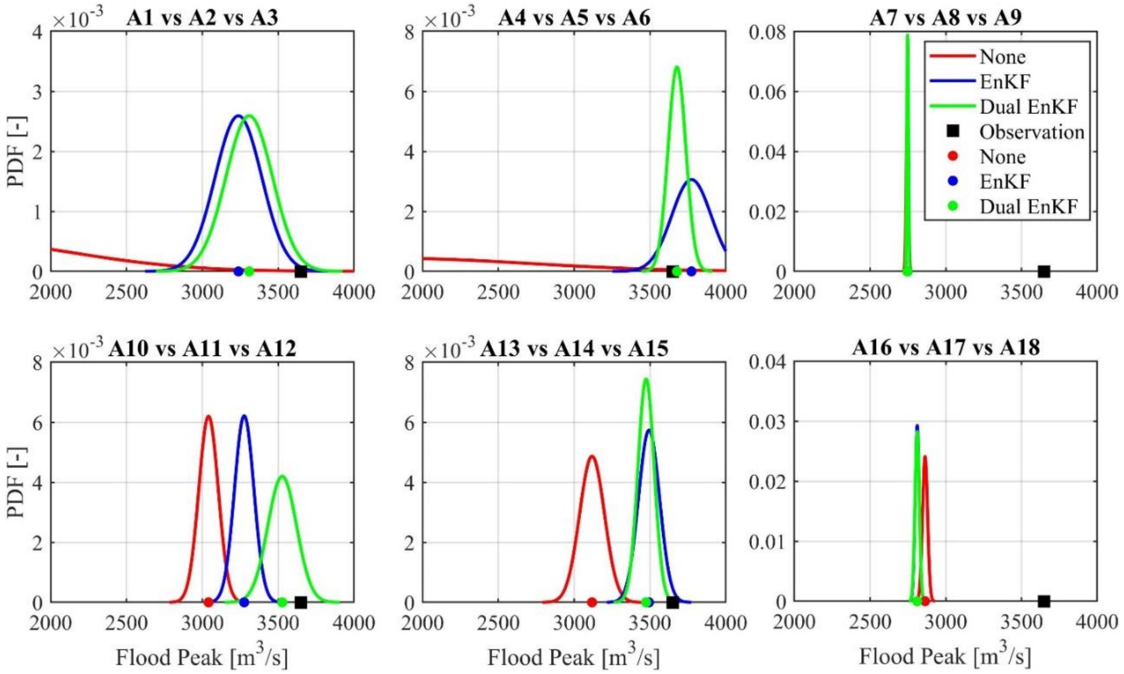


Figure 3.14. Comparisons of the three assimilation methods (none, EnKF, and Dual EnKF) for 500 ensemble flood peaks over the forecasting period. The left, middle, and right plots correspond to the approaches using NAM, PCE-I, and PCE-II, respectively. The first and second row plots correspond to the approaches using *Random* and *Selected* methods for parameter specification. The black square represents observed value at peak time; the circles are the expected values of the sample probability density functions.

From the overall inspection, it can be determined that the dual EnKF can adjust the peak of a hydrograph more accurately, and give a more confident result with a smaller uncertainty range. Therefore, we compared the distribution of flood peak values for 500 ensemble members in Fig. 3.14. This figure confirms that the joint update of states and parameters improves accuracy at flood peak more effectively than a single update of states. Also for the joint update, the expected value of the distribution was closer to the peak observation, and its variability is smaller (a narrower distribution).

Because the updating process is made twice, the dual EnKF is computationally more expensive. The computation time it takes to update states and parameters increased almost linearly. That is, the calculation time doubled or tripled for the cases of single and dual EnKF (using *Random* specification), respectively, as compared to the case without assimilations. However, for the approaches using the *Selected* specification, the calculation time did not seem to change significantly (Fig. 3.13), not because the time required for Kalman filtering was reduced, but because the time required for the parameter inference was so large that the filtering effect was masked.

3.6 Discussions

3.6.1 How can PCE be constructed for flood forecasting?

From the simulated flood forecasting results presented in Section 3.5.2, it is apparent that the manner of PCE construction has a significant impact on forecasting. The biggest difference in building PCE-I and PCE-II involves setting the range of the training sample (called experimental design). It is not surprising that a surrogate model trained for an event provides acceptable results only for the event trained. The flexibility to generalize to well-behaved outcomes for another event (e.g., a future event) is relatively low. This is why the calibrated model is often not appropriate for future forecasting. On the other hand, if a surrogate model can mimic the behavior of the original model to the greatest extent possible in a wide variety of situations and conditions, it will be able to capture its characteristics more comprehensively, thus playing a sufficient role in forecasting future events. Here we provide evidence the PCE-I model behaves like the NAM for the forecasting period, while the PCE-II behaves differently (despite both models behaving properly for the calibration period). To examine the robustness of both PCE model results, the Sobol' method (detailed in Section 2.2.1.4) was used to implement the variance-based measures of parameter sensitivities [*Sobol'*, 2001].

First, the PCE-I posterior histograms of the nine parameters obtained from GLUE for the calibration period are similar to those of the NAM, except for Lm and TG (Fig. 3.15). For these two parameters, a posterior histogram difference is a minor issue because the choice of the parameter values does not affect the end result, i.e., the sensitivity of the parameters is low. Other parameters of CQOF (1st) and CK12 (2nd) are the two most influential parameters to the model results, that is, their sensitivities are high. This result is consistent for both NAM and PCE-I (Fig. 3.16). The slight difference between the results of PCE-I and NAM, observed from the investigation of the sensitivity and the posterior distribution, is because we chose an appropriate number of training samples when constructing the PCE-I model. If one greatly increases the number of training sets, the difference in the above results will essentially disappear.

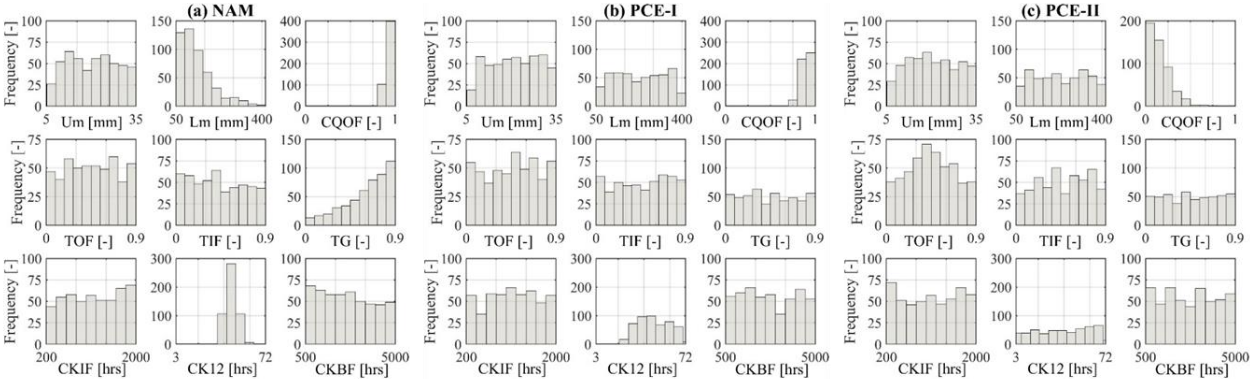


Figure 3.15. The posterior histograms for the 9 model parameters from 500 behavioral sets of 3 models (NAM, PCE-I and PCE-II) inferred by GLUE over the calibration period.

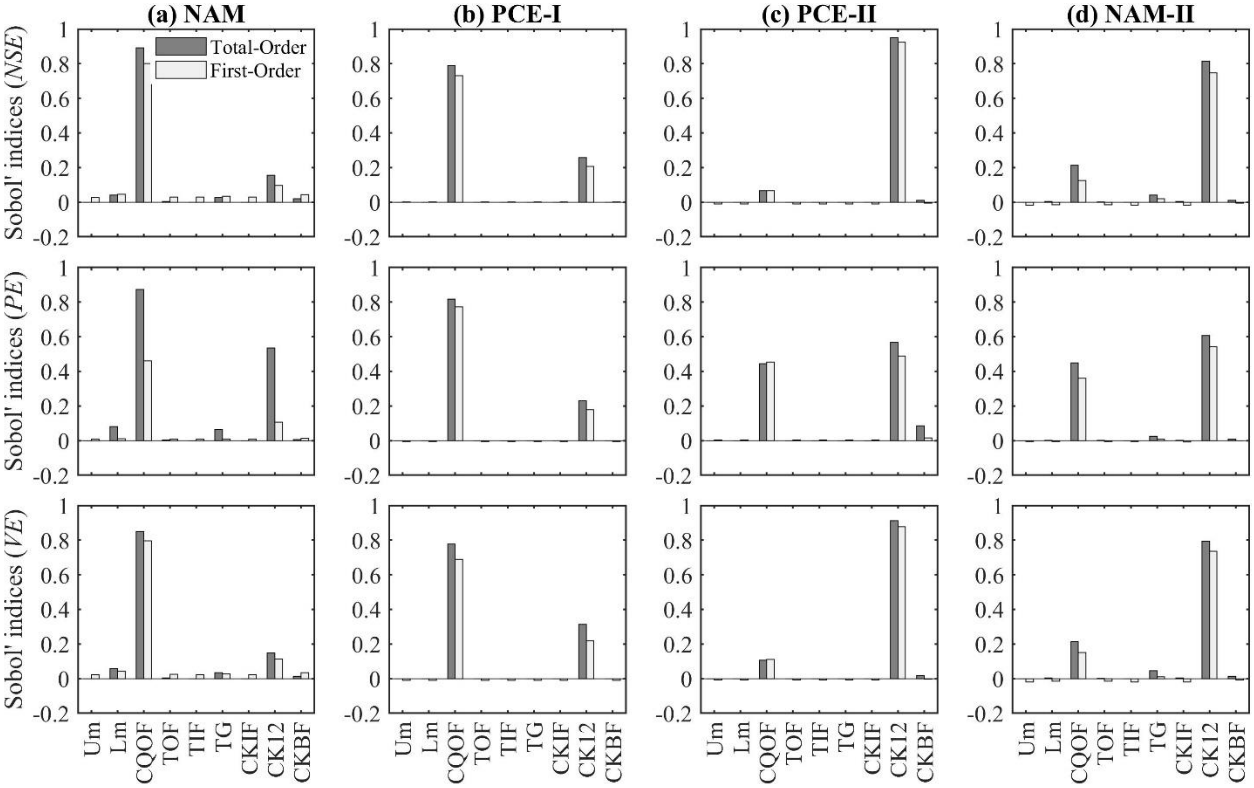


Figure 3.16. Sobol' sensitivity analysis for the 9 parameters, computed for the 3 likelihood functions of (top) NSE, (middle) PE, and (bottom) VE over the calibration period. The sensitivity results are attained based on (a, b, and c) the prior distributions of parameters for the 3 models of NAM, PCE-I, and PCE-II, respectively; and (d) the posterior distributions of parameters for NAM model. The posterior are also used to select the training parameter set for building PCE-II.

Second, the failure of PCE-II to mimic the NAM for the forecasting period can be explained largely due to the fact that PCE-II was trained using the only 100 behavioral parameter sets that were optimized for the calibration event. Model results will only vary within the boundaries that its trained data understand, and it will not be able to simulate the behavior of another event with a high skill, i.e., model "overfitting" occurs. However, over the calibration period, PCE-II always shows a good predictive performance for almost all parameter sets (compare the hydrographs of A1 to A3 with A7 to A9 in Fig. 3.9). In other words, no matter what parameter one chooses, satisfactory results are always achieved, which indicates that the influence of parameters is excluded. The posterior histograms of parameters for PCE-II (Fig. 3.15c) are almost uniform, except for the parameter of CQOF, which is the only one that can affect the end result, especially maintain the accuracy of the flood peak (note that the sensitivity of this parameter for *PE* is unusually high in Fig. 3.16c). If we change the threshold value of the likelihood function corresponding to the flood peak chosen to make the behavior set a slightly less constrained, this parameter will no longer play a role in constraining the result and follow a uniform distribution as well.

Another interesting aspect of the sensitivity test is that the sensitivity results of PCE-II differ from those of NAM and PCE-I, but are similar to those of NAM-II. The sensitivities of parameters have been altered in PCE-II. The NAM-II in Fig. 3.16d is hypothetically introduced to mimic the situations of PCE-II. Specifically, it refers to the sensitivity results when the NAM model was tested based on the posterior distributions (which are also used to select the training parameter set for building PCE-II), not the prior distributions of the parameters.

3.6.2 Is it feasible to construct a time-invariant PCE model?

A long-lasting challenge in hydrologic modeling is how to estimate parameters or state vectors optimized for all external and internal conditions. This would not be an issue for estimating previously described variables if the amount of data for calibration was sufficient. However, in the case of future forecasts during which no observation for calibration is available, it poses a problem. To tackle this challenging problem, *Fan et al.* [2016] and *Wang et al.* [2017] adopted a modeling framework in Eq. 3.1, so that PCE models should be reconstructed continuously at every time step. This method is flawless in theory, but requires additional computational resources (see efficiency comparisons in Appendix B). That is, the time to configure the PCE at every time step must be added to the total model simulation time, i.e., making the slope of Fig. 3.13 steeper. This disadvantage can be more pronounced when constructing surrogate models for complex, process-based deterministic models.

Unlike previous efforts, this study adopted an alternative modeling framework such as Eq. 3.2; that is, the PCE model is time invariant and thus developed only once over the calibration period. Therefore, during real-time forecasting, the total run time consists only of computational intervals needed for data assimilation of all ensemble members. This enhances computational efficiency significantly (see efficiency comparisons in Supplementary Material). This framework is not perfect, but the potential error that can occur by using the time-independent PCE model is minimized by coupling the data assimilation technique, thus complementing accuracy. From a comparison of the results of 18 approaches, we confirmed that the modeling framework needed for building a PCE model (especially PCE-I) is feasible. This embraces the notion that the PCE construction does not require information for future conditions but can be made with historically available data available prior to the forecasting period.

3.6.3 Do surrogate and specification sacrifice efficiency?

Our results indicate that a sophisticated combination of three independent techniques (i.e., surrogate modeling, parameter inference, and data assimilation) supplies superior predictive performance for real-time ensemble flood forecasting. The combination of many methods however leads to an essential reduction in efficiency. Because data assimilation has been shown to be necessary, we must accept efficiency deterioration. However, for surrogate modeling and parameter specification, it remains to be determined whether the additional time required by the technique combination leads to efficiency deterioration. First, for construction of the surrogate model, particularly PCE-I, the efficiency issue may not be relevant because the task does not require any observations for calibration and can be completed before the flooding season. In contrast, obtaining an ensemble of parameter sets from posterior distributions should be carried out immediately prior to the flood forecasting period, when observations are necessary. Therefore, it may take an appreciable time for completing this task, and method efficiency may be affected.

3.6.4 What are the differences between PCE and data-driven models?

Both PCE and data-driven models can provide satisfactory results for short-term forecast, but key differences between them exist. (1) PCE has a functionality of including model parameters and states as an input vector – this enables formal uncertainty quantification and model sensitivity analysis; (2) hydrologic/hydraulic model state variables (and parameters) are theoretically observable and in the case of process-based models have their own physical meaning, making it easier to physically interpret the results of PCE; (3) while purely data-driven methods are trained with observations, PCE is trained through high-fidelity samples supervised by physical relations, thus requiring fewer data samples for training; (4) data-driven models often have assumptions about the distributions governing variability of their outputs, and therefore this can lead to non-physical results (e.g., negative outputs quantifying mass, streamflow, etc.) and fail to display non-normal, bi-modal, or other complex behaviors.

3.6.5 Can modeling framework be applied to high-dimensional problems?

While the implementation and analysis of experiments is valid for the presented scope of the experimental design, one needs to proceed with care when extending this approach to more complex models. The most fundamental concern that remains is whether the proposed framework can be applied to high-dimensional problems in which fully distributed models are used. The dimension of a distributed model can be defined as the product of the number of grids cells and the number of parameters (and states). The dimension order of any truly physical models is therefore large, and extending our framework directly to such a model is not straightforward – known as the "curse of dimensionality" [*Caflisch*, 1998; *Davis and Rabinowitz*, 2007; *Sudret*, 2007]. By examining how each of the methods mentioned in the framework resolves the problem of reducing dimensions efficiently and to what extent it has been applied, the feasibility of applying the proposed framework can be estimated.

Regarding the surrogate modelling (PCE), techniques such as Bayesian compressive sensing [*Sargsyan et al.*, 2014] and sparse regression [*Blatman and Sudret*, 2008; *Blatman and Sudret*, 2010] proved capability and efficiency in many prior studies using complex models with high dimensions, up to 80 dimension [*Sargsyan et al.*, 2014]. However, these studies avoided the calculation of fully distributed problems by assuming the spatial variability of parameters to be homogeneous. Second, for the parameter specification, any optimization technique applied to high-dimensional problems could be relevant. For example, one of the large scale optimization algorithms, the competitive swarm optimizer (CSO) [*Cheng and Jin*, 2015] was employed up to the dimension of 5,000. These algorithms have been successfully optimized for problems of very large scale, but their optimizations have been applied to simple analytical functions rather than (hydrologic or meteorological) models. To our knowledge, the number of dimensions has not yet been high in problems of hydrologic optimization, in which the dimension order is almost identical to the number of parameters. The spatial variability of parameters is not fully addressed in most studies, although a "multiplier" concept [*Pokhrel et al.*, 2008]. Last, EnKF is made possible in problems of higher dimensionality through covariance localization. It is mainly applied in meteorological models with many parameters, and the number of dimensions can be up to the order of millions, e.g., 2,592,000 [*Fujita et al.*, 2007]. The localization technique was able to reduce the dimensions efficiently.

3.7 Conclusions

This Chapter presents a new robust, accurate, and efficient modeling framework that consists of the novel integration of three individual techniques: surrogate modeling, parameter inference, and data assimilation. This unified framework is suited for ensemble flood forecasts quantifying prediction uncertainty. The strengths of each technique are (i) the use of PCE offers significant computational savings; (ii) the inference of parameters before data assimilation allows for faster convergence, smaller uncertainties, and greater accuracy of the end results; and (iii) the Kalman filters assimilate errors that occur in real-time flood forecasting. Based on the results of the 18 refined approaches according to the permutations of the above methods, the following conclusions can be drawn:

- Of the two methods for PCE construction, only PCE-I (constructed based on prior, uniform distributions) is acceptable for forecasting, although both methods reproduce observations of the calibration period well. Note that PCE-II (constructed based on posterior distributions) does not provide satisfactory results, even when coupled with other inference and assimilation techniques. The results obtained from PCE-I are similar, and in some cases even superior to those based on the original deterministic NAM model. The PCE used is a single model constructed before the forecast period and thus does not change over time — this is a unique feature different from previous studies in which PCE was rebuilt at each calibration or forecasting time step.
- Especially for short-range forecasting, model parameter input and state initialization plays a crucial role. In some previous studies, posterior distributions were employed to derive a parameter ensemble before forecasting, but the effect of such parameter specification was not quantified for the data assimilation. *Selected* parameter specification (made through the GLUE framework in this study) offers improved accuracy and predictability of forecast outcomes over the *Random* parameter specification. However, it is less computationally efficient, and the issue is expected to be especially problematic when using complex deterministic models.
- The usefulness of single and dual EnKFs is demonstrated through comparisons of the 18 approaches. Both techniques have excellent overall performance, but the dual EnKF showed a slightly better performance than the single EnKF. There was a remarkable improvement in reproducing the hydrograph peak values (Table 3.3). In the absence of assimilation, the *Selected* approach offers superior results and if it cannot be used, data assimilation must be applied.

- The computational time discussed in this study consists of three principal components: surrogate building time, parameter inference time, and data assimilation time. Our conclusions may marginally vary depending on the particular model used and the region in which it is applied, but here the efficiency improvement from using the surrogate modeling technique overwhelms any efficiency deterioration derived from the other two components. That is, the use of the surrogate model makes it possible to effectively address computational efficiency. This feasibility is maximized when many ensemble outcomes are needed and when complex, physically-based models should be simulated.
- From the comprehensive analyses presented above, A15 is our first choice and A14 is the second. When only a deterministic model is used, we recommend A12 (or A11). Using the unified framework developed here, real-time and ensemble flood forecasting are promising directions, allowing for satisfactory measures of accuracy, predictability, and efficiency. Ultimately, the framework developed in this dissertation contributes to a shift in modeling paradigm arguing that complex, high-fidelity, physical hydrologic and hydraulic models should be increasingly adopted for real-time and ensemble flood forecasting.

CHAPTER IV

A novel surrogate data assimilation for real-time ensemble flood forecasting

"Knowing is not enough, we must apply.
Willing is not enough, we must do"
- *(Lee, B)*

4.1 Introduction

Making accurate and timely predictions of floods, one of the natural disasters that cause enormous economic damage and casualties, is a major task in hydrology [*Moradkhani and Sorooshian*, 2008]. In order to alert communities immediately and support emergency response plans, forecasting flood in real-time plays a crucial role; however, it suffers from inherent difficulties due to epistemic and aleatoric uncertainties associated with future conditions of rainfall forcing, initial and boundary conditions, and model parameters [*Beven*, 1989; *Ajami et al.*, 2007; *Kim et al.*, 2016b; *Kim et al.*, 2016c; *Beven et al.*, 2018; *Dwelle et al.*, 2019]. As the development of in situ or remote sensing techniques over several decades makes it possible to collect real-time observation, data assimilation (DA) has proven to be one of the most effective ways to improve the performance and quantify the uncertainty of real-time predictions [*Liu et al.*, 2012]. The central idea of DA is to find a way to reduce the bias of hydrological model states and/or parameters sequentially by incorporating real-time observations into pre-forecasted results [*Evensen*, 1994; *Clark et al.*, 2008; *Moradkhani and Sorooshian*, 2008]. At present, DA techniques are becoming more and more sophisticated, from simple rule-based direct insertion to advanced smoothing and sequential techniques [*Liu et al.*, 2012].

Despite the effectiveness of data assimilation, the computational burden required to perform model evaluations in real time remains an obstacle [*Liu et al.*, 2012; *Houtekamer and Zhang*, 2016; *Bannister*, 2017; *Loos et al.*, 2020]. Since DA includes the process of predicting and updating states (and model outputs), the number of model evaluations should be carried out multiple times compared to predicting only once when DA is not applied. If a model takes a lot of computation time, this problem can become even more critical. Examples are when running hydrological models coupled with the multi-dimensional governing equations of Navier-Stokes [*Marshall et al.*, 1997; *Giraldo and Restelli*, 2008; *Tossavainen et al.*, 2011], Saint-Venant [*Kim et al.*, 2012a; *Kim et al.*, 2012b], Richards [*Maxwell et al.*, 2007; *Kollet et al.*, 2010], and Hairsine-Rose [*Kim et al.*, 2013; *Kim and Ivanov*, 2014] on atmosphere, surface, subsurface, and surface erosion, respectively – their CPU runtimes required for a 1-day simulation generally reach up to the order of days. Furthermore, most DA techniques use ensemble representations for the covariances of forecast (e.g., streamflow) and analysis error of model states (and/or parameters). Maintaining a larger set of ensembles for initial states, model noises, and perturbed observations helps reduce sampling errors that can occur when the state (and parameter) variables are non-Gaussian and non-linear [*Evensen*, 2003; *Li and Xiu*, 2008; *Liu et al.*, 2012; *Slivinski and Snyder*, 2016; *Bannister*, 2017]. The size of the ensemble can be at least $\mathcal{O}(10^2)$ to more than $\mathcal{O}(10^8)$ [*Houtekamer and Zhang*, 2016], resulting in a higher computational barrier. This is indeed not desirable in situations such as flood forecasting where real-time decisions are needed.

Parallel computing, a computational cluster, or cloud computing infrastructure using thousands of processors could be a solution to cope with this concern [*Neal et al.*, 2010;

Houtekamer et al., 2014; *Wittmann et al.*, 2017; *Glenis et al.*, 2018]. However, for problems with fine resolutions or high dimensions, even parallel computations require significant runtime [*Cintra and Velho*, 2018; *Echeverribar et al.*, 2019; *Wing et al.*, 2019; *Hosseiny et al.*, 2020]. The efficiency depends heavily on computer facilities with high processor requirements [*Houtekamer et al.*, 2014]. Another approach that has attracted attention recently is the use of surrogate models, which are used widely in many research areas to mimic nonlinear dynamic models [*Laloy et al.*, 2013; *Dwelle et al.*, 2019; *Tran et al.*, 2020; *Zhang et al.*, 2020]. This surrogate model performs thousands of simulations in seconds, resulting in improved computational efficiency that dramatically reduce CPU runtime [*Razavi et al.*, 2012b; *Asher et al.*, 2015; *Mohanty*, 2015]. The most common methods used to build surrogate models include genetic programming, support vector machines (SVM), artificial neural networks (ANN), Gaussian process emulation (GPE, also called Kriging), and polynomial chaos expansion (PCE) [*Simpson et al.*, 2001; *Wang and Shan*, 2007; *Rajabi*, 2019]. In particular, PCE has shown its merit and attractiveness in recent studies [*Wang and Shan*, 2007; *Rajabi*, 2019], and it turns out to be more effective than the other data-driven methods [*Razavi et al.*, 2012b; *Rajabi*, 2019; *Torre et al.*, 2019]. For example, PCE performs better than the data-driven models (e.g., ANN or SVM) when applied to small training sets and minimal configurations [*Torre et al.*, 2019]. Another merit of using PCE occurs when predicting extreme events beyond the boundaries of training data [*Flood and Kartam*, 1994; *Minns and Hall*, 1996; *Tokar and Johnson*, 1999]. Rare events can be captured more completely because PCE is built based on high-fidelity samples supervised by physical models [*Schöbi et al.*, 2017; *Dubreuil et al.*, 2018; *Tran et al.*, 2020]. Due to its advantages, PCE has been used extensively in many types of hydrological problems from simple to complex [e.g., *Sochala and Le Maître*, 2013;

Sargsyan et al., 2014; *Wang et al.*, 2017; *Dwelle et al.*, 2019; *Tran and Kim*, 2019; *Tran et al.*, 2020].

In the literature, conventional ways of constructing PCE are to build a ‘single’ PCE that can mimic the entire model process, for ‘each’ computational time step; however, these are problematic. When tackling a high-dimensional problem with a large number of uncertain parameters/states, its PCE construction may be challenging due to the exponential increase in the number of PCE coefficients [*Blatman and Sudret*, 2010; *Sargsyan et al.*, 2014; *Konakli and Sudret*, 2016; *Dwelle et al.*, 2019; *Tran et al.*, 2020; *Tran and Kim*, 2021b]. If one provides necessary design (sampling) points to have substantial and sufficient search space for all the dimensions, that points for the model evaluation could be overly large. Advanced techniques such as the least angle regression [*Blatman and Sudret*, 2011] and the Bayesian compressive sensing [*Sargsyan et al.*, 2014] have been able to solve some of the high-dimensional (~ 80 dimension) problems by calculating coefficients only for the most relevant PCE basis terms; but, dealing with a larger dimension is still intricate. Additionally, when making a real-time prediction, the conventional approach requires building a new PCE at every time step prior to forecasting [*Wang et al.*, 2018; *Hu et al.*, 2019a; *Tran et al.*, 2020]. From the context of real-time flood predictions that should be provided within a very short time (usually less than an hour), generating PCE every time is a huge burden. Therefore, finding alternative ways to reduce the dimension or to avoid the recurring PCE construction is the right direction.

Another issue of PCE construction is that to mimic the original model and yield a reliable outcome for a wide range of conditions, PCE should be trained through as many training sets (called experimental design, $\boldsymbol{\mathcal{X}}$) as possible [*Schöbi et al.*, 2017]. Choosing a suitable size for the experimental design is also a challenge [*Razavi et al.*, 2012b]. In order to circumvent this problem, *Blatman and Sudret* [2010] proposed a sequential experimental design (SED) scheme to limit the size of the experimental design when constructing the PCE, minimizing its building time. Specifically, the size of the experimental design (N), one of two important parameters, is increased until a given statistic for computing the error between the surrogate and original models reaches a target value. This scheme allowed us to successfully determine the size of $\boldsymbol{\mathcal{X}}$. However, the order of the polynomial (p), the other parameter that directly affects the performance of the PCE [*Dwelle et al.*, 2019; *Tran and Kim*, 2019], was not considered in that SED scheme. The value of the polynomial degree had to be selected ad-hoc or by trial and error. Additionally, the SED scheme uses only a single convergence criterion for a relative error (e.g., leave-one-out cross-validation error, LOO (see Eq. (2.21))) and sets a small target value (e.g., 10^{-5}) to define acceptance level. However, its inherent assumption that the relative error should decrease monotonically as N or p increases is not always satisfied [*Hu and Youn*, 2010; *Sargsyan et al.*, 2014; *Diaz et al.*, 2018; *Dwelle et al.*, 2019; *Torre et al.*, 2019]. The criterion proposed by the SED scheme may not reach a small target for problems where the error does not decrease continuously, and thus infinite iterations may occur. We need to investigate whether this criterion is appropriate for flood simulations and how to modify the scheme if it is not.

In this Chapter, we present a robust surrogate data assimilation approach based on polynomial chaos expansion theory. This approach includes novel solutions to significantly reduce the computational cost of data assimilation and to effectively determine the optimal N and p of PCE construction. It is organized as follows. Eight filters replacing the Ensemble Kalman filter are developed and necessary modifications to the SED scheme are presented in Section 4.2. From the synthetic and real experimental setups in Section 4.3, the performance of the proposed surrogate filters is evaluated in Section 4.4. Discussions and conclusions drawn from comprehensive analyses continue in Sections 4.5 and 4.6, respectively.

4.2 Methods

4.2.1 Single and dual ensemble Kalman filters

Since the ensemble Kalman filter (EnKF) has become the most broadly used DA technique in many discipline due to its ease of implementation in solving diverse DA conundrums [*Weerts and El Serafy*, 2006; *Clark et al.*, 2008; *Liu et al.*, 2012; *Pathiraja et al.*, 2018; *Tran et al.*, 2020], EnKF was chosen as an original filter to be substituted in this study. Both single and dual EnKFs were used to perform the experiments. Both filters have in common that they update model states at each time step. The main difference is whether these filters update model parameters. Because the internal structure and principles of EnKFs have already been described in detail in the literature, this study will briefly address the key equations that are necessary to describe the surrogate filter we will propose later. More details of EnKFs can be referred to Section 3.2.3.

Regarding a nonlinear (deterministic or statistical) model on discrete time domains, a transition equation for its state ($\boldsymbol{x}$) is described by:

$$\boldsymbol{x}_t^{i-} = f\left(\boldsymbol{x}_{t-1}^{i+}, \boldsymbol{\theta}_{t-1}^{i+}, \boldsymbol{u}_{t-1}^{i}\right) + w_{t-1}^{i}, \qquad i = 1, \dots n \tag{4.1}$$

Different to Chapter III, where parameters are not changed after the forecast step, in this Chapter, we assume that model parameters also follow a random walk, the *i*-th ensemble parameters at time *t* are treated by adding Gaussian noise, τ_{t-1}^{i} with covariance $\boldsymbol{E}_{t-1}^{\boldsymbol{\theta}}$ to $\boldsymbol{\theta}_{t-1}^{i+}$ [*Moradkhani et al.*, 2005c]:

$$\boldsymbol{\theta}_t^{i-} = \boldsymbol{\theta}_{t-1}^{i+} + \tau_{t-1}^{i}, \qquad \tau_{t-1}^{i} \sim N\left(0, \boldsymbol{E}_{t-1}^{\boldsymbol{\theta}}\right) \tag{4.2}$$

The *i*-th prediction (model output) at time *t*, y_t^{i-} is the function of the model states, parameters, and forcings:

$$y_t^{i-} = h\left(\boldsymbol{x}_t^{i-}, \boldsymbol{\theta}_t^{i-}, \boldsymbol{u}_t^{i}\right) \tag{4.3}$$

The *i*-th ensemble parameters at time *t*, $\boldsymbol{\theta}_t^{i+}$ are corrected by using the *Kalman gain,* $K_t^{\boldsymbol{\theta}}$ and the *i*-th observations at time *t*, $y_t^{obs,i}$:

$$\boldsymbol{\theta}_t^{i+} = \boldsymbol{\theta}_t^{i-} + K_t^{\boldsymbol{\theta}}\left(y_t^{obs,i} - y_t^{i-}\right) \tag{4.4}$$

With these updated ensemble parameters $\boldsymbol{\theta}_t^{i+}$, the prediction is updated once more with an equation similar to Eq. (4.3):

$$y_t^{i+} = h\left(\boldsymbol{x}_t^{i-}, \boldsymbol{\theta}_t^{i+}, \boldsymbol{u}_t^{i}\right) \tag{4.5}$$

Finally, the *i*-th ensemble model states that at time *t*, $\boldsymbol{x}_t^{i+}$ are corrected by using the *Kalman gain* $\mathbf{K}_t^{x}$ and perturbed observations:

$$\boldsymbol{x}_t^{i+} = \boldsymbol{x}_t^{i-} + K_t^{x}\left(y_t^{obs,i} - y_t^{i+}\right) \tag{4.6}$$

The equations above from Eq. (4.1) to Eq. (4.6) all correspond to the Dual EnKF. For the single EnKF, all the equations are identical except for Eqs. (4.2), (4.4), and (4.5): the noise parameter in Eq. (4.2) is disregarded, Eq. (4.4) is assumed as $\boldsymbol{\theta}_t^{i+} = \boldsymbol{\theta}_t^{i-}$, and Eq. (4.5) is then simplified to $y_t^{i+} = y_t^{i-}$.

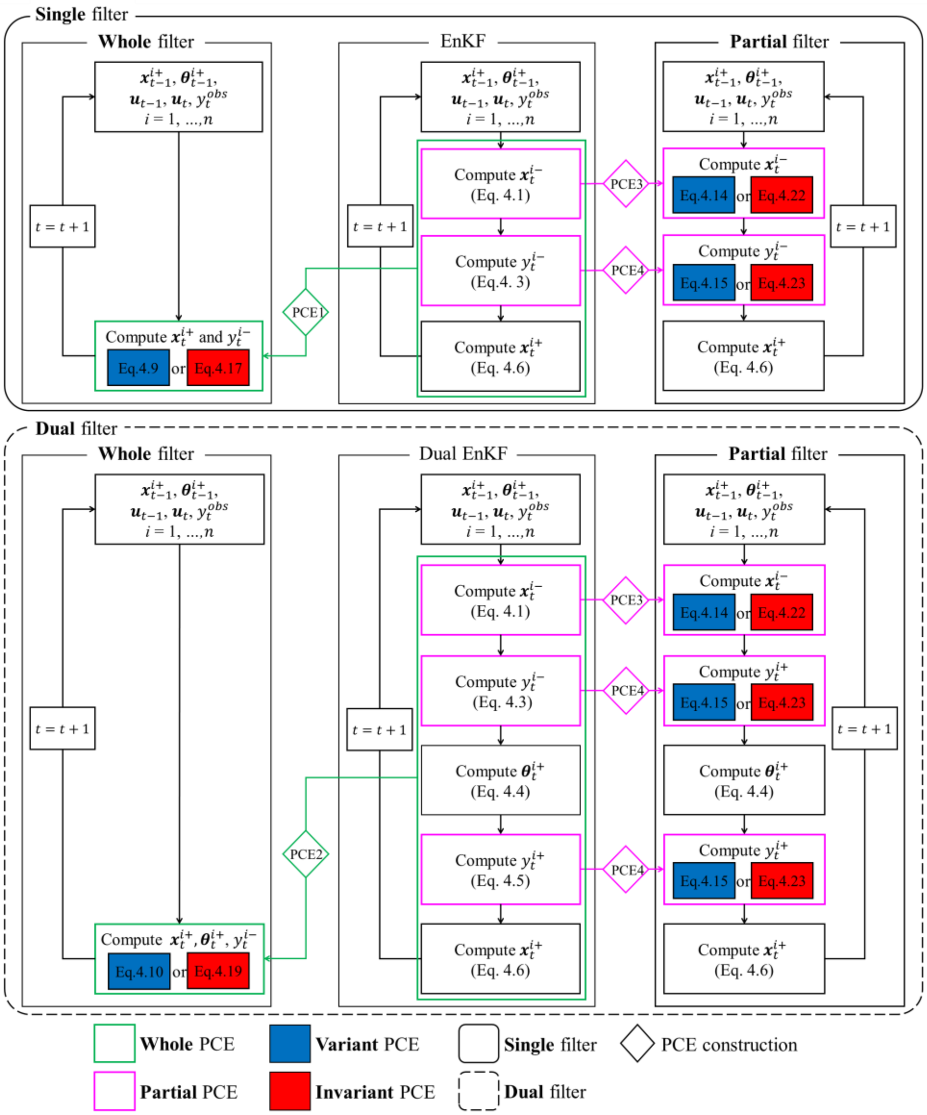
Combining the above six equations, the final form of EnKFs can be expressed as:

$$\left[\boldsymbol{x}_t^{i+}, y_t^{i-}\right] = \text{EnKF}\left(\boldsymbol{x}_{t-1}^{i+}, \boldsymbol{\theta}_{t-1}^{i+}, \boldsymbol{u}_{t-1}^{i}, \boldsymbol{u}_t^{i}, y_t^{obs,i}\right) \tag{4.7}$$

$$\left[\boldsymbol{x}_t^{i+}, \boldsymbol{\theta}_t^{i+}, y_t^{i-}\right] = \text{Dual EnKF}\left(\boldsymbol{x}_{t-1}^{i+}, \boldsymbol{\theta}_{t-1}^{i+}, \boldsymbol{u}_{t-1}^{i}, \boldsymbol{u}_t^{i}, y_t^{obs,i}\right) \tag{4.8}$$

4.2.2 The development of eight surrogate filters

Here, we present eight surrogate filters based on polynomial chaos expansion (PCE) theory that can be substituted for EnKFs. The proposed filters expect to achieve high accuracy at low computational cost and inherit the same fundamental assumptions as the EnKF implementations. The differences among the eight filters will be described below and in Fig. 4.1.



Single filter
Whole filter
EnKF
Partial filter
$\boldsymbol{x}_{t-1}^{i+}, \boldsymbol{\theta}_{t-1}^{i+}, \boldsymbol{u}_{t-1}, \boldsymbol{u}_t, y_t^{obs}$ $i = 1, \ldots, n$
$t = t + 1$
Compute $\boldsymbol{x}_t^{i+}$ and y_t^{i-}
Eq.4.9 or Eq.4.17
PCE1
Compute $\boldsymbol{x}_t^{i-}$ (Eq. 4.1)
Compute y_t^{i-} (Eq.4. 3)
Compute $\boldsymbol{x}_t^{i+}$ (Eq. 4.6)
PCE3
PCE4
Compute $\boldsymbol{x}_t^{i-}$
Eq.4.14 or Eq.4.22
Compute y_t^{i-}
Eq.4.15 or Eq.4.23
Compute $\boldsymbol{x}_t^{i+}$ (Eq. 4.6)
Dual filter
Whole filter
Dual EnKF
Partial filter
$\boldsymbol{x}_{t-1}^{i+}, \boldsymbol{\theta}_{t-1}^{i+}, \boldsymbol{u}_{t-1}, \boldsymbol{u}_t, y_t^{obs}$ $i = 1, \ldots, n$
$t = t + 1$
Compute $\boldsymbol{x}_t^{i+}, \boldsymbol{\theta}_t^{i+}, y_t^{i-}$
Eq.4.10 or Eq.4.19
PCE2
Compute $\boldsymbol{x}_t^{i-}$ (Eq. 4.1)
Compute y_t^{i-} (Eq. 4.3)
Compute $\boldsymbol{\theta}_t^{i+}$ (Eq. 4.4)
Compute y_t^{i+} (Eq. 4.5)
Compute $\boldsymbol{x}_t^{i+}$ (Eq. 4.6)
PCE3
PCE4
PCE4
Compute $\boldsymbol{x}_t^{i-}$
Eq.4.14 or Eq.4.22
Compute y_t^{i+}
Eq.4.15 or Eq.4.23
Compute $\boldsymbol{\theta}_t^{i+}$ (Eq. 4.4)
Compute y_t^{i+}
Eq.4.15 or Eq.4.23
Compute $\boldsymbol{x}_t^{i+}$ (Eq. 4.6)
Whole PCE
Partial PCE
Variant PCE
Invariant PCE
Single filter
Dual filter
PCE construction

Figure 4.1. Flowchart for the construction of the eight proposed surrogate filters. The eight filters consist of the permutations of 2 × 2 × 2 subcases: Whole or Partial surrogate structures (left vs. right panels) × Invariant or Variant surrogate building systems (blue vs. red boxes) × Single or Dual assimilating targets (top vs. bottom panels). The rhombuses show the extent to which EnKF has been replaced.

4.2.2.1 Different surrogate structures: whole vs. partial surrogate filters

The first criterion for distinguishing among the eight surrogate filters is to determine what to replace in the existing EnKF system. One possibility is to construct a single surrogate that can mimic the entire EnKF process. Given all the input variables on the right side of Eqs. (4.7) and (4.8), the output variable on the left side is computed through the surrogate filters. This is similar to the conventional approach, hereafter called Surrogate Whole Filter (SuWF). The other new approach is to configure multiple surrogates, each of which mimics a specific process in the EnKFs. Theoretically, it is possible to distinguish all the processes of EnKFs and construct surrogates for all of them, but two independent surrogates are only built for two of the processes in this study. These processes were chosen because they are the most unfavorable in terms of computational efficiency but have a great impact on accuracy [*Li and Xiu*, 2009]. The processes are those computing ensemble states and streamflow forwarded in time through the propagators $f(.)$ and $h(.)$ in Eqs. (4.4.1), (4.3), and (4.5). Regarding the EnKF processes other than these two, the same applies as with the existing EnKFs. This is hereafter called the Surrogate Partial Filter (SuPF).

4.2.2.2 Different building systems: variant vs. invariant surrogate filters

The second criterion is whether or not the PCE employed to create the surrogate filters changes over time. One conventional type of PCE is a time-variant PCE (VaPCE), which is continuously reconstructed based on new information about forcings and streamflow at all forecasting steps of data assimilation [*Fan et al.*, 2016; *Wang et al.*, 2017; *Wang et al.*, 2018; *Dwelle et al.*, 2019; *Hu et al.*, 2019a; *Tran and Kim*, 2019]. The collection of the training set is much more straightforward than in the other type because precipitation forcing and streamflow are known during the assimilation process (that is, $\boldsymbol{u}_{t-1}, \boldsymbol{u}_t$ and y_t^{obs} are real data) [*Tran et al.*, 2020].

We will refer to the surrogate filters created using the time-variant PCE as Variant Surrogate Filters (VaSuFs, i.e., VaSuWF and VaSuPF).

The other type of PCE is built only once over the calibration period and is used later for the forecasting period [*Tran et al.*, 2020]. This time-invariant PCE (hereafter referred to as InPCE) has the advantage of maximizing applicability and efficiency because it does not need to be re-built during real-time forecasting. One notes that InPCE should be made to better represent the behavior of the original model under a wide range of conditions. Trained for a limited dataset from the past, the PCE model has shown excellent performance with other events similar to the calibration sets. However, for events that differ from those in the training series, it is challenging to construct a surrogate model that mimics the original model [*Tran et al.*, 2020]. This issue can be addressed by training with as much data as possible, but attaining large amounts of data is still far away for some (ungauged) domains. We therefore propose a new procedure of collecting training sets to build an invariant PCE that does not require measurement data. In this study, we assume that all input variables of a filter (model) in Eqs. (4.7) and (4.8), i.e., model states, model parameters, rainfall measurements, and observed discharges, are uncertain and vary within a particular range. Instead of arranging them in a deterministic way, their input values are stochastically perturbed through a sampling process. The Latin hypercube (LHS) sampling technique [*McKay et al.*, 1979b] is used in this work. This procedure makes the invariant PCE suitable for as many input conditions as possible. We will refer to the surrogate filters created by the invariant PCE as Invariant Surrogate Filters (InSuFs, i.e., InSuWF and InSuPF).

Such a differentiating criterion is related to the amount of information needed to generate a PCE, that is, whether to select training information for just a single time step or all periods. It thus involves setting the extent to which the generated PCE can replace the original model. In other words, a PCE generated using a wider range of training data can mimic the results of the original filter in a wider range of simulation conditions. A more detailed comparison and discussion on this subject can be found in Chapter III.

4.2.2.3 Different assimilating targets: single vs. dual surrogate filters

The last criterion is based on the goal of data assimilation: whether to update only state vectors (like single EnKF in Eq. (4.7)) or to update model parameters as well as the state vectors (like Dual EnKF in Eq. (4.8)). Surrogate filters that replace single and Dual EnKF can be called Single Surrogate filters (SuFs, i.e., InSuWF, InSuPF, VaSuWF, and VaSuPF) and Dual Surrogate filters (Dual SuFs, i.e., Dual InSuWF, Dual InSuPF, Dual VaSuWF, and Dual VaSuPF), respectively.

The mathematical denotations of these eight surrogate filters are presented in Appendix A for clarity. This includes the denotations for both surrogate filters and their corresponding PCEs. The latter specifically consists of 4 variant PCEs (VaPCE1, VaPCE2, VaPCE3, and VaPCE4) and 4 invariant PCEs (InPCE1, InPCE2, InPCE3, and InPCE4).

4.2.2.4 Denotation of eight surrogate filters

The eight surrogate filters proposed (Fig. 4.1) consist of 2 × 2 × 2 subcases (whole or partial × variant or invariant × single or dual). Denotation of the filters are described below.

The Single Variant Surrogate Whole filter (VaSuWF) has the same mathematical form as Eq. (4.7). The entire set of processes of the EnKF is replaced with a (first) variant PCE (named VaPCE1), where $\mathrm{VaSuWF} = \mathrm{VaPCE1}$.

$$\left[\boldsymbol{x}_t^{i+}, y_t^{i-}\right] = \mathrm{VaSuWF}\left(\boldsymbol{x}_{t-1}^{i+}, \boldsymbol{\theta}_{t-1}^{i+}, \boldsymbol{u}_{t-1}^{i}, \boldsymbol{u}_t^{i}, y_t^{obs,i}\right), \quad i = 1, \dots n \tag{4.9}$$

$$\left[\boldsymbol{x}_t^{i+}, y_t^{i-}\right] = \mathrm{VaPCE1}\left(\boldsymbol{x}_{t-1}^{i+}, \boldsymbol{\theta}_{t-1}^{i+}, \boldsymbol{u}_{t-1}^{i}, \boldsymbol{u}_t^{i}, y_t^{obs,i}\right), \quad i = 1, \dots n \tag{4.10}$$

The Dual Variant Surrogate Whole filter (Dual VaSuWF) has the same mathematical form as Eq. (4.8). The entire set of processes of the Dual EnKF is replaced with a (second) variant PCE (VaPCE2), where Dual VaSuWF = VaPCE2.

$$\left[\boldsymbol{x}_t^{i+}, \boldsymbol{\theta}_t^{i+}, y_t^{i-}\right] = \text{Dual VaSuWF}\left(\boldsymbol{x}_{t-1}^{i+}, \boldsymbol{\theta}_{t-1}^{i+}, \boldsymbol{u}_{t-1}^{i}, \boldsymbol{u}_t^{i}, y_t^{obs,i}\right), \quad i = 1, \dots n \tag{4.11}$$

$$\left[\boldsymbol{x}_t^{i+}, \boldsymbol{\theta}_t^{i+}, y_t^{i-}\right] = \mathrm{VaPCE2}\left(\boldsymbol{x}_{t-1}^{i+}, \boldsymbol{\theta}_{t-1}^{i+}, \boldsymbol{u}_{t-1}^{i}, \boldsymbol{u}_t^{i}, y_t^{obs,i}\right), \quad i = 1, \dots n \tag{4.12}$$

The Single Variant Surrogate Partial filter (VaSuPF) has the same mathematical form as Eq. (4.7). However, the process of computing $\boldsymbol{x}_t^{i-}$ in the latter EnKF is replaced with a (third) variant PCE (VaPCE3) and the process of computing y_t^i is replaced with a (fourth) variant PCE (VaPCE4), where VaSuPF includes VaPCE3 and VaPCE4.

$$\left[\boldsymbol{x}_t^{i+}, y_t^{i-}\right] = \mathrm{VaSuPF}\left(\boldsymbol{x}_{t-1}^{i+}, \boldsymbol{\theta}_{t-1}^{i+}, \boldsymbol{u}_{t-1}^{i}, \boldsymbol{u}_t^{i}, y_t^{obs,i}\right), \quad i = 1, \dots n \tag{4.13}$$

$$\boldsymbol{x}_t^{i-} = \mathrm{VaPCE3}\left(\boldsymbol{x}_{t-1}^{i+}, \boldsymbol{\theta}_{t-1}^{i+}, \boldsymbol{u}_{t-1}^{i}\right) + w_{t-1}^{i}, \quad i = 1, \dots n \tag{4.14}$$

$$y_t^i = \mathrm{VaPCE4}\left(\boldsymbol{x}_t^{i-}, \boldsymbol{\theta}_t^{i}, \boldsymbol{u}_t^{i}\right), \quad i = 1, \dots n \tag{4.15}$$

The Dual Variant Surrogate Partial filter (Dual VaSuPF) has the same mathematical form as Eq. (4.8). However, the process of computing $\boldsymbol{x}_t^{i-}$ in the latter Dual EnKF is replaced with the third variant PCE (VaPCE3), as shown in Eq. (4.14), and the two processes of computing y_t^i are replaced with the same fourth variant PCE (VaPCE4) as in Eq. (4.15).

$$\left[\boldsymbol{x}_t^{i+}, \boldsymbol{\theta}_t^{i+}, y_t^{i-}\right] = \text{Dual VaSuPF}\left(\boldsymbol{x}_{t-1}^{i+}, \boldsymbol{\theta}_{t-1}^{i+}, \boldsymbol{u}_{t-1}^{i}, \boldsymbol{u}_t^{i}, y_t^{obs,i}\right), \quad i = 1, \dots n \tag{4.16}$$

Another four invariant filters are set up similarly to VaSuFs except for using the time-invariant PCE (InPCE). The fifth filter, Single Invariant Surrogate Whole filter (InSuWF), also has the same mathematical form as Eq. (4.7). The entire set of processes of the EnKF is replaced with an (first) invariant PCE (named InPCE1), where InSuWF = InPCE1.

$$\left[\boldsymbol{x}_t^{i+}, y_t^{i-}\right] = \text{InSuWF}\left(\boldsymbol{x}_{t-1}^{i+}, \boldsymbol{\theta}_{t-1}^{i+}, \boldsymbol{u}_{t-1}^{i}, \boldsymbol{u}_t^{i}, y_t^{obs,i}\right), \quad i = 1, \dots n \tag{4.17}$$

$$\left[\boldsymbol{x}_t^{i+}, y_t^{i-}\right] = \text{InPCE1}\left(\boldsymbol{x}_{t-1}^{i+}, \boldsymbol{\theta}_{t-1}^{i+}, \boldsymbol{u}_{t-1}^{i}, \boldsymbol{u}_t^{i}, y_t^{obs,i}\right), \quad i = 1, \dots n \tag{4.18}$$

The Dual Invariant Surrogate Whole filter (Dual InSuWF) has the same mathematical form as Eq. (4.8). The entire set of processes of the Dual EnKF is replaced with an (second) invariant PCE (InPCE2), where Dual InSuWF = InPCE2.

$$\left[\boldsymbol{x}_t^{i+}, \boldsymbol{\theta}_t^{i+}, y_t^{i-}\right] = \text{Dual InSuWF}\left(\boldsymbol{x}_{t-1}^{i+}, \boldsymbol{\theta}_{t-1}^{i+}, \boldsymbol{u}_{t-1}^{i}, \boldsymbol{u}_t^{i}, y_t^{obs,i}\right), \quad i = 1, \dots n \tag{4.19}$$

$$\left[\boldsymbol{x}_t^{i+}, \boldsymbol{\theta}_t^{i+}, y_t^{i-}\right] = \text{InPCE2}\left(\boldsymbol{x}_{t-1}^{i+}, \boldsymbol{\theta}_{t-1}^{i+}, \boldsymbol{u}_{t-1}^{i}, \boldsymbol{u}_t^{i}, y_t^{obs,i}\right), \quad i = 1, \dots n \tag{4.20}$$

The Single Invariant Surrogate Partial filter (InSuPF) has the same mathematical form as Eq. (4.7). However, the process of computing $\boldsymbol{x}_t^{i-}$ in the EnKF is replaced with an (third) invariant PCE (InPCE3) and the process of computing y_t^i is replaced with an (fourth) invariant PCE (InPCE4).

$$\left[\boldsymbol{x}_t^{i+}, y_t^{i-}\right] = \text{InSuPF}\left(\boldsymbol{x}_{t-1}^{i+}, \boldsymbol{\theta}_{t-1}^{i+}, \boldsymbol{u}_{t-1}^{i}, \boldsymbol{u}_t^{i}, y_t^{obs,i}\right), \quad i = 1, \dots n \tag{4.21}$$

$$\boldsymbol{x}_t^{i-} = \text{InPCE3}\left(\boldsymbol{x}_{t-1}^{i+}, \boldsymbol{\theta}_{t-1}^{i+}, \boldsymbol{u}_{t-1}^{i}\right) + w_{t-1}^{i}, \quad i = 1, \dots n \tag{4.22}$$

$$y_t^i = \text{InPCE4}\left(\boldsymbol{x}_t^{i-}, \boldsymbol{\theta}_t^{i}, \boldsymbol{u}_t^{i}\right), \quad i = 1, \dots n \tag{4.23}$$

The Dual Invariant Surrogate Partial filter (Dual InSuPF) has the same mathematical form as Eq. (4.8). However, the process of computing $\boldsymbol{x}_t^{i-}$ in the Dual EnKF is replaced with a third invariant PCE (InPCE3), as in Eq. (4.22), and the two processes of computing y_t^i are replaced with the same fourth invariant PCE (InPCE4), as in Eq. (4.23).

$$\left[\boldsymbol{x}_t^{i+}, \boldsymbol{\theta}_t^{i+}, y_t^{i-}\right] = \text{Dual InSuPF}\left(\boldsymbol{x}_{t-1}^{i+}, \boldsymbol{\theta}_{t-1}^{i+}, \boldsymbol{u}_{t-1}^{i}, \boldsymbol{u}_t^{i}, y_t^{obs,i}\right), \quad i = 1, \dots n \tag{4.24}$$

4.2.3 PCE based surrogate filters

To develop the surrogate filters explained in Section 4.2.2 and handle the computational burden mentioned above, a PCE is used. Given inputs $\mathbf{X}_t^i$ and outputs $\mathbf{Y}_t^i$ of $\boldsymbol{\mathcal{M}}$ shown in Eqs. (4.10), (4.12), (4.14), (4.15), (4.18), (4.20), (4.22), and (4.23), the relationship between $\boldsymbol{X}_t^i$ and $\boldsymbol{Y}_t^i$ is described as:

$$\boldsymbol{Y}_t^i = \boldsymbol{\mathcal{M}}\left(\boldsymbol{X}_t^i\right) \approx \boldsymbol{\mathcal{M}}^{PCE}\left(\boldsymbol{X}_t^i\right) = \sum_{\alpha \in \mathbb{N}^{N_X}} \varepsilon_\alpha \Psi_\alpha\left(\boldsymbol{X}_t^i\right), \quad i = 1, \dots n \tag{4.25}$$

where the numbers of inputs $\boldsymbol{X}_t^i$ and outputs $\boldsymbol{Y}_t^i$ (each member of the PCE outputs is called a quantity of interest, QoI) in Eq. (4.25) are N_X and N_Y, respectively, which vary depending on the aforementioned PCEs. Specifically, the values of N_X are $N_S + N_P + 2N_I + 1$, $N_S + N_P + 2N_I + 1$, $N_S + N_P + N_I$, and $N_S + N_P + N_I$ for PCE1, PCE2, PCE3, and PCE4, respectively, regardless of whether the PCEs are variant or invariant, whereas the corresponding values of N_Y are $N_S + 1$, $N_S + N_P + 1$, N_S, and 1 for the 4 PCEs, respectively. In this Chapter, LAR is used to construct surrogate filters. More information and discussion about LAR can be found in Section 2.3 of this dissertation.

4.2.4 Optimization of PCE hyper-parameters

The constructed PCE includes two hyper-parameters: the size of experimental design N and the degree of polynomials p. It is important to set the values of the hyper-parameters prior to estimating the PCE coefficients, which directly affects the ability to capture the behavior of the original filter [*Blatman and Sudret*, 2010]. However, such a determination is not straightforward and often has been decided by trial and error [*Hu and Youn*, 2010; *Laloy et al.*, 2013; *Sochala and Le Maître*, 2013; *Wang et al.*, 2017; *Dwelle et al.*, 2019; *Tran and Kim*, 2019]. To address this problem, we propose a procedural version in which the ultimate values of N and p are identified by continuously increasing their values until any convergence criteria are met. This method is the extension of the *sequential experimental design* [*Blatman and Sudret*, 2010].

4.2.4.1 Sequential experimental design – polynomial degree scheme

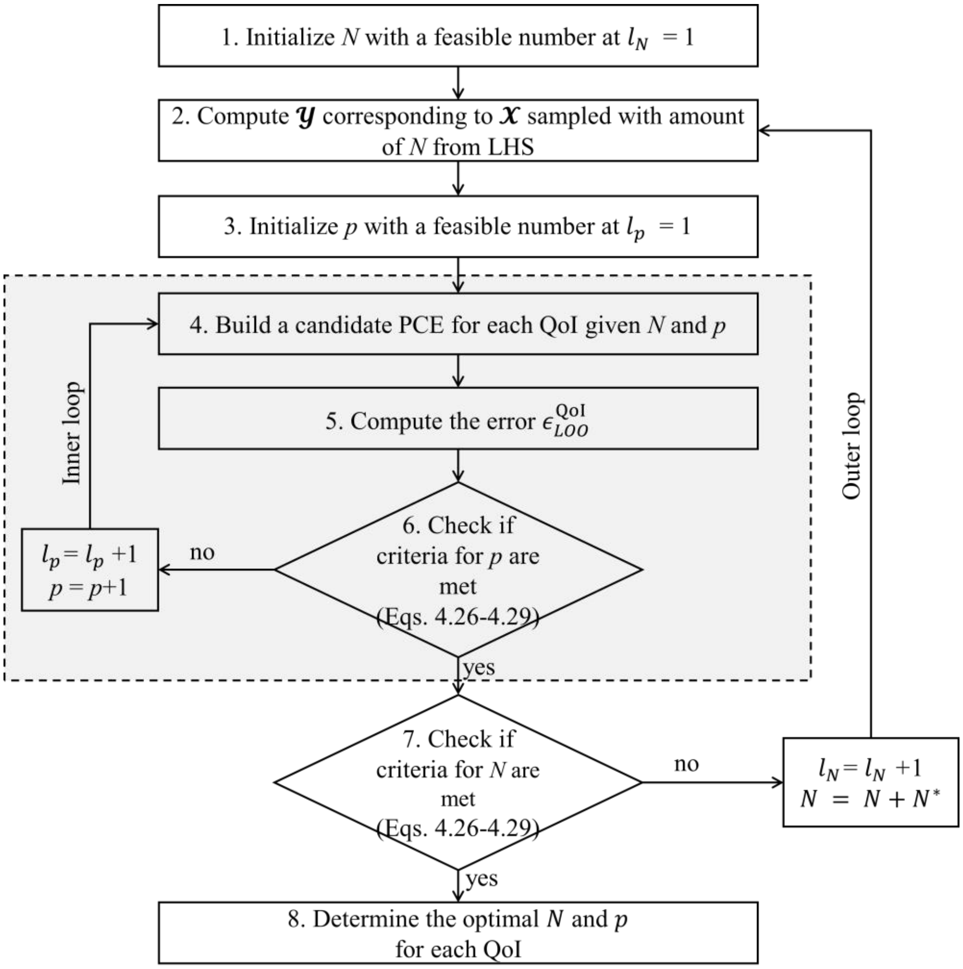

1. Initialize N with a feasible number at $l_N = 1$
2. Compute $\boldsymbol{\mathcal{Y}}$ corresponding to $\boldsymbol{\mathcal{X}}$ sampled with amount of N from LHS
3. Initialize p with a feasible number at $l_p = 1$
4. Build a candidate PCE for each QoI given N and p
5. Compute the error $\epsilon_{LOO}^{\text{QoI}}$
6. Check if criteria for p are met (Eqs. 4.26-4.29)
no
$l_p = l_p + 1$
$p = p+1$
Inner loop
yes
7. Check if criteria for N are met (Eqs. 4.26-4.29)
no
$l_N = l_N + 1$
$N = N + N^*$
Outer loop
yes
8. Determine the optimal N and p for each QoI

Figure 4.2. Flowchart of the sequential experimental design – polynomial degree (SED-PD) scheme proposed for optimizing N and p in building PCE for each quantity of interest (QoI). The SED-PD scheme includes two iterative loops. The first (inner) loop is inside the grey box, aiming to optimize the value of p. The second (outer) loop includes the first one, allowing for the determination of the optimal values of N and p. N^* is an increment that determines the amount of sample added to the existing sample for each iteration. The leave-one-out cross-validation error ($\epsilon_{LOO}^{\text{QoI}}$) is used.

In the original *sequential experimental design* (SED) scheme suggested by *Blatman and Sudret* [2010], the enrichment of experimental design ($\boldsymbol{\mathcal{X}}$) is stopped when an accuracy metric reaches a target error. That is, a single stopping criterion was employed. Because some studies do not meet its intrinsic assumption that the accuracy metric decreases monotonically [*Hu and Youn*, 2010; *Sargsyan et al.*, 2014; *Diaz et al.*, 2018; *Dwelle et al.*, 2019; *Torre et al.*, 2019], this study extends the existing scheme so that multiple stopping (convergence) criteria can be satisfied sequentially by enriching both N and p in the so-called *sequential experimental design – polynomial degree* (SED-PD) scheme. Specifically, this SED-PD consists of two subsequent iterative cycles to determine the optimum values of N and p, in which the sub-loop of p is influenced by the iteration of N enrichment. A stepwise description of the SED-PD scheme is provided as follows, and an associated flowchart is sketched in Fig. 4.2.

1. The first step is to initialize the size N of the experimental design with a feasible number. If one chooses an initial value of N that is too small, it will take a long time to converge; conversely, an initial value that is too large may not converge at all. Therefore, selecting an appropriate number for the initial value of N is an important task. In general, determining the optimal size of the experimental design depends largely on the complexity of the original model, as well as on the computational budget available [*Razavi et al.*, 2012b]. The relevant literature has not reported a well-established rule of thumb for the initialization of N to build the PCE [*Blatman and Sudret*, 2010; *Schöbi et al.*, 2017; *Diaz et al.*, 2018; *Dubreuil et al.*, 2018; *Dwelle et al.*, 2019; *Torre et al.*, 2019]. Here, we attempt to appraise an approximate order of an initial value for N: regarding the four variant PCEs (i.e., VaPCE1, 2, 3, and 4) in Eqs. (4.10), (4.12), (4.14), and (4.15), N can be expressed in the order of the number of ensemble members N_{PCE} (i.e., initial $N = N_{PCE}$) because

these PCEs reconstructed in each computational time have a single time step (the number of time steps, $N_T = 1$). Regarding the other invariant PCEs (i.e., InPCE1, 2, 3, and 4), the smallest starting value could be the product of the number of time steps, N_T and the number of ensemble members N_{PCE} (i.e., initial $N = N_T \times N_{PCE}$). Such a number indicates that an original filter (model) needs to be computed over the number of ensemble members of the input variables per each time step.

2. Given N, the experimental design $\boldsymbol{\mathcal{X}}$ is sampled using the LHS technique, and the corresponding response $\boldsymbol{\mathcal{Y}}$.

3. The other hyper-parameter, the polynomial degree *p*, needs to be initialized as well. Unlike the initialization of N, the starting value of p can be chosen to be 1 without much effort.

4. Given the values of N and p, a candidate surrogate for each QoI, $\boldsymbol{\mathcal{M}}^{PCE,\mathrm{QoI}}$ is constructed.

5. To quantify the difference between the results of the original and surrogate filters, judge the degree of convergence, and evaluate the performance of the latter filters, a statistic is introduced and will be computed for each QoI for each iteration. Following the study by *Blatman and Sudret* [2010], the accuracy metric, LOO, is computed for each QoI, hereafter specified as $\epsilon_{LOO}^{\mathrm{QoI}}$ (see Eq. (2.21)). This leave-one-out cross-validation error metric, $\epsilon_{LOO}^{\mathrm{QoI}}$ is designed to quantify how exactly the surrogate model behaves compared to the original model by computing a deviation between these model outputs; and to detect an overfitting phenomenon more easily [*Blatman and Sudret,* 2010].

6. The inner iterative cycle refers to Steps (4) through (6) (Fig. 4.2) and aims to determine an optimal p given a value of N. If the convergence (stopping) criteria for p are not satisfied, the inner loop is executed again (i.e., l_p increased by one) for the new p increased by one. Such

iterations continue until convergence criteria based on the accuracy metric are met. The details for the criteria will be described in the next section.

7. Once the inner loop is finished, the second bigger iterative cycle begins, comprising Steps (2) through (7). This outer loop includes the former inner loop, determining optimal values of both hyper-parameters. For the fixed value of p determined in the first loop, a similar decision is made as to whether other criteria for N are met. If they are not, the iterative algorithm goes back to Step (2) with the new N increased by N^* (i.e., l_N increased by one). New samples with size N^* are added to the existing samples of the experimental design. Then, Steps (2) through (7) are repeated until these criteria for N are satisfied. Note that each time a new N is chosen and the outer loop executes, p is newly determined within the inner loop. Eventually, both criteria must be satisfied to complete the SED-PD algorithm; optimal values of N and p are determined for each QoI.

4.2.4.2 Convergence (stopping) criteria

A set of multiple convergence (stopping) criteria based on the error metric have been proposed to stop the loop iterations of the SED-PD scheme. This is similar to an optimization problem that can maximize the accuracy of the PCE while minimizing the computational cost required for PCE construction. Following the successive procedure of SED-PD, convergence criteria are applied twice to the selections of N and p, respectively. Four criteria are proposed to ensure stopping the enrichment of p and N (Fig. 4.3). The increase of p or N is stopped if any of the following four convergence criteria are satisfied:

(1) One can stop if the $\epsilon_{LOO}^{\mathrm{QoI}}$ value reaches a sufficiently small target error. It is acceptable that the difference between two comparing models is negligible. The threshold error ϵ_{th}^{lower}

targeted in this study is set to 10^{-5}, similar to the previous study [*Blatman and Sudret*, 2011]. The optimal values are specifically determined by selecting the values of p and N at the moment when $\epsilon_{LOO}^{\text{QoI},l}$ at iteration l ($l = \{l_p, l_N\}$) (see Fig. 4.3) is smaller than the lower threshold. The corresponding mathematical expression is:

$$\text{stop if } \epsilon_{LOO}^{\text{QoI},l} \leq \epsilon_{th}^{lower} \tag{4.26}$$

(2) The first criterion is valid only with the expectation that the error decreases monotonically. However, due to the complexity of the model, which is affected by many sources of uncertainty, the first criterion may not be met within a finite number of iterations. Besides, an excessive increase in the number of N and p does not mean that they always provide a better surrogate model, which can lead to an increase in $\epsilon_{LOO}^{\text{QoI}}$ called over-fitting. Such a phenomenon can be avoided by addition of another criterion. That is, one can stop if the $\epsilon_{LOO}^{\text{QoI}}$ increases in three consecutive iterations and $\epsilon_{LOO}^{\text{QoI},l-2}$ is smaller than ϵ_{th}^{upper}:

$$\text{stop if } \epsilon_{LOO}^{\text{QoI},l-2} \leq \epsilon_{LOO}^{\text{QoI},l-1} \leq \epsilon_{LOO}^{\text{QoI},l} \text{ and } \epsilon_{LOO}^{\text{QoI},l-2} \leq \epsilon_{th}^{upper} \tag{4.27}$$

where $\epsilon_{LOO}^{\text{QoI},l-2}$, $\epsilon_{LOO}^{\text{QoI},l-1}$, and $\epsilon_{LOO}^{\text{QoI},l}$ are the error estimates computed at the successive iterations of $l-2$, $l-1$, and l, respectively. To avoid instances where the iterations stop very early or the optimal $\epsilon_{LOO}^{\text{QoI}}$ value is still large, we also introduced another threshold, ϵ_{th}^{upper}, as a safeguard. In this study, the value of ϵ_{th}^{upper} is set to be 10^{-1}. The optimal values are then determined as the p or N values at the iteration when $\epsilon_{LOO}^{\text{QoI}}$ is the smallest. For example, in Fig. 4.3, p or N value at $l-2$ is selected.

(3) Another possible case where the scheme will not converge is when the error is not small and rarely decreases. By calculating the degree of reduction, i.e., slope, one can take into account

divergence in such a case. This can be stopped if the slope in Eq. (4.28) calculated using the errors estimated in three consecutive iteration steps is insignificant. Thus, the third stopping criterion is:

$$\text{stop if } \frac{\left|\epsilon_{LOO}^{\text{QoI},l} - \epsilon_{LOO}^{\text{QoI},l-1}\right|}{\left|\epsilon_{LOO}^{\text{QoI},l-2} - \epsilon_{LOO}^{\text{QoI},l-1}\right|} \leq \epsilon_{th}^{slope} \; and \; \epsilon_{LOO}^{\text{QoI},l-1} \leq \epsilon_{th}^{upper} \tag{4.28}$$

where 0.05 is used as the value of ϵ_{th}^{slope}, as in prior studies [*Echard et al.*, 2011; *Schöbi et al.*, 2017; *Dubreuil et al.*, 2018]. Like the second criterion, the value of $\epsilon_{LOO}^{\text{QoI}}$ is restricted to be smaller than ϵ_{th}^{upper}. The p or N value is determined at the iteration when $\epsilon_{LOO}^{\text{QoI}}$ is the smallest. For example, p or N values at $l-1$ are selected in Fig. 4.3.

(4) The fourth and final criterion is given in case that divergence or overflow can still occur even though the above three safety measures have been implemented. According to the computational power employed, the p and N values should be limited to avoid infinite iterations of the SED-PD when the above three criteria do not work. That is, the fourth stopping criterion is:

$$\text{stop if}\, p \geq \; p_{max} \; or \; \text{ if}\, N \geq \; N_{max} \tag{4.29}$$

where a maximum degree of polynomials, p_{max}, of 15 is used in this study. The maximum number of experimental design, N_{max}, is set to be $(p_{max}+1)^{Nx}$, analogous to the total number of model evaluations when using the Gaussian quadrature method [*Sudret*, 2008]. The p or N value is determined at the iteration when $\epsilon_{LOO}^{\text{QoI}}$ is the smallest. For example, p or N values at $l-1$ (not l_{max}) are selected in Fig. 4.3.

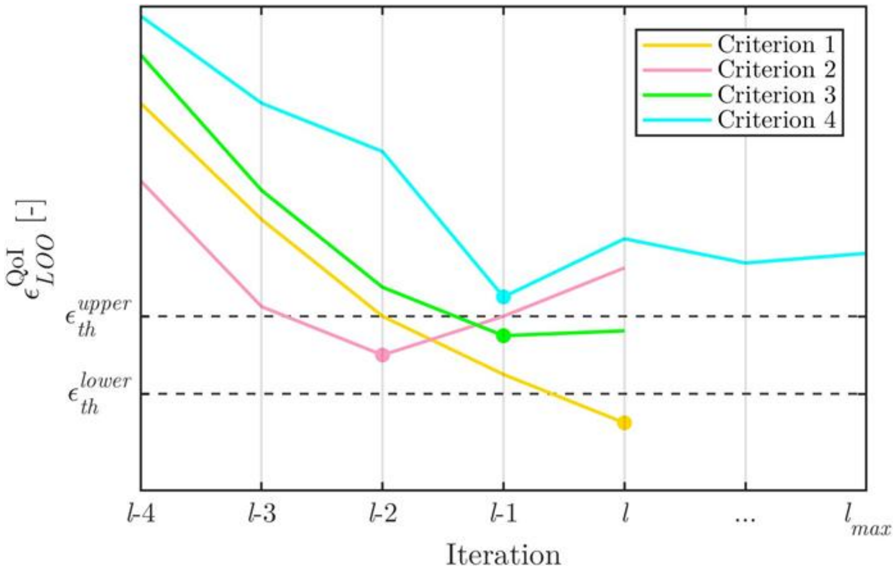


Figure 4.3. A schematic illustration for stopping iterations given four convergence criteria proposed by the SED-PD scheme. The circle point in each line refers to the iteration at which the value of p or N should be determined.

4.3 Experimental configurations

4.3.1 Flood events

All experiments were carried out in the Vu Gia watershed in central Vietnam (see Fig. 4.4 and Section 3.4.1 for more detail) with NAM model (see Section 2.2.1.2). Three flood events from the rainy season in 2016 were selected for verification and application of the surrogate filters. Precipitation data was observed at the Thanh My and Kham Duc stations, and discharge data at the outlet of the watershed, at the Thanh My station (Fig. 4.4a). All the hourly data available in this domain were taken into account and the data are provided by the Vietnam National Centre for Hydro-Meteorological (Fig. 4.4b). The areal average of rainfall used for a hydrological model was computed through the Thiessen method, while the potential evapotranspiration was not considered due to its insignificant effects on flooding.

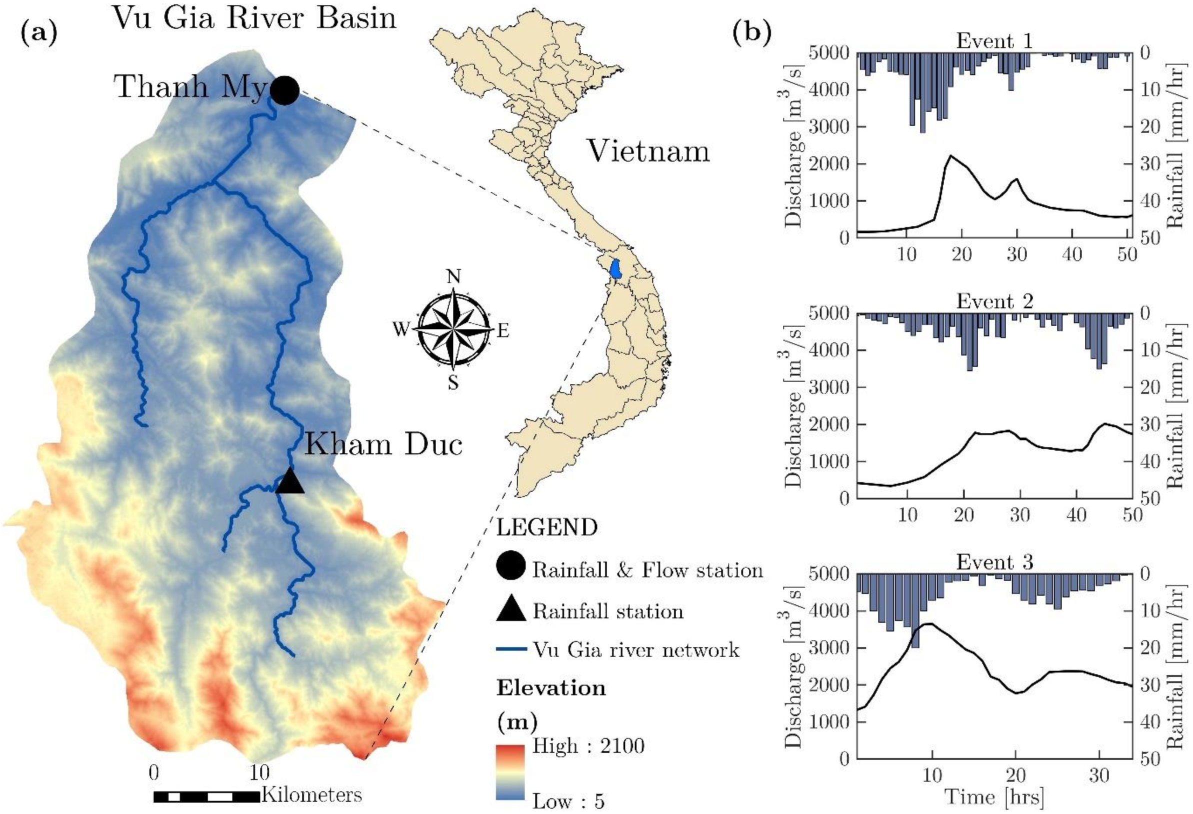


Figure 4.4. (a) Geographical location and topographic characteristics of the Vu Gia watershed, located in central Vietnam, and (b) three flood events used for testing the surrogate filters, including average rainfall over basin (navy bars) and observed discharge (black lines) at the outlet, the Thanh My station.

4.3.2 Modeling procedure for the variant PCE construction

Building the variant PCEs should be performed in real time for the present forcings at each time step. The first task is to initialize the state vector and specify the parameters of the hydrological model: zero initial states are assumed and parameters are sampled from the posterior distributions estimated by the generalized likelihood uncertainty estimation (GLUE) [*Beven and Freer*, 2001] (for more details, see Section 2.2.1.3). Since a much smaller number of training samples (e.g., 50) can make the PCE successfully [*Tran and Kim*, 2019], the initial N of the experimental design $\boldsymbol{\mathcal{X}}$ is given as the number of ensembles, N_{PCE} of 10. In this type of construction, real-time forcings and observations are employed for the EnKF assimilations. Then, the SED-PD scheme stops iterating if criteria for N are met; otherwise, it continues the iterations with the N increased by $N^* = 10$. Note that to construct the variant PCEs, stochastically-varying rainfall and discharge is not essential – real data for the three events is directly employed. Four variant PCEs are built at each time step; that is, 51, 50, and 34 variant PCEs are built for the first, second, and third flood events, respectively.

4.3.3 Modeling procedure for the invariant PCE construction

The first task to constructing invariant PCEs is also to initialize the state vector and specify the parameters of the NAM model. This study initializes the ensemble of the states with zero and specifies the ensemble of the parameters with samples chosen randomly from the Uniform (prior) distribution. These ensembles start with the initial $N = N_T \times N_{PCE} = 100 \times 100 = 10{,}000$. A number of N filter evaluations are followed for rainfall forcings and streamflow observations. As described in Section 4.2.2.2, stochastic inputs of rainfall and discharge are preferred to build the invariant PCEs. We generated the stochastic samples under the assumption that these two input variables followed a Uniform distribution over bounded intervals. The lower bounds of the possible ranges for observed rainfall and discharge were set to be zero, while the upper bounds were subject to the domain. In this study, values of 50 mm/hour and 7,000 m^3/s were employed for the upper bounds of rainfall and streamflow, respectively. These values were determined from the maximum hourly rainfall in data available since 2015 and an extreme corresponding to a 100-year return period. The latter flood frequency analysis was made by applying the Pearson type-III distribution to annual flood peaks for data from 1976 to 2016. Then, according to the SED-PD scheme, additional N^* filter evaluations are included if criteria for N are not met. The N for the second (next) iteration becomes $N = N + N^* = (100 + 10) \times (100 + 10) = 12{,}100$. This size continues to grow until the criteria are satisfied. Note that each of the invariant PCEs is unique and can be applied to any event, including the three events given.

4.3.4 Synthetic and real data assimilation experiments

The first of three events was used as a synthetic experiment, the objective of which is to judge the performance of certain data assimilation techniques whether evaluating the convergence of the parameters, quantifying the parameter range adequately, and minimizing the predictive uncertainty [*Moradkhani*, 2008]. Serving as a control run, this synthetic experimental dataset was generated through a free run using observed rainfall (Event 1), pre-specified (also referred to as "true") parameters in Table 4.1, and discharges computed from the former two. Other flood events (i.e., 2 and 3) were utilized for application of the proposed filters to real-time flood forecasting, in which the performance of the eight surrogate filters are compared and validated with the single and dual ensemble Kalman filters.

Table 4.1. Description, initial range, and predefined value of parameters of NAM.

Parameter	Description	Range	Predefined Value
Um [mm]	Maximum water content in surface storage	[5, 35]	10.99
Lm [mm]	Maximum water content in lower zone/root storage	[50, 400]	302.79
CQOF [-]	Overland flow coefficient	[0, 1]	0.99
CKIF [hrs]	Interflow drainage constant	[200, 2000]	0.22
TOF [-]	Overland flow threshold	[0, 0.9]	0.81
TIF [-]	Interflow threshold	[0, 0.9]	0.60
TG [-]	Groundwater recharge threshold	[0, 0.9]	1237.08
CK12 [hrs]	Time constant for routing interflow/overland flow	[3, 72]	13.77
CKBF [hrs]	Time constant for base flow	[500, 5000]	3482.37

All data assimilation experiments were performed with n of 500, which was considered to be a rational ensemble size to adequately represent the uncertainty bounds. The initialization of the states given the ensemble size was simply set to be zero. The specification of the parameters for the same size was performed by using the posterior distribution by GLUE for the real data experiment, and by using the prior Uniform distribution for the range given in Table 2.3 for the synthetic experiment.

To achieve the most reliable ensemble prediction over the entire forecasting period, it is necessary to assume noise for quantities that contain uncertainty in the data assimilation framework [*Renard et al.*, 2010; *DeChant and Moradkhani*, 2012]. In this study, 500 perturbations were applied to the precipitation and streamflow observations in both synthetic and real data experiments to account for uncertainties. Specifically, we assume a log-normal error distribution with a relative error of 25% for precipitation. The streamflow observation error is assumed to be normally distributed with a relative error of 15% at each time step. It is also assumed that the model parameters follow a random walk by adding a small amount of noise following a normal distribution with a relative error of 1%.

4.3.5 Evaluation measures for accuracy and efficiency

To evaluate the accuracy and predictability of the proposed surrogate filters, both deterministic and probabilistic measures were selected. For deterministic metrics, the NSE and PE are selected as defined in Eqs. (2.9) and (2.10), respectively. For probabilistic measures, the Brier Score (*BS*) (see Eq. (3.36)), the continuously ranked probability score ($CRPS$), and $Spread$ are adopted. Wherein, $CRPS$ measures the proximity of the forecast distribution and the measurement distribution at a single time step [*Gneiting and Raftery*, 2007]. In this study, the temporal mean of $CRPS$, $\overline{CRPS}$ is used for comparison:

$$\overline{CRPS} = \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}} \int_{-\infty}^{\infty} \left[F(y_t^{-}) - F\left(y_t^{obs}\right)\right]^2 dy \tag{4.30}$$

where $F(y_t^{-})$ and $F\left(y_t^{obs}\right)$ are the empirical cumulative distribution of n ensemble predictions y_t^{i-} and the actual observation y_t^{obs} at time *t*, respectively. The value of $\overline{CRPS}$ is non-negative, and has a value of zero if two distributions are identical.

Since a reliable forecast with an excessively high dispersion is not desired, the $Spread$ can be considered. This parameter is equal to the square root of the average ensemble variance over the evaluation period [*Fortin et al.*, 2014; *Liu et al.*, 2019] and is non-negative with the best value of zero. It has the same unit as streamflow:

$$Spread = \sqrt{\frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(y_t^{i-} - y_t^{obs}\right)^2\right]} \tag{4.31}$$

Regarding the modeling efficiency, runtime at each time step (RT_t) and cumulative runtime ($RT_{cum,t}$) are established as:

$$RT_t = RT_{build,t} + RT_{run,t} \times n \tag{4.32}$$

$$RT_{cum,t} = \sum_{t=1}^{t} RT_t \tag{4.33}$$

where $RT_{build,t}$ is the runtime needed for building a filter at each time step of the assimilation; $RT_{run,t}$ is the runtime to perform the filter for assimilation at each time step for one ensemble member. The runtime $RT_{build,t}$ consists of the time $RT_{\mathcal{X},t}$ required to configure the experimental design ($\boldsymbol{\mathcal{X}}$) (i.e., model evaluations) and the time $RT_{opt,t}$ required to determine the optimal hyper-parameters and coefficients of PCE. The latter runtime is the summation of $RT_{opt,t}^{\mathrm{QoI}_m}$ for each QoI because a filter is made after PCE construction over each QoI. In contrast, the former runtime chooses the largest time among all $RT_{\boldsymbol{\mathcal{X}},t}^{\mathrm{QoI}_m}$ because the PCE is built by recycling all the previously performed model evaluations, i.e., it shares the experimental design with the largest N of all QoIs. That is, $RT_{build,t}$ for the variant filters ($RT_{build,t}^{Va}$) is written as:

$$RT_{build,t} = RT_{build,t}^{Va} = \max\left(RT_{\boldsymbol{\mathcal{X}},t}^{\mathrm{QoI}_m} \middle| m = 1, \dots, N_{Y,filter}\right) + \sum_{m=1}^{N_{Y,filter}} RT_{opt,t}^{\mathrm{QoI}_m} \qquad (4.34)$$

where m is an index for the number of PCE outputs of the filter ($N_{Y,filter}$). $N_{Y,filter}$ for SuWF, Dual SuWF, SuPF, and Dual SuPF are $N_S + 1$, $N_S + N_P + 1$, $N_S + 1$, and $N_S + 1$ respectively, regardless of whether the filters are Variant or Invariant Surrogate filters. Note that because the variant surrogate filters need to re-build at every time step of the assimilation, $RT_{build,t}^{Va}$ is subject to the time step t. On the other hand, $RT_{build,t}$ for the invariant filters (RT_{build}^{In}) does not need the subscript t and is independent of time because their construction could be done before forecasting. Thus, RT_{build}^{In} is similarly expressed as:

$$RT_{build,t} = RT_{build}^{In} = \max\left(RT_{\boldsymbol{\mathcal{X}}}^{\mathrm{QoI}_m} \middle| m = 1, \dots, N_{Y,filter}\right) + \sum_{m=1}^{N_{Y,filter}} RT_{opt}^{\mathrm{QoI}_m} \qquad (4.35)$$

4.4 Results

4.4.1 Optimization of the PCE hyper-parameters

Hyper-parameters, N and p must be predetermined to construct PCE for each QoI, but their optimal values to maximize the performance of constructing PCE are unknown. Here, we present the results of the SED-PD scheme, which can be a guideline for other studies. The results are shown in Fig. 4.5, including the convergence criteria used for stopping the scheme and the number of iterations needed for optimizing N and p. Table 4.2 presents the optimal values of N and p, the error $\epsilon_{LOO}^{\mathrm{QoI}}$ at stopping, and the building time. These results are subject to vary depending on the surrogate solutions proposed but are significantly different between the variant and invariant PCEs.

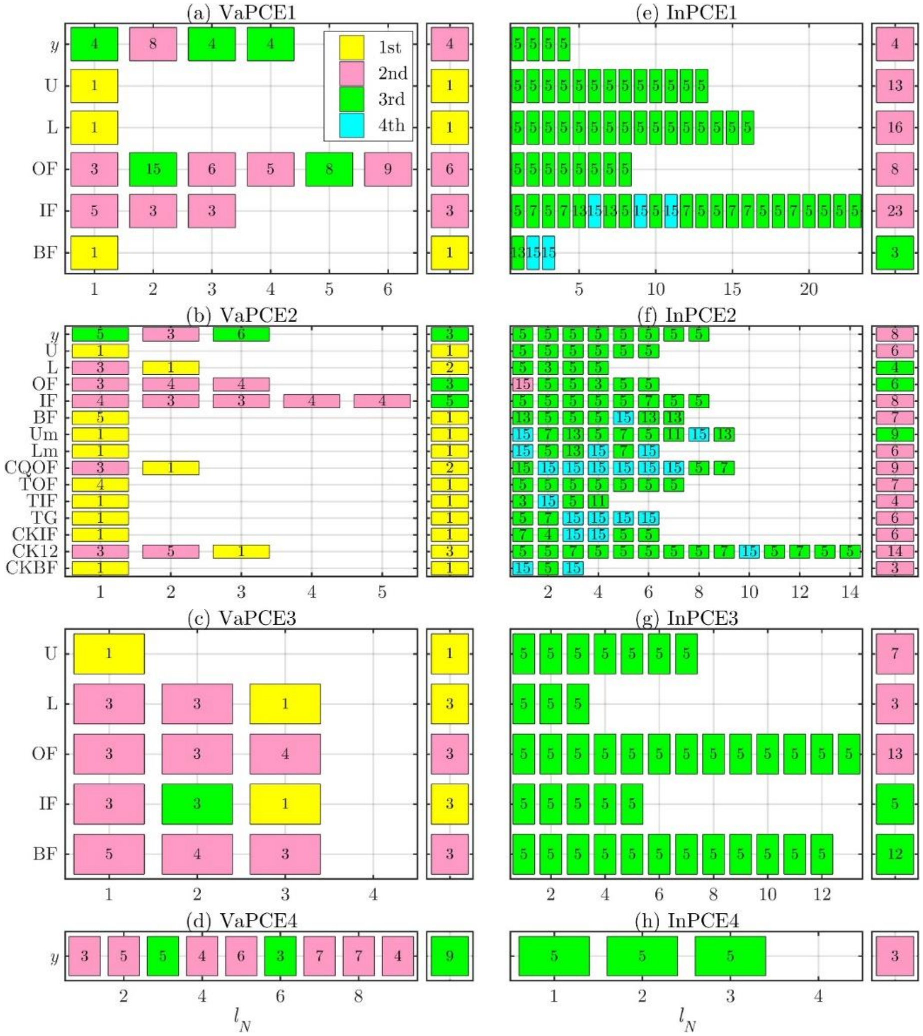


Figure 4.5. Illustrations of the criteria (colored in subplots) used for stopping and the number of iterations (numbered) for each QoI (in y-axis) in optimizing the hyper-parameters, p (the left larger box of each subplot) and N (the right box of each subplot). Subplots, (a) to (h) correspond to constructing different PCEs using the SED-PD scheme. One of the four stopping criteria is shown with colors in each cell for p and N. The number of iterations, l_p and l_N, for optimizing p and N is written inside each cell.

The SED-PD adopts four criteria for attaining the optimal values of N and p. For the selection of p, the second (~57%, in magenta) or the first (~28%, in yellow) stopping criterion is used for the variant PCEs. That is, over-fitting predominantly happens or $\epsilon_{LOO}^{\mathrm{QoI}}$ smaller than the threshold of 10^{-5} exists. The number of iterations l_p is smaller and varies from 1 to 5, specifically 1 (~25 %), 3 (~29 %), 4 (~20%), and 5 (~12 %) (see Fig. 4.5). In contrast, the third (~87%, in green in Fig. 4.5) or fourth (~13%, in cyan) criterion is frequently used to construct the invariant PCEs. That is, the optimal p (p_{opt}^{QoI}) is largely determined when the consecutive values of $\epsilon_{LOO}^{\mathrm{QoI}}$ remain unchanged or the degree of polynomials reaches its maximum value of 15. This implies that it is difficult for the value of $\epsilon_{LOO}^{\mathrm{QoI}}$ to reach its ideal predefined value, ϵ_{th}^{lower}. The number of iterations l_p to optimize p is generally greater than 5, with the most commonly identified numbers of iterations being 5 (~72%), 7 (~16%), and 15 (~29%). These results confirm that finding an optimal p when constructing VaPCEs requires fewer iterations and is much faster than in making InPCEs.

For determining the optimal N of each QoI (N_{opt}^{QoI}), a similar approach using four criteria was made. In building VaPCEs, the most commonly used criteria were 1 (~67%), 2 (~19%), and 3 (~15%), and the numbers of iterations l_N required were only 1 (~48%) or 3 (~30%). These results indicate that the values of $\epsilon_{LOO}^{\mathrm{QoI}}$ easily reach the desired values so the l_N is lower. In contrast, in constructing InPCEs, the criteria used were 2 (~78%) and 3 (~22%), and the values of l_N were mostly from 3 to 8, but ranged up to 23. In summary, the results of the optimization of N and p demonstrate that it is much easier to identify appropriate N and p values in constructing VaPCEs than InPCEs.

Table 4.2. Results of the SED-PD scheme in constructing four PCEs for each QoI, including the optimal sizes of the experimental design N ($N_{opt,t}^{\text{QoI}}$, N_{opt}^{QoI}) and the polynomial degree p ($p_{opt,t}^{\text{QoI}}$, p_{opt}^{QoI}), the magnitude of the accuracy metric at stopping ($\epsilon_{LOO,t}^{\text{QoI}}, \epsilon_{LOO}^{\text{QoI}}$), the runtime needed to perform N model evaluations ($RT_{\boldsymbol{X},\boldsymbol{t}}^{\text{QoI}}, RT_{\boldsymbol{X}}^{\text{QoI}}$), and the runtime to estimate the PCE coefficients ($RT_{opt,t}^{\text{QoI}}, RT_{opt}^{\text{QoI}}$).

	QoI	VaPCE*					InPCE				
		$N_{opt,t}^{\text{QoI}}$	$p_{opt,t}^{\text{QoI}}$	$\epsilon_{LOO,t}^{\text{QoI}}$	Runtime [sec]		N_{opt}^{QoI}	p_{opt}^{QoI}	$\epsilon_{LOO}^{\text{QoI}}$	Runtime [sec]	
					$RT_{\boldsymbol{X},\boldsymbol{t}}^{\text{QoI}}$	$RT_{opt,t}^{\text{QoI}}$				$RT_{\boldsymbol{X}}^{\text{QoI}}$	RT_{opt}^{QoI}
PCE1	y	20	6	7.69E-04	3.2	1.52	12100	5	0.0608	1936	108
	U	10	1	4.63E-31	1.6	0.05	40000	5	0.0294	6400	364
	L	10	1	7.26E-08	1.6	0.05	52900	5	0.0180	8464	1123
	OF	40	3	3.38E-05	6.4	6.72	22500	5	0.0869	3600	278
	IF	10	1	1.00E-04	1.6	0.53	90000	5	0.0079	14400	24712
	BF	10	1	1.26E-06	1.6	0.05	12100	15	0.0111	1936	4403
PCE2	y	30	2	1.11E-04	9.3	0.86	22500	5	0.0869	6975	278
	U	10	1	5.18E-06	3.1	0.05	16900	5	0.0388	5239	103
	L	20	1	3.26E-06	6.2	0.20	14400	5	0.0241	4464	86
	OF	30	2	3.70E-05	9.3	0.60	19600	5	0.0674	6076	1116
	IF	50	2	1.18E-04	15.5	1.33	22500	7	0.0122	6975	292
	BF	10	5	3.01E-07	3.1	0.24	19600	15	0.0094	6076	5021
	Um	10	1	5.18E-06	3.1	0.05	28900	15	0.0076	8959	5489
	Lm	10	1	2.69E-06	3.1	0.04	16900	15	0.0098	5239	5090
	CQOF	20	1	1.86E-07	6.2	0.23	25600	15	0.0083	7936	11486
	CKIF	10	4	1.25E-08	3.1	0.18	19600	5	0.0129	6076	97
	TOF	10	1	6.45E-07	3.1	0.06	12100	15	0.0084	3751	1403
	TIF	10	1	6.88E-06	3.1	0.05	16900	15	0.0084	5239	6801
	TG	10	1	6.73E-08	3.1	0.05	16900	15	0.0091	5239	3038
	CK12	30	1	4.23E-08	9.3	0.47	44100	7	0.0088	13671	4036
	CKBF	10	1	6.74E-08	3.1	0.05	10000	15	0.0109	3100	2376
PCE3	U	10	1	4.85E-32	1.6	0.05	19600	5	0.0595	3136	68
	L	30	1	9.96E-06	4.8	0.33	10000	5	0.0364	1600	148
	OF	10	1	3.96E-04	1.6	0.61	40000	5	0.0531	6400	63
	IF	30	1	8.65E-06	4.8	0.36	16900	5	0.0022	2704	567
	BF	10	3	2.34E-05	1.6	0.66	40000	5	0.0001	6400	130
PCE4	y	80	2	3.37E-04	12.8	15.39	10000	5	0.0528	1600	565

*The results at $t = 50$ are shown over Event 2.

Table 4.2 reports the values of N_{opt}^{QoI}, p_{opt}^{QoI}, $\epsilon_{LOO}^{\text{QoI}}$, $RT_{\boldsymbol{X}}^{\text{QoI}}$ and RT_{opt}^{QoI} in constructing eight PCEs for each QoI. The N_{opt}^{QoI} and p_{opt}^{QoI} values of the variant PCEs are smaller than 100 and 3, respectively, and their $\epsilon_{LOO}^{\text{QoI}}$ values are mostly smaller than 10^{-4}. Conversely, the N_{opt}^{QoI} and p_{opt}^{QoI} values of the invariant PCEs are greater than 10,000 and 5, respectively, and their $\epsilon_{LOO}^{\text{QoI}}$values are greater than 10^{-3}. PCEs generated based on more sampling data (i.e., larger N_{opt}^{QoI}) and more complex models (i.e., higher p_{opt}^{QoI}) do not necessarily provide better results, i.e. smaller errors. Furthermore, the differences in the time required to implement the SED-PD are evident. The build times of the invariant filters are significantly greater than those of the variant filters, by factors of approximately 2962, 3027, 263, and 263 times in the comparisons of InSuWF versus VaSuWF, Dual InSuWF versus Dual VaSuWF, InSuPF versus VaSuPF, and Dual InSuPF versus Dual VaSuPF, respectively (see Table 4.2). These results indicate that a more universal (invariant) PCE built with enormous margins of all uncertain variables takes significantly more time than a specific (variant) PCE in which all forcings and observations are confirmed in real time.

4.4.2 Parameter specification for data assimilation

In Chapter II, we emphasized that the *Selected* parameter specification (sampled from a posterior) provides improved accuracy and predictability of forecast outcomes over the *Random* parameter specification (sampled from a prior). A wide range of random parameters show undesirable and inaccurate forecasts. In this work, GLUE was also used to determine the posterior distribution of the parameters. A likelihood function, L including Nash–Sutcliffe efficiency (NSE), peak error (PE), and volume error (VE) is selected as a likelihood function.

$$L = (1 - NSE) + \frac{PE}{100} + \frac{VE}{100} \tag{4.36}$$

Since individual acceptance thresholds for the NSE, PE, and VE were determined to be 0.8, 5%, and 5%, respectively, according to the robust analysis [*Tran and Kim*, 2019], the posterior parameter sets are obtained when L is smaller than 0.3. The GLUE was applied to Event 1, and then its posterior distribution was assumed to be the initial parameter values needed in the following section.

We estimated the marginal and pairwise marginal posterior distributions of parameters by using GLUE, as illustrated for Event 1 in Fig. 4.6a. The posterior distributions of Lm, CQOF, and CK12 have a pointed shape, indicating that those parameters are highly sensitive to the L and are easily identifiable. The remaining six parameters are almost equally distributed over the entire parameter range, and such flat distributions indicate that their parameters are relatively insensitive to the L and are more uncertain. The same conclusion also can be drawn for the results of the pairwise posterior distributions; namely, that any combinations of six insensitive parameters are evenly distributed over the squared domains, except for the combinations of Lm, CQOF, and CK12.

Additionally, a sensitivity analysis (SA) was carried out in order to identify critical parameters effecting predictive accuracy. By using Sobol' variance-based global sensitivity analysis (see Section 2.2.1.4), both overall interaction of each parameter through the total-order (main) sensitivity index and pairwise (joint) interaction between parameters are shown in Fig. 4.6b. The results also confirmed that Lm, CQOF, and CK12 are the most sensitive parameters to the objective function, while other parameters have relatively low sensitivity. Regarding the joint sensitivities, the diagonal interactions among Lm, COQF, TG, and CK12 illustrated by the connecting lines with larger width and lower opacity principally affect the results of L as compared to other interactions. The main and joint sensitivities using Sobol' indices were consistent with the results inferred from the aforementioned posterior distribution. It is worth noting these results because the most influential parameters can be primarily examined for data assimilation of real-time predictions.

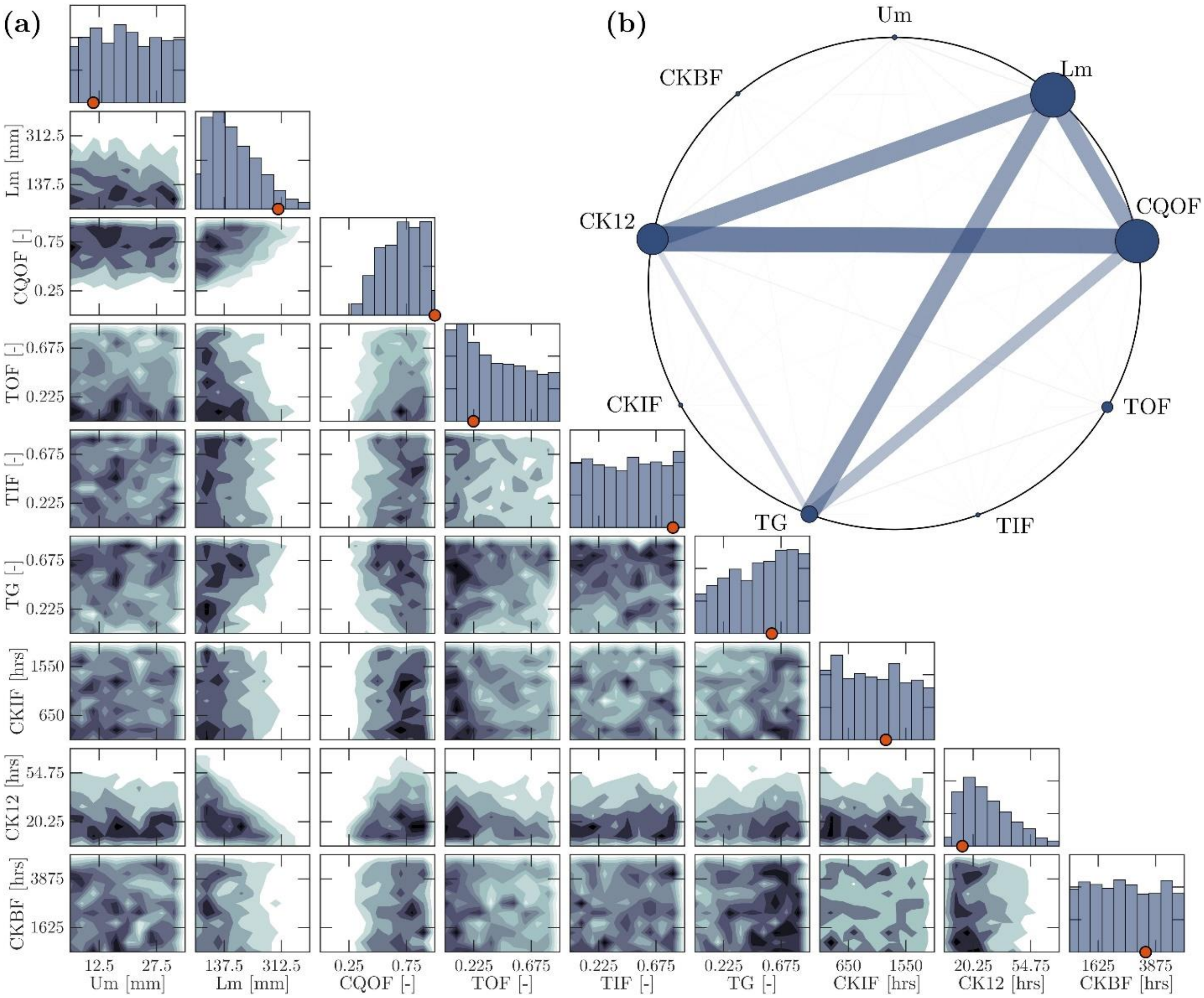


Figure 4.6. (a) The marginal and pairwise marginal posterior distributions of nine parameters over Event 1 by using GLUE; (b) the qualitative representation of Sobol' sensitivity analysis for the nine parameters based on the likelihood function (L). In (a), the red nodes are selected as the predefined, "true" parameter values, which will be used in the synthetic experiment. In (b), the diameter of the nodes around the circle is proportional to the total-order sensitivity, and the width and opacity of the lines connecting the nodes are proportional to the pairwise interaction sensitivity.

4.4.3 Data assimilation of the synthetic experiment

Synthetic experiments are often employed to examine whether parameters converge satisfactorily, whether the range of parameters is adequately quantified, and whether predictive uncertainty is minimized [*Moradkhani*, 2008]. First, we ensured that the model parameters updated from dual data assimilation can converge to the predefined parameter values for Event 1. Fig. 4.7 shows the time evolution of posterior distributions of nine parameters for five dual filters. The most easily identifiable parameters are CQOF and CK12, while the rest of the parameters could not reduce the large uncertainty range over time, as also inferred from the results of GLUE and sensitivity analysis in Section 4.4.2. All dual filters except for Dual InSuWF successfully provided the posterior distribution of these parameters that almost converged to the predefined values at the end stage of assimilation.

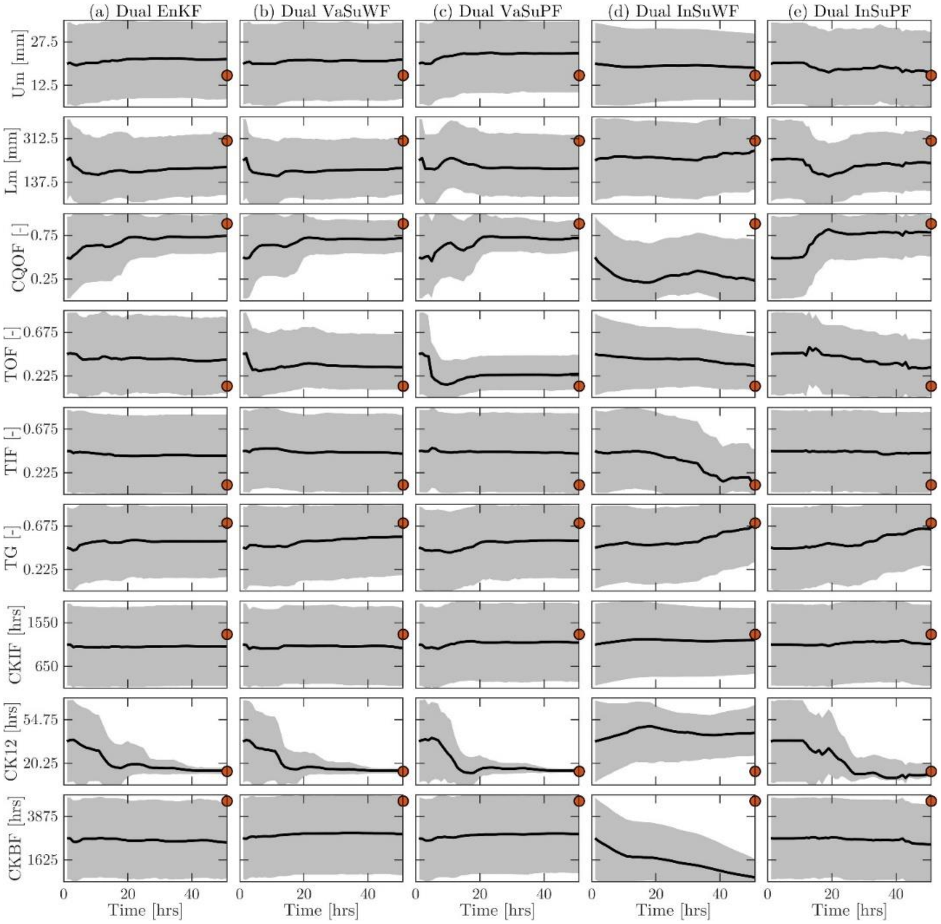


Figure 4.7. The time evolution of posterior distributions of nine parameters for five dual filters: (a) Dual EnKF, (b) Dual InSuWF, (c) Dual InSuPF, (d) Dual VaSuWF, and (e) Dual VaSuPF. Data assimilation with the ensemble size of 500 was performed for the synthetic experiment over Event 1. Shaded areas and black lines represent the 90% confidence intervals and the mean values of ensemble parameters, respectively. Red nodes refer to the predefined value of parameters in Table 4.1.

Compared to the tendency of Dual EnKF to converge to the predefined parameters, the dual variant filters (Dual VaSuWF and Dual VaSuPF) can accurately update the posterior parameters in terms of convergence speed and degree. On the other hand, the dual invariant partial filter (Dual InSuPF) provides a slightly different converging tendency wherein the magnitude of the uncertainty of CK12 is larger than Dual EnKF, and the convergence is slower. Specifically, the mean value of ensemble members of CK12 (black line in Fig. 4.7) converges to its predefined value at about 20 hours in Dual EnKF (as well as both Dual Variant filters), but it takes an additional 7 hours in Dual InSuPF. The other dual invariant whole filter (Dual InSuWF) completely fails to estimate the parameter posterior distribution, because the influential parameters converge to a lesser extent and the identified posterior distributions do not converge to the predefined values (Fig. 4.7d).

Ensemble streamflow predictions and their error measures were compared for ten filters in the synthetic experiment over Event 1 in Fig. 4.8. (1) Deterministic and stochastic error measure values in Figs. 4.8b to 4.8d indicate that all surrogate filters are functioning properly to improve the accuracy of streamflow predictions. The results of NSE, PE, BS, $\overline{CRPS}$, and $Spread$ when using DA are significantly better than those without DA. (2) The single and dual surrogate filters were compared, clearly demonstrating the effect of simultaneously updating the states and parameters on the accuracy. All dual filters, except for Dual InSuWF, produce superior results to single filters and provide almost the same results as the original Dual EnKF. For example, the performance of four dual filters increases by at least about 8, 65, 48, 68, and 51% for the ensemble median of NSE ($\widetilde{NSE}$), the ensemble median of PE ($\widetilde{PE}$), BS, $\overline{CRPS}$, and $Spread$, respectively. The interquartile uncertainty range of NSE is also reduced to 0.02 – 0.07 in dual filters from 0.03

– 0.39 in single filters, and that of PE is reduced to 2 – 10% in dual filters from 10 – 30% in single filters. (3) Dual InSuWF showed no performance improvement for NSE and PE compared to InSuWF, while there were improvements of about 36 and 22% for $\overline{CRPS}$ and $Spread$, respectively. Dual InSuWF is less accurate than Dual EnKF, while InSuWF has more accurate prediction performance than EnKF.

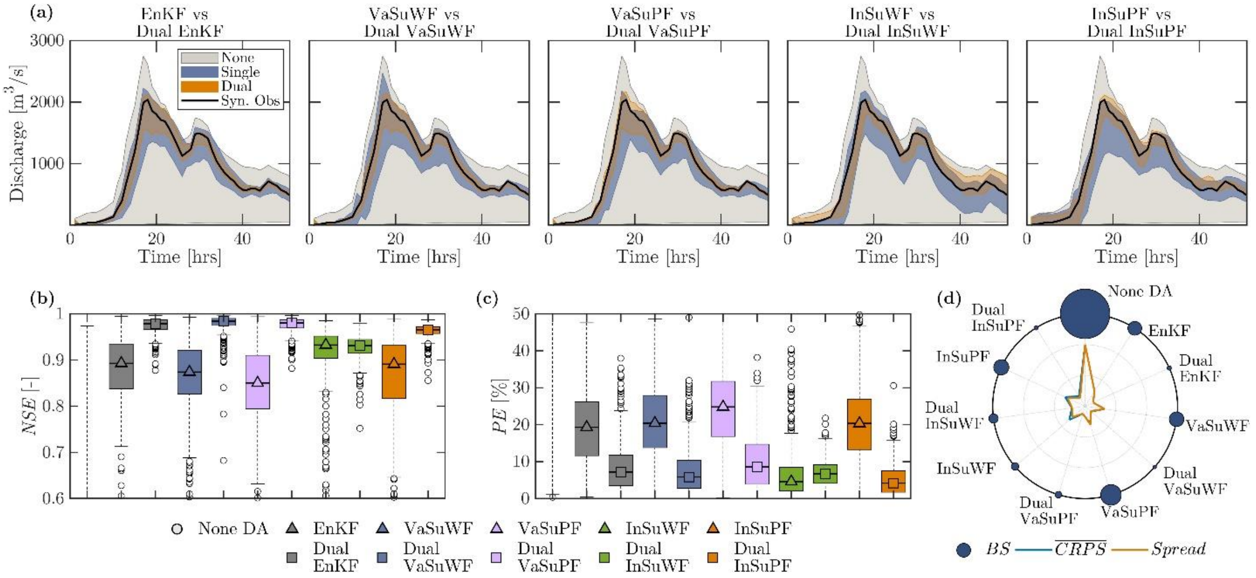


Figure 4.8. (a) Comparisons of 500 ensemble streamflow predictions with 90% confidence intervals, and (b to d) comparisons of the accuracy metrics for ten filters in the synthetic experiment over Event 1 for (b) NSE, (c) PE, and (d) BS, $\overline{CRPS}$, and $Spread$. In each boxplot, the central mark is the median, the edges of the box are the 25th and 75th percentiles, and the upper and lower whiskers are the maximum and minimum except for outliers (dot symbols). In (d), the circle plots qualitatively represent probabilistic measures for BS (dark blue nodes), $\overline{CRPS}$ (blue lines), and $Spread$ (yellow lines). The size of nodes corresponds to the magnitude of BS; the distance from points on the blue and yellow lines to the center of the circle corresponds to the magnitude of $\overline{CRPS}$ and $Spread$.

In summary, the analyses above in the synthetic experiment indicate that all of the four single filters (InSuWF, InSuPF, VaSuWF, and VaSuPF) worked similarly to the original single filter, EnKF, in terms of accuracy for streamflow forecasting. Furthermore, three of the four dual filters (Dual InSuPF, Dual VaSuWF, and Dual VaSuPF) showed equivalent performance to the original dual filter, Dual EnKF in terms of increased accuracy and parameter posterior estimation. However, Dual InSuWF failed to converge to predefined parameters during the assimilation process, nor did it improve predictive performance.

4.4.4 Data assimilation of the real experiment

In this section, two data assimilation experiments using real rainfall and streamflow observations were conducted to further examine the performance of proposed filters in a real-time forecasting framework. The forecasting results over Events 2 and 3 are reported from Figs. 4.9 to 4.12. In general, the results of the real data assimilation experiment have equivalent conclusions to those of the synthetic experiment. Qualitative inspections for observation consistency and uncertainty interval from Figs. 4.9 and 4.10 reveal that all surrogate filters provided similar results to EnKFs, and dual filters are more accurate and have narrower uncertain ranges than single filters.

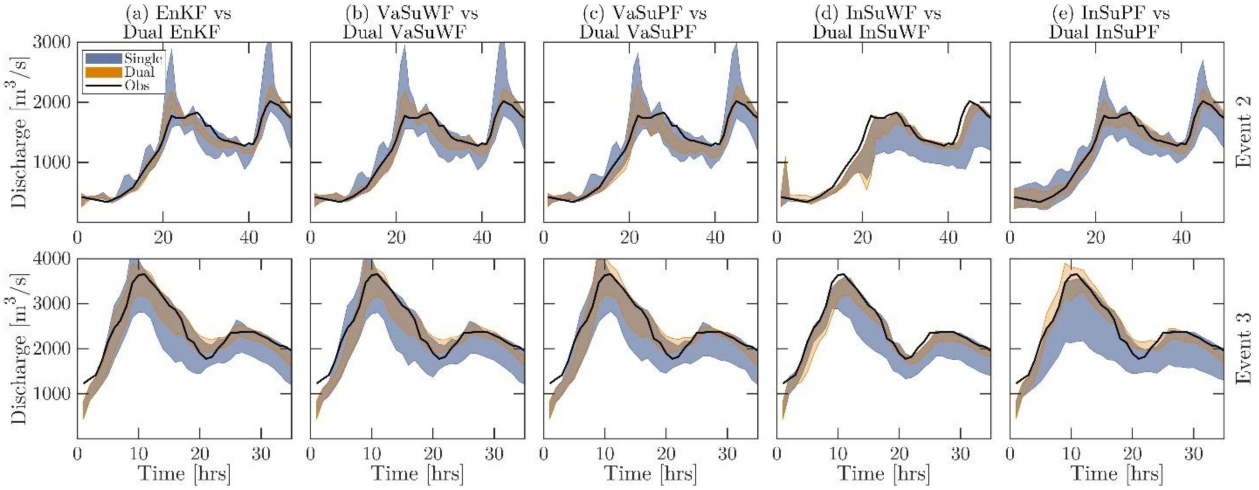


Figure 4.9. Comparisons of 500 ensemble streamflow predictions with 90% confidence intervals for ten filters in the real data assimilation experiments over (top) Event 2 and (bottom) Event 3.

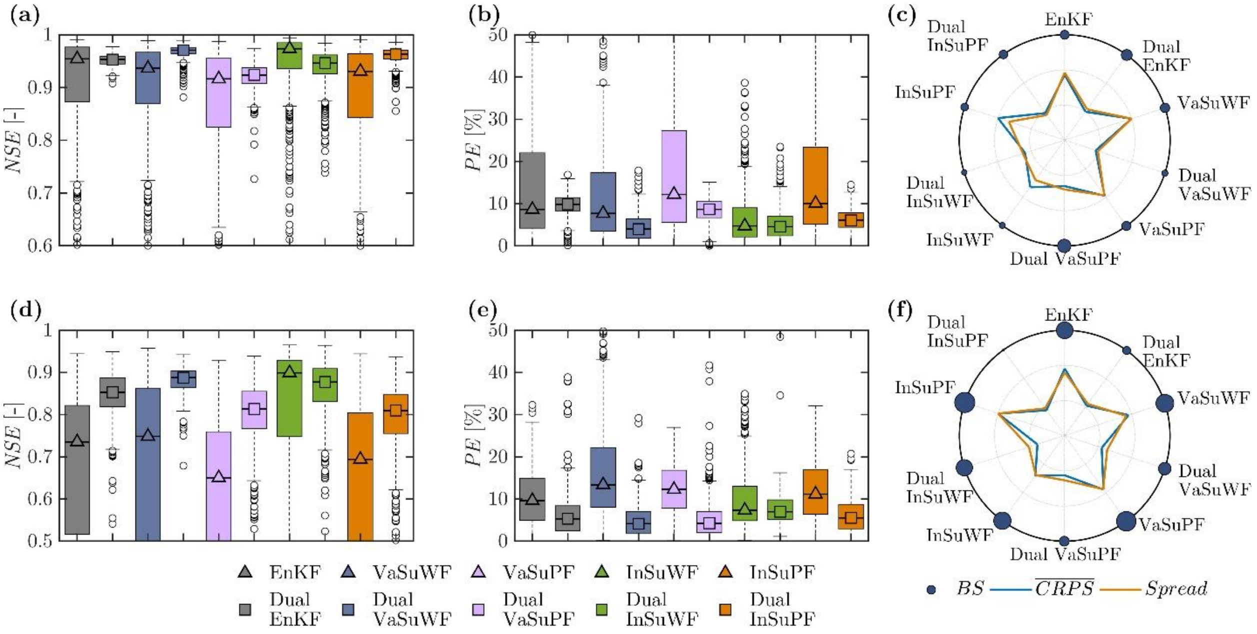


Figure 4.10. Comparisons of the accuracy metrics for ten filters with an ensemble size of 500 in the real data assimilation experiments over (a to c) Event 2 and (d to f) Event 3: (a,d) NSE, (b,e) PE, and (c,f) BS, $\overline{CRPS}$, and $Spread$. The boxplots demonstrate the median (central mark), the 25th and 75th percentiles (the edges of the box), and the maximum and minimum (the upper and lower whiskers) except for outliers (dot symbols). The circle plots qualitatively represent probabilistic measures for BS (dark blue nodes), $\overline{CRPS}$ (blue lines), and $Spread$ (yellow lines). The size of nodes corresponds to the magnitude of BS; the distance from points on the blue and yellow lines to the center of the circle corresponds to the magnitude of $\overline{CRPS}$ and $Spread$.

Since it is evident that streamflow predictions were improved by using data assimilation, we omitted the comparison for the absence of DA and showed the comparison between filters regarding the reliability of surrogate filters. Three paired comparisons were performed based on three standards of surrogate filter construction described in Section 4.2.2: Whole versus Partial, Variant versus Invariant, and Single versus Dual. Such comparisons can be done easily with the help of a relative 'difference' metric (Δ) between the values of the evaluation metrics (Metric), including $\widetilde{NSE}$, $\widetilde{PE}$, BS, $\overline{CRPS}$, and $Spread$. This difference metric is defined in the unit of percentage as:

$$\Delta = \frac{\left|\mathrm{Metric(SuFs1)} - \mathrm{Metric}_{\mathrm{ideal}}\right| - \left|\mathrm{Metric(SuFs2)} - \mathrm{Metric}_{\mathrm{ideal}}\right|}{\left|\mathrm{Metric(SuFs1)} - \mathrm{Metric}_{\mathrm{ideal}}\right|} \times 100 \qquad (4.37)$$

where SuFs1 denotes the former group filters, i.e., whole, variant, and single filters, while SuFs2 denotes the latter filters, i.e., partial, invariant, and dual filters. $\mathrm{Metric}_{\mathrm{ideal}}$ represents the ideal (perfect) values of the metrics of NSE, PE, BS, $\overline{CRPS}$, and $Spread$, that is 1, 0, 0, 0, and 0, respectively. The positive (or negative) values of Δ indicate that the prediction results of the latter group filters are more (or less) accurate than those computed by the former filters.

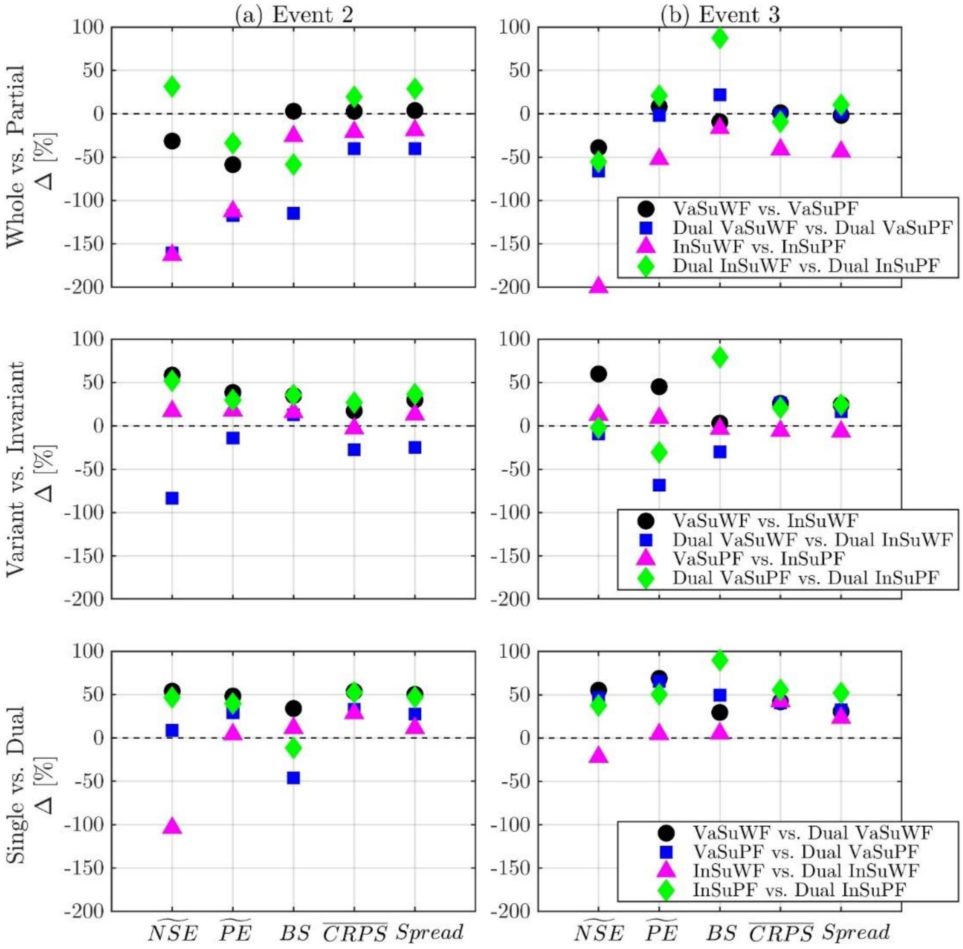


Figure 4.11. Three paired comparisons of a relative 'difference' metric (Δ) in Eq. (4.37) for the five evaluation metrics (*x*-axis) over (a) Event 2 and (b) Event 3. These comparison pairs are based on three standards of surrogate filter construction described in Section 4.2.2. The positive (negative) values of Δ indicate that the prediction results of Partial, Invariant, and Dual filters are more (less) accurate than those computed by Whole, Variant, and Single filters, respectively.

The results of Δ for the three paired comparisons are illustrated in Fig. 4.11 and also reported in detail in Table C.1 (Appendix C). First, the results of Δ between whole and partial filters are mostly negative, revealing that whole filters outperform partial filters by up to 203, 117, 115, 41, and 44% for $\widetilde{NSE}$, $\widetilde{PE}$, BS, $\overline{CRPS}$, and $Spread$, respectively, over both Events 2 and 3. The only exception for this tendency can be found in Dual InSuWF, which does not show an obvious superiority over Dual InSuPF. As an example for Event 2, $\widetilde{NSE}$, $\overline{CRPS}$, and $Spread$ of Dual InSuWF are inferior (i.e., positive Δ values) to those of Dual InSuPF by about 31, 19, and 29%, respectively, while $\widetilde{PE}$ and BS are improved (i.e., negative Δ values) by about 34 and 59%, respectively (Fig. 4.11). Regarding the second paired comparisons of building systems between variant and invariant filters, three out of four invariant filters have better performance than the corresponding variant filters (Fig. 4.11). In particular, the values of Δ for all metrics have positive values ranging up to 60, 45, 27, 79, and 37% for $\widetilde{NSE}$, $\widetilde{PE}$, BS, $\overline{CRPS}$, and $Spread$, respectively. Conversely, for the forecasting results of the remaining invariant filter, Dual InSuWF (i.e., blue in Fig. 4.11), it is hard to conclude which building systems are superior. The performance results are mixed depending on the metrics, e.g., results of $\widetilde{PE}$ are 68% worse than those in Dual VaSuWF while those of $\overline{CRPS}$ are 27% better for Event 3. Third, convincing evidence was found that dual filters outperform single filters. As also seen in Section 4.4.3, the dual updates of parameter and state significantly enhance the forecasting results and narrow their uncertainty spreads. Quantitatively, the 'difference' metric results of dual filters are improved by up to 9, 48, 34, 54, and 50% over Event 2 for $\widetilde{NSE}$, $\widetilde{PE}$, BS, $\overline{CRPS}$, and $Spread$, respectively. For Event 3, these improvements are more substantial, with improvements of up to 38, 69, 90, 56, and 52%, respectively (Fig. 4.11). In summary, the Dual VaSuWF and Dual InSuPF filters have proven to be superior to the others in providing accurate forecasting results, followed by Dual VaSuPF, with prediction results closest to the above two filters.

The above comparisons were made for prediction results for short lead times (LT) of 1 hour. As predictions for greater lead times are usually in demand, additional analysis was performed to examine whether the surrogate filters can provide reliable and accurate streamflow predictions for larger lead times of 1 to 6 hours. As expected, the forecasting performance for ten filters decreases with the lead time – decreasing for all metrics. Such a tendency is clearly shown in Fig. C.1 (in Appendix C), where the evaluation metrics at each lead time are compared with those at the lead time of 1 hour for ten filters. Specifically, in Event 2, the ranges of degradation of $\widetilde{NSE}$, $\widetilde{PE}$, BS, $\overline{CRPS}$, and $Spread$ at a lead time of 6 hours compared to those at lead time of 1 hour are 181-1430, 0-469, 0-853, 51-154, and 46-152%, respectively. In Event 3, these ranges are 235-1206, 105-395, 115-585, 51-203, and 66-196%. Interestingly, some filters (all single filters and Dual InSuWF) have much worse predictability with respect to lead time, while all dual filters except for Dual InSuWF are not as good, but better than the former filters.

In order to compare the degree of performance deterioration for lead time among 10 filters, another relative 'difference' metric (Γ) is computed as:

$$\Gamma = \frac{|\mathrm{Metric_{best}} - \mathrm{Metric_{ideal}}| - |\mathrm{Metric}(\mathrm{SuFs}, LT) - \mathrm{Metric_{ideal}}|}{|\mathrm{Metric_{best}} - \mathrm{Metric_{ideal}}|} \times 100 \qquad (4.38)$$

where $\mathrm{Metric}(\mathrm{SuFs}, LT)$ denotes the evaluation metric of a surrogate filter at a lead time (LT) that is varied from 1 to 6 hours. $\mathrm{Metric_{best}}$ represents the best of the 60 values (corresponding to 10 filters × 6 different lead times) closest to the $\mathrm{Metric_{ideal}}$. The negative values of Γ indicate the degree of performance deterioration as compared to the best value.

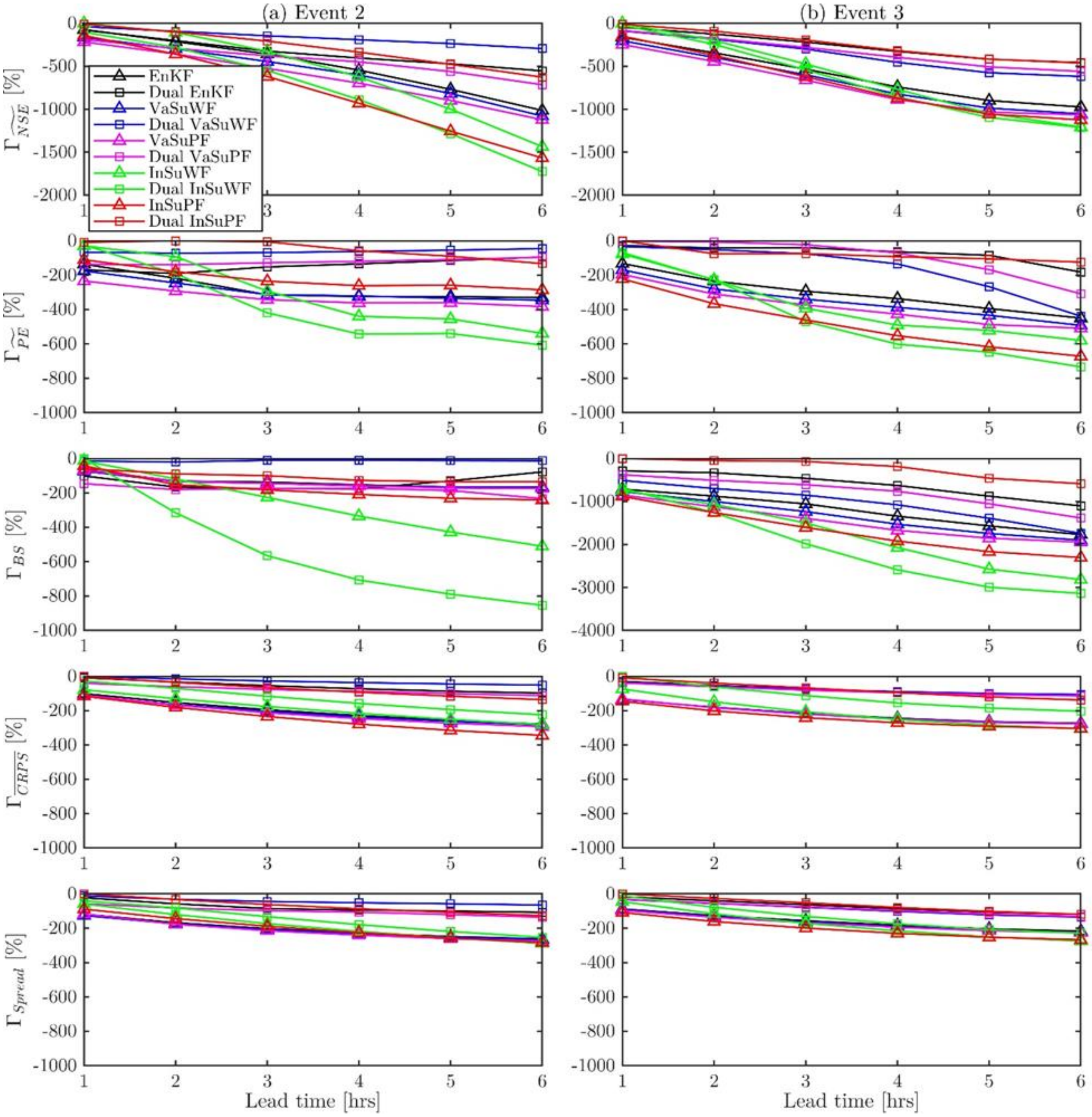


Figure 4.12. The comparisons of a relative 'difference' metric (Γ) in Eq. (4.38) for the five evaluation metrics (y-axis) with different lead times from 1 to 6 hours (*x*-axis) over (a) Event 2 and (b) Event 3. These values in each subplot were compared for the best of 60 values (10 filters × 6 lead times) closest to the ideal value of each evaluation metric. The negative values of Γ indicate the degree of performance deterioration as compared to the best value.

Compared to the best performance of $\widetilde{NSE}$, its performance degradation at the longest lead time of 6 hours stretches from 293, 714, and 626% in Dual VaSuWF, Dual VaSuPF and Dual

InSuPF, respectively, up to 1724 % in Dual InSuWF. Such a degradation is highest in $\widetilde{NSE}$, and then in BS and $\widetilde{PE}$ (Fig. 4.12). Another interesting phenomenon is that the performance between the 10 filters is not very different for a lead time of 1 hour, but the performance difference between the filters increases significantly as the lead time increases. For example, the performance differences of Γ at 1 hour lead time are about 216, 233, 146, 115, and 126% for $\widetilde{NSE}$, $\widetilde{PE}$, BS, $\overline{CRPS}$, and $Spread$, respectively. At a lead time of 6 hours, the differences extend up to 1724, 606, 853, 343, and 288% (Fig. 4.12a). Filters that perform well at longer lead times are Dual VaSuWF, Dual InSuPF, and Dual VaSuPF, which means that filters that performed better at a lead time of 1 hour outperform other surrogate filters at longer lead times.

4.4.5 Evaluation of the superiority of computational performance to EnKF

Fig. 4.13 demonstrates the superiority of SuFs to EnKFs in terms of efficiency by computing the cumulative and instantaneous runtime ($RT_{cum,t}$ and RT_t) of ten filters for Event 2. From Fig. 4.13a displaying the cumulative runtime versus time for n of 500, it can be seen that the calculation speed of InSuFs is much faster than that of EnKFs, whereas the speed of VaSuFs is slightly faster. For example, at the end of forecasting ($t = 50$), the best filter, Dual InSuPF is about 500 times more efficient than Dual EnKF. Since the building time ($RT_{build,t}$) of both of these two filters is equal to zero, the difference of 500 times is the same as the difference in the running time ($RT_{run,t}$) of the two filters (i.e., about 4.6×10^{-4} and 0.23 secs; see the slope of Eq. (4.32) written in the legend of Fig. 4.13b). Comparing the runtime of VaSuFs to EnKFs is the case when the additional runtime required for building the filters ($RT_{build,t}$) offsets the efficiency of the runtime in running the filters ($RT_{run,t}$). When is small, the efficiency improvement of Dual VaSuPF and Dual VaSuWF to Dual EnKF is relatively low (e.g., only 4 and 6 times faster for $n = 500$), but as

n gets larger, this improvement becomes much greater (e.g., 65 and 100 times faster for $n = 10{,}000$).

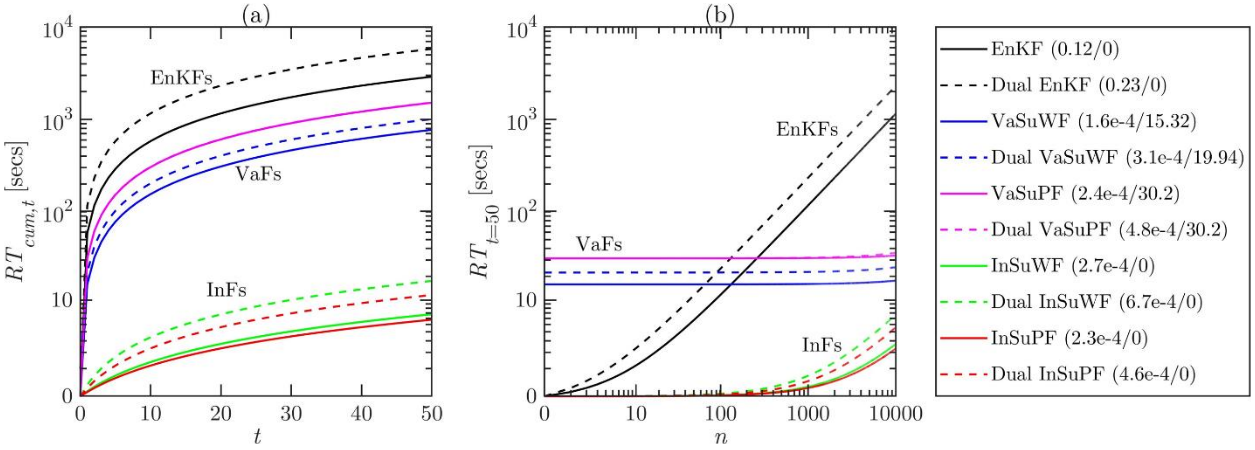


Figure 4.13. Comparisons of runtime for ten filters: (a) the cumulative runtime ($RT_{cum,t}$) with respect to forecasting timestep (t) and (b) the runtime at the end of forecasting ($RT_{t=50}$) with respect to the number of ensemble (n). In (a), $RT_{cum,t}$ is computed over the fixed ensemble size of 500 while in (b), $RT_{t=50}$ is computed for the fixed time of 50. $RT_{t=50}$ shows a linear relationship with the ensemble size; its slope ($RT_{run,t}$ in Eq. (4.32)) and intercept ($RT_{build,t}$) for all filters are written in the parentheses on the legend. Note that $RT_{build,t}$ for the invariant filters (RT_{build}^{In}) are set to be zero to compare only the time required for forecasting, because these filters can be built prior to forecasting. The results of Event 2 were used.

4.5 Discussions

4.5.1 Is a Partial surrogate approach more promising?

The primitive rationale of the Partial approach was to individually replace the time-consuming processes (e.g., Eqs. 4.1, 4.3, and 4.5) in the original filter (model). The rest of the EnKF processes (Eqs. 4.4 and 4.6), which take less time but play an important role, remain the same. Since Kalman gain (K) in that Partial filter was directly calculated and reflected in updating the parameters and states, it was possible to present more accurate results than the conventional Whole approach that blackboxed this process. Thus, a remarkable question to be addressed is how to select the process to be replaced among processes included in original filters (models) when designing a surrogate filter (model). Considering the trade-off between the time taken to execute the process and its physical importance, a flexible surrogate model design will be possible.

What are the central advantages of the Partial approach in terms of efficiency, other than the aforementioned accuracy improvement? The Partial approach can effectively reduce the number of dimensions of the PCE input. In this study, the total number of dimensions, N_X was $N_S + N_P + N_u$ in the Partial filter, while $N_S + N_P + 2N_u + N_{obs}$ (the number of observations) in the Whole approach. The number of dimensions is reduced by $N_u + N_{obs}$ from 17 (Whole) to 15 (Partial). Although the decrease in the number of dimensions, 2, may seem small, its contribution from the perspective of PCE coefficients is by no means small. That is, the number of PCE coefficients decreases significantly from 26,334 to 15,504 (~ 40 % reduction) estimated by the N_X and a common p of 5 from Eq. (2.3), thus resulting in the smaller size of experimental design (e.g., from 90,000 to 40,000 for building Dual InSuFs; see Table 4.2). If forcings and observations with different values for space are considered (i.e., if N_u and N_{obs} are not equal to 1), such a reduction effect by the Partial approach will be even greater.

How can the Partial approach be extended to a fully distributed model with much larger dimensions rather than a lumped model? In this case, the total number of dimensions is as large as the total dimensions of the lumped model multiplied by the number of computational cells (N_{cell}). That is, $N_X = (N_S + N_P + 2N_u + N_{obs}) \times N_{cell}$ in the Whole approach. Note that such a number is incredibly too high. A pragmatic solution by the Partial approach is to create independent surrogate PCEs as many as N_{cell}. It is enabled if each cell is treated as a separate process and thus is superseded with an independent PCE. This ultimately has the effect of turning the problem of generating one PCE with the entire dimension into a problem of generating several (N_{cell}) PCEs with dimensions of N_X/N_{cell}. Even more productive solution is enabled by combining this Partial approach with Karhunen-Loève (KL) decomposition [*Karhunen*, 1946], where it can lump cells with high correlation between model outputs in space (or time) into one group. Once the spatially-correlated groups are identified, the Partial approach is then employed to construct a PCE for that group cells. Surely, results simulated by the constructed PCE can be pertained only to the portion of domain determined previously.

4.5.2 Is building a 'universal' PCE achievable and consequential?

A well-known characteristic and challenge of data-driven models is that they cannot be applied to domains outside the scope of trained data. Basically, data-driven models have only one unique model suitable for each training data space. Likewise, PCE also has one optimal model for each data set (experimental design). If new data needs to be taken into account for future forecasting, as with most studies in the past [*Sargsyan et al.*, 2014; *Bazargan et al.*, 2015; *Wang et al.*, 2018; *Dwelle et al.*, 2019; *Hu et al.*, 2019a; *Tran and Kim*, 2019; *Zhang et al.*, 2020], it is natural to create a new PCE model like the time-variant approach. However, this approach cannot be applied when measurement (or forecasted) data is scarce, and even if the data is sufficient, there is a very critical disadvantage that a time-consuming operation must be repeated whenever data is altered. One of the goals of this study was to determine whether a 'universal' invariant PCE that could be applied to a wide range of rainfall events could be generated.

It is apparent that the higher the degree of generalization of rainfall events represented by the experimental design, the more the PCE can be applied to various conditions. An effortless way to expand the scope of the data space (i.e., the Invariant approach) was proposed and verified based on the results of the synthetic and real experiments. A 'universal' surrogate filter using the

Invariant approach (Dual InSuPF) is the most efficient filter and one of the three surrogate filters that provide the most reliable predictability. These results confirm that a single ‘universal’ PCE could be constructed across a wide range of random data space (e.g., for rainfall and streamflow) and applied to new input space. This idea of using a ‘random input generator’ allows for creating as many hypothetical events as possible that can happen in that region. This does not require any specific historical event data, so it has the advantage of being easily applied even if there is no historical data.

What are the practical implications of the universal Invariant filter? While many studies have highlighted the benefits of PCE in saving computational costs, most studies have been limited to hindcasting based on historical data (e.g., focusing on uncertainty quantification or sensitivity analysis) [*Wu et al.,* 2014; *Meng and Li,* 2018; *Miller et al.,* 2018; *Dwelle et al.,* 2019]. To the best of our knowledge, no surrogate model (or filter) was applied to real-time flood forecasting because there is a downside of having insufficient time to rebuild a new PCE at each computation step for inputs provided in real time. Therefore, the idea of making a unique surrogate filter during non-flood season can bridge the gap between hindcasting and real-time forecasting.

4.5.3 Is an advanced SED-PD necessary in constructing PCE?

In this dissertation, we propose a SED-PD scheme, an advanced version of SED introduced by *Blatman and Sudret* [2010], that was employed to determine the optimal values of N and p in estimating PCE coefficients. As a result of investigating the accuracy errors below and in the literature [*Hu and Youn,* 2010; *Sargsyan et al.*, 2014; *Diaz et al.*, 2018; *Dwelle et al.*, 2019; *Torre et al.*, 2019], we found that it is inappropriate to apply the SED originally developed for finite element problems directly to hydrologic problems of interest. Here, we underscore the limitations of SED and the necessities of SED-PD.

Fig. 4.14 clearly shows the difference between the hypothetical results obtained using SED and those from SED-PD, in terms of evolution of the error over runtime. The results of SED were derived specifically for the VaPCE1 and InPCE1 constructions of QoI = y; the p values were fixed from 1 to 7, and the maximum iterations for N ($l_{N,max}$) were limited to 1000 and 100 for VaPCE1 and InPCE1, respectively; there was a single stopping criterion (the first criterion) where the lower threshold ϵ_{th}^{lower} is 10^{-5}. Interestingly, in both VaPCE1 and InPCE1, $\epsilon_{LOO}^{\mathrm{QoI}}$ could not reach to the threshold value, which implies that one cannot create any PCE that satisfies the desired condition within a limited time when using SED. Apparently, such an optimization using SED should be amended.

Three possibilities are discussed for why SED-PD is necessary in optimizing the hyper-parameters (N and p). First, the use of a single criterion cannot guarantee convergence if the target error is not reached. Errors computed are no longer reduced even if a large number of iterations (e.g., up to 1000 and 100 of $l_{N,max}$, equivalently 10000 and 1188100 of N, for VaPCE1 and InPCE1 in Fig. 4.14a and 4.14c) have been implemented. Compared to the SED-PD case, SED has not been able to obtain the desired convergence even after spending an enormous amount of time, or has to invest an almost infinite amount of time until convergence (see SED vs. SED-PD comparisons when stopping iterations: $RT_{build,t=50}^{Va,QoI=y}$ are infinite vs. 7.92 secs for VaPCE1 in Fig. 4.14a, and $RT_{build}^{In,QoI=y}$ are infinite vs. 2.81×10^3 secs for InPCE1 in Fig. 4.14c). On the other hand, SED-PD can determine the optimal N and p quickly through the four criteria proposed, ensuring system convergence and computational stability.

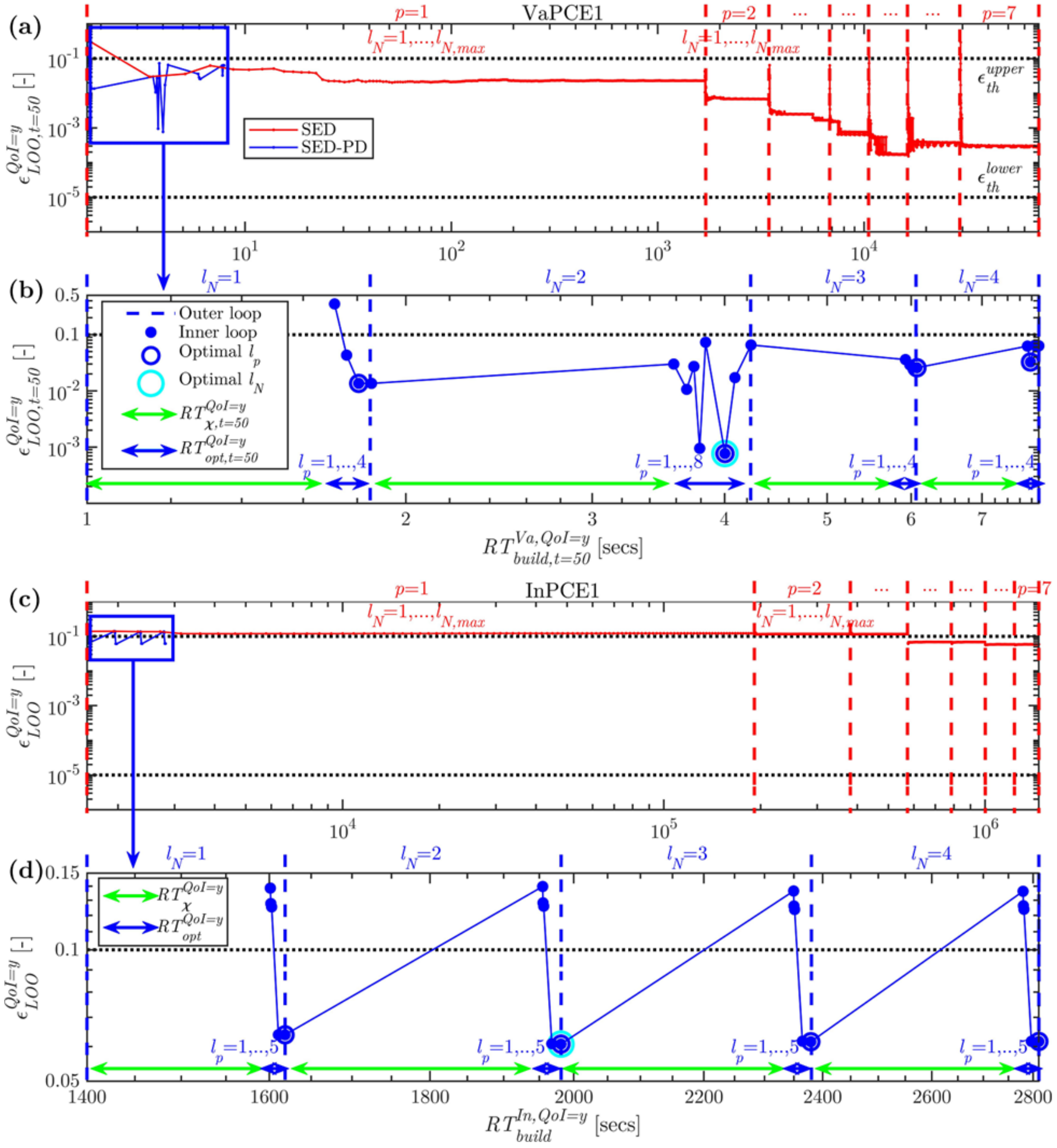


Figure 4.14. Illustrations of building runtime ($RT^{Va,QoI=y}_{build,t=50}$, $RT^{In,QoI=y}_{build}$) versus error ($\epsilon^{QoI=y}_{LOO,t=50}$, $\epsilon^{QoI=y}_{LOO}$) under SED and SED-PD in constructing PCE for QoI = y (streamflow) for (a, b) VaPCE1 and (c, d) InPCE1. In SED, the polynomial degree p was fixed at values, in turn, from 1 up to 7; iterations for N were performed up to $l_{N,max}$ of 1000 for VaPCE1 and 100 for InPCE1. In SED-PD, zoomed-in subplots b and d, the blue dashed line separates each iteration of the outer loop (l_N); the solid blue point separates the iteration of the inner loop (l_p). The blue empty circle marks the optimal p value in each inner loop; and the cyan empty circle marks the optimal N and p in constructing PCE. The green double arrow denotes the runtime needed for attaining the experimental design ($RT^{\mathrm{QoI}}_{\boldsymbol{\chi}}$), while the blue double arrow represents the runtime to implement SED-PD (RT^{QoI}_{opt}). For VaPCE1, the results at $t = 50$ are shown for Event 2.

The second possibility is related to whether or not to include the optimization process of the p value in SED. Since the optimal value of p is unknown, a common way to reveal an acceptable N value in SED is to start from p = 1 and increase the value by 1 until the optimal N is determined (as in Fig. 4.14a and 4.14c). Or, if one can assume an appropriate p value, the optimal N for that random p value can be determined. Since the optimal p value that causes the smallest error can vary from case to case (for example, optimal p is determined as 6 when $l_N = 2$ for VaPCE1 and as 5 when $l_N = 2$ for InPCE1; see the cyan empty circles in Fig. 4.14b and 4.14d), SED always requires additional analysis (of similar form to Fig. 4.14a and 4.14c). However, in SED-PD, dual optimizations for both p and N are adopted such that the optimal value of p is automatically identified at each iteration l_N (like the blue empty circles in Fig. 4.14b and 4.14d). Such a dual optimization system ultimately improves the existing approach of SED by which the p value had to be selected ad-hoc or by trial and error.

Note that the performance of SED is highly influenced by the PCE types and the lower threshold, ϵ_{th}^{lower} (see Fig. 4.14). In particular, the success or failure of the SED optimization process depends on the latter threshold value. Therefore, one might wonder how the optimization result will change if a larger value is chosen for ϵ_{th}^{lower} of SED. In this regard, Fig. 4.15 demonstrates the effects of the lower threshold, ϵ_{th}^{lower} on the runtime for building the PCE: the larger the threshold value, the sooner SED can stop and the higher the probability of attaining an optimal N value. The empty red squares in Fig. 4.15 indicate that there is no probability to get its optimum that satisfies the criterion within the maximum number of iterations given p. In order to avoid SED failure, a feasible threshold greater than the minimum value of $\epsilon_{LOO}^{\text{QoI}}$, say about 1.5 × 10^{-4} for VaPCE1 and about 5.5 × 10^{-2} for InPCE1, must be selected in advance. Then, a question arises of how to pre-determine the value of ϵ_{th}^{lower} for each PCE construction. A practical and general answer to this question is to perform an additional analysis similar to Fig. 4.15 that uses trial and error with different thresholds – this analysis is however unnecessary in SED-PD.

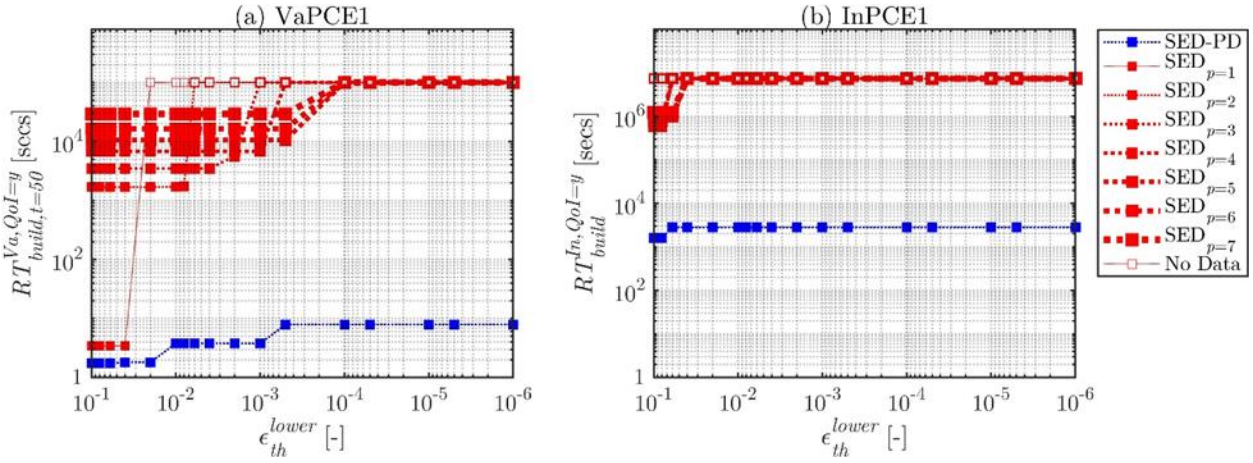


Figure 4.15. Effects of the lower threshold, ϵ_{th}^{lower}, on the runtime for building (a) VaPCE1 and (b) InPCE1 for QoI = y (streamflow) when using SED and SED-PD. The empty red squares indicate that there is no optimal N value that satisfies the criterion within the maximum number of iterations given p. For VaPCE1, the results at $t = 50$ are shown for Event 2.

4.5.4 Is the surrogate filter broadly applicable to geophysical science?

Surrogate filter approaches proposed can be applied to various geophysical fields that require data assimilation to improve the accuracy and efficiency of real-time predictions. All DA techniques generally consist of ("prediction step") predicting the values of current state variables given information at the previous time step, and ("update (analysis) step") updating the predictands by calculating the error between the predicted values and the currently observed values. Although the DA techniques differ in how they calculate and analyze the error in detail, the fact that states, parameters, and forcings are transited to the predictions through a propagator (e.g., Eqs. (4.1), (4.3), and (4.5)) is identical. Since the Partial filter proposed replaces these equations, it can be applied seamlessly to other DA techniques without much modification. Obviously, the Whole filter can be generated only with evaluation results of the original filter, so there is no limitation in applying it to other DA applications. Moreover, these approaches are all based on any geophysical model that should be used for the prediction and update. Thus, the more complex governing equations the model contains, the more the computational effect using the surrogate filter will be maximized.

4.6 Conclusions

The main objectives of this Chapter are: (1) to present a robust and efficient data assimilation technique in the framework of hydrologic flood forecasting, embracing the merits of the ensemble Kalman filter (EnKF) and polynomial chaos expansion (PCE) in order to produce reliable streamflow predictions with significantly reduced runtime; (2) to underscore the advantages of the novel partial and invariant approaches in making a surrogate filter, by investigating the accuracy and efficiency of the eight surrogate filters categorized according to different surrogate structures (whole and partial), building systems (variant and invariant), and assimilating targets (single and dual); (3) to propose an advanced dual optimization system with multiple stopping criteria, named *sequential experimental design-polynomial degree* (SED-PD), that simultaneously determines the hyper-parameters of N and p necessary for the PCE construction. The following are the principal results and conclusions of this study.

- The SED-PD scheme has evolved into a dual optimizing system and requires four stopping criteria, dealing with two issues that originally occurred in SED during PCE construction. In particular, the inherent assumption of SED that its accuracy error should decrease monotonically with iterations is not always satisfied. Thus, the multiple criteria were needed to ensure convergence of the optimization process and avoid the possibility of infinite iterations. Additionally, the exclusion of polynomial degree from the optimization process leads to a practical issue, wherein the value of polynomial degree had to be selected ad-hoc or by trial and error. The dual optimization system proposed resolves this existing issue of SED.

- A comprehensive investigation into how to configure a surrogate filter has been carried out. Conventionally, the Whole (replacing entire processes of the original filter) and Variant (requiring reconstruction at each time step) approaches have been employed. However, we have confirmed that this traditional approach deteriorates forecasting performance in terms of accuracy and efficiency. A novel Partial (replacing part of the original filter) and Invariant (valid for whole time periods) approach is proposed for the filter construction, which outperforms the conventional approach. The Partial approach can directly reduce the number of dimensions by turning the problem of generating one PCE with the entire dimension into a problem of generating several PCEs with a reduced dimension. The Invariant approach making a unique surrogate filter during non-flood season can bridge the gap between hindcasting and real-time forecasting.
- Specific results from the synthetic and real data assimilation experiments are (1) Dual SuFs except for Dual InSuWF successfully mimic the convergence characteristics of Dual EnKF in updating model parameters; (2) The comparing results of eight surrogate filters show that Dual VaSuWF, Dual VaSuPF, and Dual InSuPF illustrate the most superior performance, equivalent to that of Dual EnKF; (3) These three filters perform relatively well at longer lead times as well, although forecasting performance decreases with lead time for all filters.

- Regarding efficiency, the use of surrogate filters dramatically improves the computational performance. In particular, Dual VaSuWF, Dual VaSuPF, and Dual InSuPF are about 6, 4, and 500 times faster than Dual EnKF, respectively, (when comparing the cumulative runtime ($RT_{cum,t}$) over event 2 with the ensemble size of 500). This efficiency gain is more pronounced when original filters being replaced are time-consuming or larger ensemble sizes are employed. Since the calculation speed of the generated PCE is related to the time of the arithmetic operation level, it is always fast regardless of how complicated and time-consuming the original filter is.
- Based on in-depth analyses, the Dual Invariant Partial filter (i.e., Dual InSuPF) is the best one, being superior in terms of usefulness, effectiveness, and robustness as an EnKF replacement. Therefore, the proposed surrogate filter will be a promising alternative tool for performing computationally-intensive data assimilation in high-dimensional problems. Ultimately, it not only provides equivalently accurate forecasting results in real time, but also significantly reduces the computational burden of larger ensemble predictions.

CHAPTER V

A new surrogate model enables predictions for extreme events that deviate significantly from the training dataset

> "The only way of finding the limits of the possible is by going beyond them into the impossible"
>
> - *(Clarke, AC)*

5.1 Introduction

Floods have long been studied in practical and scientific fields because they cause severe damage to the environment and societies around the world [*Hirabayashi et al.*, 2013; *Ward et al.*, 2013]. Timely and accurate predictions of extreme flood events play a pivotal role in decision-making processes to mitigate risk [*Ward et al.*, 2013; *Sanders et al.*, 2020]. However, predictions are confronted with various uncertainties due to incomplete understanding of the actual natural systems [*Kim et al.*, 2016c; *Kim et al.*, 2016b] and unknown distributions of the parameters that are difficult to measure directly [*Beven and Binley*, 1992; *Kavetski et al.*, 2006; *Moradkhani and Sorooshian*, 2008; *Kim and Ivanov,* 2014; *Kim et al.*, 2016a]. Rigorous enforcement of high-fidelity predictions through understanding, quantifying, and reducing such uncertainties often requires a calibration, optimization, or assimilation process that adjusts a modeling system to the available measurements (e.g., streamflow) [*Beven and Freer*, 2001; *Vrugt and Robinson*, 2007;

Moradkhani and Sorooshian, 2008; *Tran and Kim*, 2021a]. However, this inverse type of modeling generally entails significant computational resources because it relies on a considerable number of repeated model runs for various scenarios [*Liu and Gupta*, 2007; *Keating et al.*, 2010; *Beven and Binley*, 2014; *Zhang et al.*, 2020]. Recently, a very attractive solution has been developed and applied that can drastically reduce the computational costs for the model run. This approach substitutes a high-cost deterministic model with a cheap-to-run "surrogate" model that reproduces comparable physical properties but has a lower computational cost [*Razavi et al.*, 2012b; *Asher et al.*, 2015; *Tran et al.*, 2020].

Surrogate models that originated in a wide range of disciplines are being developed and implemented for water resources problems [*Razavi et al.*, 2012a; *Sargsyan et al.*, 2014; *Christelis and Hughes*, 2018; *Dwelle et al.*, 2019; *Hu et al.*, 2019a; *Tran et al.*, 2020; *Wang et al.*, 2020]. The central premise for a surrogate model to provide results consistent with the original model is to be able to approximate the relationship between input and output similar to the original model [*Wang and Shan*, 2007; *Smith*, 2013; *Asher et al.*, 2015; *Rajabi*, 2019]. However, the current surrogate models based on this relationship cannot provide a reliable prognosis for outliers (or extremes) beyond the training data space, although they generally have excellent predictive power for regions within the training data [*Razavi et al.*, 2012b; *Asher et al.*, 2015; *Matos et al.*, 2017; *Tran et al.*, 2020]. This is because the surrogate model is adapted locally to a constrained number of training points (also referred to as design sites), and thus only sites close to the training space can be diagnosed [*Bowden et al.*, 2012]. For example, in the context of water resource problems, there will be a high probability of extreme events that, due to climate change, have not been experienced in the past [*Prein et al.*, 2016; *Bao et al.*, 2017; *Bloschl et al.*, 2020]. There is also the possibility that extreme events that deviate from recorded events will occur due to climate internal variability, even assuming that climate stationarity is maintained [*Kim et al.*, 2015; *Matos et al.*, 2017; *Kim et al.*, 2018; *Kim et al.*, 2019b; *Doi and Kim*, 2020; 2021]. Therefore, a common solution for ensuring the predictive power in the entire data space is to expand the data range of the design site to cover all possible cases [*Schöbi et al.*, 2017]. However, obtaining sufficient collections of extreme events for training is unfeasible, so one needs to develop an alternative solution to ensure predictability for the events beyond the data space.

Generalized likelihood uncertainty estimation (GLUE) is an uncertainty quantification framework that has been frequently used in different research disciplines over the past 30 years (more than 2,700 citations as of July, 2021 from the Web of Science). GLUE estimates the posterior distribution of model parameters following the concept of "equifinality" whereby various combinations of parameter values can provide equivalently accurate predictions [*Beven and Binley*, 1992; *Beven*, 2006]. Many studies have performed in-depth analyses to compare the performance of GLUE and other Bayesian inference methods, evaluate the effects of formal and informal likelihood functions, and improve the performance of GLUE with advanced sampling techniques. However, setting up an acceptance threshold — one of the most important features for determining the posterior distribution in the GLUE implementation — appears to have received less attention. This acceptance threshold has a direct impact on the accuracy and computational cost of GLUE [*Blasone et al.*, 2008a; *Stedinger et al.*, 2008], but has been arbitrarily determined as either an allowable degree of simulation error or a fixed ratio of the total number of simulations [*Blasone et al.*, 2008b; *Vrugt et al.*, 2008c]. The former approach is to repeat simulations continuously until the number that satisfies the desired accuracy is reached, while the latter is to select only the top few percent of the simulation results performed. Basically, the ability to obtain good behavioral results from both approaches is proportional to the number of simulations. That is, performing as many simulations as possible is a way to achieve better results, but this is not always feasible and effective, especially for models with many uncertain parameters and high computational cost [*Beven and Binley*, 2014]. Therefore, it is of interest to better understand how the trade-off between the accuracy and efficiency of GLUE varies according to the use of different acceptance thresholds.

The literature demonstrates that revolutionary surrogate modeling has been applied for many models and has proven its strengths in a wealth of publications. However, those studies simply focused on the development of new surrogate models and compared their accuracy with other models (e.g., kriging, support vector machines, or radial basis functions) [*Razavi et al.*, 2012a; *Schöbi et al.*, 2015; *Rajabi*, 2019; *Xing et al.*, 2019; *Zhang et al.*, 2020] based on multiple accuracy scores in their construction processes (e.g., relative mean squared error, leave-one-out error, or leave-one-out cross-validation) [*Blatman and Sudret*, 2011; *Lüthen et al.*, 2020]. Training a surrogate model with data of sufficiently large size can usually achieve high accuracy, but unfortunately this may greatly reduce its efficiency [*Razavi et al.*, 2012b]. Recalling that surrogate models are fashioned to offset the expensive computational cost of the original models, one cannot simply sacrifice efficiency for a slight improvement in accuracy. It is therefore paramount to strike a balance between accuracy and efficiency in order to fully exploit the power of surrogate models.

In this Chapter, a primary goal was to gain comprehensive knowledge of building a well-organized surrogate model that can provide reliable ensemble results, even for extreme events that deviate significantly from the training data space. For this purpose, we here present a new surrogate model named polynomial chaos-kriging (PCK) that combines the advantages of two well-known surrogate models, polynomial chaos expansion (PCE) and kriging. While the PCK model has been used in the field of structural engineering with some common, simple benchmarks such as Ishigami, Rosenbrock, borehole, four-branch, or Sobol' functions [*Schöbi et al.*, 2014; *Kersaudy et al.*, 2015;

Cortesi et al., 2019; *Du and Leifsson*, 2019; *Leifsson et al.*, 2020; *Nagawkar et al.*, 2020], little research has been done with it on water resources or geophysical problems such as extreme flood prediction. We also present a unified modeling framework that applies GLUE to the construction of the proposed surrogate model. In this framework, we investigate and discuss the effects of the acceptance threshold types on the model accuracy and efficiency. Finally, we propose a new "performance score" that indicates how much better the accuracy and efficiency of the surrogate model are over the original model, thereby providing a guideline for selecting an appropriate surrogate emulator.

5.2 Methods

5.2.1 Surrogate modeling: Polynomial chaos-kriging

We designed a surrogate model, polynomial chaos-kriging (PCK), with the aim of combining the merits of both polynomial chaos expansion (PCE) for capturing the global tendency of the original model with a set of orthogonal polynomials [*Sudret*, 2008], and kriging for handling the local approximation at training points via Gaussian processes [*Echard et al.*, 2011]. Once we have a training dataset for the input-output relationship of the original hydrologic model, we endeavor to construct a surrogate model ($\boldsymbol{\mathcal{M}}^{\text{su}}$) that supplants the original model $\boldsymbol{\mathcal{M}}$. Here, the PCK consists of two mathematical terms with respect to $\boldsymbol{X}$:

$$\boldsymbol{Y} \approx \boldsymbol{\mathcal{M}}^{\text{su}}(\boldsymbol{X}) = \text{PCK}(\boldsymbol{X}) = \sum_{\alpha=1}^{N_\Psi} \varepsilon_\alpha \Psi_\alpha(\boldsymbol{X}) + \sigma^2 Z(\boldsymbol{X}) \tag{5.1}$$

Note that the first term on the right side $\sum_{\alpha=1}^{N_\Psi} \varepsilon_\alpha \Psi_\alpha(\boldsymbol{X})$ is equivalent to the mathematical expression of PCE in Eq. (3.2) and serves as a trend function within the formulation in Eq. (D.1). σ^2 is the variance (or kriging variance) of the Gaussian process $Z(\boldsymbol{X})$ with zero-mean and unit-variance. The Gaussian process $Z(\boldsymbol{X})$ is characterized by an autocorrelation function (ACF) between two arbitrary input samples $\boldsymbol{X}$ and $\boldsymbol{X}'$, i.e., $R\left(\boldsymbol{X},\boldsymbol{X}'\right) = R\left(\left|\boldsymbol{X}-\boldsymbol{X}'\right|;\boldsymbol{\delta}\right)$ and its hyperparameters $\boldsymbol{\delta}$, which represent the amplitude and the lengths of the correlation [*Bachoc*, 2013]. Various ACFs can be used, such as linear, Dirac, exponential, squared exponential (or Gaussian), and Matérn [*Santner et al.*, 2003; *Bachoc*, 2013; *Schöbi et al.*, 2015]. In this work, the Matérn ACF is adopted as it was favored in previous studies [*Schöbi et al.*, 2015; *Lataniotis et al.*, 2020; *Wang*, 2021].

The PCK was established through three primary procedures [*Schöbi et al.*, 2015]. The first procedure was to estimate the PCE coefficients (ε_{α}) using LAR method (detailed in Chapter II). The second procedure is to determine the parameter $\boldsymbol{\delta}$ of the autocorrelation function and the Gaussian process variance $\boldsymbol{\sigma^2}$. The parameter $\boldsymbol{\delta}$ can be obtained through a maximum-likelihood estimate [*Marrel et al.*, 2008] or a leave-one-out cross-validation [*Bachoc*, 2013]. The latter method was used in this work because it provides more robust results [*Bachoc*, 2013] as:

$$\widehat{\boldsymbol{\delta}} = \mathrm{argmin}_{\boldsymbol{\delta}}[\boldsymbol{\mathcal{Y}}^{\mathrm{T}}\mathbf{R}(\boldsymbol{\delta})^{-1}\mathrm{diag}(\mathbf{R}(\boldsymbol{\delta})^{-1})^{-2}\mathbf{R}(\boldsymbol{\delta})^{-1}\boldsymbol{\mathcal{Y}}] \tag{5.2}$$

where $\widehat{\boldsymbol{\delta}}$ denotes the optimal $\boldsymbol{\delta}$, $\mathbf{R}(\boldsymbol{\delta}) = R\left(\left|\mathcal{X}^{(k1)} - \mathcal{X}^{(k2)}\right|;\boldsymbol{\delta}\right)$ is the correlation matrix of two samples $\boldsymbol{\mathcal{X}}^{(k1)}$ and $\boldsymbol{\mathcal{X}}^{(k2)}$ among the experimental design $\boldsymbol{\mathcal{X}}$ with $k1 = k2 = 1, \dots, N$.

The third procedure is to optimize the Gaussian variance σ^2 given $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\mathcal{Y}}$, $R\left(\boldsymbol{X},\boldsymbol{X}'\right)$, $\widehat{\boldsymbol{\delta}}$, and $\varepsilon_{\alpha}\,\Psi_{\alpha}(\boldsymbol{X})$, wherein $\varepsilon_{\alpha}\,\Psi_{\alpha}(\boldsymbol{X})$ are considered candidates for the trend part of kriging, with the number of N_{Ψ} candidates ($\boldsymbol{\alpha} = 1, \dots, N_{\Psi}$). An iterative algorithm is then employed with N_{Ψ} iterations. The initialization ($\boldsymbol{\alpha} = 1$) is a PCK with one single polynomial (i.e., $\varepsilon_1\,\Psi_1(\boldsymbol{X})$) in the trend part. Iteratively, the polynomials are added one-by-one to the trend part. At each iteration, the Gaussian variance $\boldsymbol{\sigma^2}$ is estimated as follows:

$$\sigma^2 = \frac{1}{N}(\boldsymbol{\mathcal{Y}} - \boldsymbol{F}\varepsilon_{\alpha})^{\mathrm{T}}\mathbf{R}\big(\widehat{\boldsymbol{\delta}}\big)^{-1}(\boldsymbol{\mathcal{Y}} - \boldsymbol{F}\varepsilon_{\mathrm{\alpha}})\ , \boldsymbol{\alpha} = 1, \dots, N \tag{5.3}$$

where $\mathbf{R}\big(\widehat{\boldsymbol{\delta}}\big) = R\big(\left|\mathcal{X}^{(k1)} - \mathcal{X}^{(k2)}\right|; \widehat{\boldsymbol{\delta}}\big)$ is the optimal correlation matrix. $\boldsymbol{F}$ is the information matrix computed as Eq. (2.8). For each iteration, a surrogate model $\mathrm{PCK}_{\boldsymbol{\alpha}}$ with a different variance $\boldsymbol{\sigma^2}$ is constructed. Among the number of N_{Ψ} constructed PCKs, an optimal PCK with a minimal deviation from the original model result is selected [*Schöbi et al.*, 2015]. To quantify this deviation, a leave-one-out error (ϵ) in Eq. (5.4) is used [*Blatman and Sudret*, 2011; *Vapnik*, 2013]:

$$\epsilon = \frac{1}{N}\sum_{k=1}^{N}\left(\boldsymbol{\mathcal{M}}\big(\mathcal{X}^{(k)}\big) - \mathrm{PCK}_{\boldsymbol{\mathcal{X}}^{(-k)}}\big(\mathcal{X}^{(k)}\big)\right)^2 \tag{5.4}$$

where $\mathrm{PCK}_{\boldsymbol{\mathcal{X}}^{(-k)}}$ denotes the PCK model built by using the experimental design $\boldsymbol{\mathcal{X}}^{(-k)} = \left\{\mathcal{X}^{(1)}, \mathcal{X}^{(2)}, \dots, \mathcal{X}^{(k-1)}, \mathcal{X}^{(k+1)}, \dots, \mathcal{X}^{(N)}\right\}$ with the size of $N - 1$.

5.2.2 Parameter inference using GLUE

Parameter inference is to deduce which values of uncertain parameters $\boldsymbol{\theta}$ are likely to provide predictions consistent with observations. From the inference results, one can generate a posterior distribution of parameters given the observed streamflow and demonstrate the possible ensemble outcomes for that distribution. The posterior distribution of $\boldsymbol{\theta}$ conditioned on observations y^{obs} is generally expressed using Bayes' rule as [*Tarantola*, 2005]:

$$\Pi(\boldsymbol{\theta}|y^{obs}, \boldsymbol{x}, \boldsymbol{u}) \propto \mathcal{L}(y^{obs}, \boldsymbol{x}, \boldsymbol{u}|\boldsymbol{\theta})\rho(\boldsymbol{\theta}) \tag{5.5}$$

where $\rho(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$ generated based on their prior knowledge; $\mathcal{L}(y^{obs}, \boldsymbol{x}, \boldsymbol{u}|\boldsymbol{\theta})$ is the likelihood, which represents the conditional probability of the model results given the set of parameters; and $\Pi(\boldsymbol{\theta}|y^{obs}, \boldsymbol{x}, \boldsymbol{u})$ is the posterior distribution of $\boldsymbol{\theta}$.

This study employs the GLUE framework, which is straightforward to implement and allows flexibility in choosing an informal likelihood function and its cutoff threshold [*Beven*, 2006]. To characterize deviations for the shape, peak, and volume of a flood hydrograph simultaneously, a combination of Nash-Sutcliffe efficiency (NSE), peak error (PE), and volume error (VE) is presented as a likelihood function (L):

$$L = \frac{\left(\frac{\sum_{t=1}^{\mathrm{T}}(y_t^{obs}-y_t)^2}{\sum_{t=1}^{\mathrm{T}}(y_t^{obs}-\overline{y^{obs}})^2}\right)+\left(\frac{\left|y_{max}^{obs}-y_{max}\right|}{y_{max}^{obs}}\right)+\left(\frac{\left|V^{obs}-V\right|}{V^{obs}}\right)}{3} \tag{5.6}$$

where y_t^{obs} and y_t are the observed and simulated streamflow at time t, respectively; T is the total number of time steps over the flood event; y_{max}^{obs} and y_{max} are the observed and simulated streamflow at peak, respectively; V^{obs} and V denote the total volume of observed and simulated hydrographs, respectively. Note that the sub-equations in the three parentheses represent the complementary NSE $(1 - NSE)$, $PE,$ and $VE,$ respectively. This likelihood function has a value in the range of 0 to 1, and the closer this value is to the minimum, the smaller the error. The cutoff threshold can be specified as either an allowable deviation of the likelihood function (here named "accuracy-aimed threshold") or a fixed ratio of the total number of simulations (here named "efficiency-aimed threshold") [*Beven and Freer*, 2001; *Vrugt et al.*, 2008c]. The entire simulations performed are divided into the runs that satisfy (*behavioral*) or do not meet (*non-behavioral*) this threshold condition, where the behavioral runs are leveraged to trace out $\Pi(\boldsymbol{\theta}|y^{obs}, \boldsymbol{x}, \boldsymbol{u})$.

5.2.3 A framework of a surrogate model-based uncertainty quantification

Figure 5.1 presents a general framework for constructing PCK and inferring the model parameters in a computationally efficient manner by coupling PCK and GLUE. Generally, one uses the set of inputs and outputs (or the experimental design and model response) of an original hydrological model (Box A) to construct a PCK surrogate model (Box B) that allows for fast computation of the inverse inference for uncertain parameters of the hydrological model (Box C).

Specifically, in Box (A), the experimental design $\boldsymbol{\mathcal{X}}$ consists of N sets of $\boldsymbol{\theta}$, $\boldsymbol{x}$, and $\boldsymbol{u}$, wherein the $\boldsymbol{\theta}$ values are chosen randomly from the uniform (prior) distribution $\rho(\boldsymbol{\theta})$ using Latin hypercube sampling (LHS), the states $\boldsymbol{x}$ are initialized with zero vectors, and the forcing $\boldsymbol{u}$ values are collected by using the climate data (i.e., rainfall) from historical events. The corresponding response $\boldsymbol{\mathcal{Y}}$ values are then obtained by applying $\boldsymbol{\mathcal{X}}$ to the hydrological model $\boldsymbol{\mathcal{M}}$.

Box (B) illustrates the construction procedure of PCK given $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$. The PCE coefficients (ε_{α}) are first estimated by LAR given $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}$, and the polynomial degree (p). The hyperparameter $\boldsymbol{\delta}$ of the autocorrelation function $R\left(\boldsymbol{X}, \boldsymbol{X}'\right)$ is then optimized as Eq. (5.2). Once both ε_{α} and $\boldsymbol{\delta}$ are determined, an iterative algorithm begins to optimize the Gaussian variance σ^2 (Eq. (5.3)) and construct a surrogate PCK (Eq. (5.1)), and this continues until N_{Ψ} number of iterations. Among the N_{Ψ} constructed PCKs, the PCK having the smallest leave-one-out error computed in Eq. (5.4) is selected as the optimal PCK.

Once a surrogate model has been constructed, one can use it for computationally inexpensive parameter inference. GLUE (Box C) provides the posterior (or behavioral) distribution of the parameters, $\Pi(\boldsymbol{\theta}|y^{obs}, \boldsymbol{x}, \boldsymbol{u})$, and the uncertain interval of the simulated streamflow y. The latter predictions match with the observed streamflow y^{obs} and satisfy the acceptance threshold. In present work, two types of acceptance thresholds, the accuracy-aimed threshold and the efficiency-aimed threshold, are unitized. In GLUE, $\boldsymbol{\theta}$ and $\boldsymbol{x}$ are initialized similarly to Box A, while the forcing $\boldsymbol{u}$ values are used from potential future events that include possible “extreme” events.

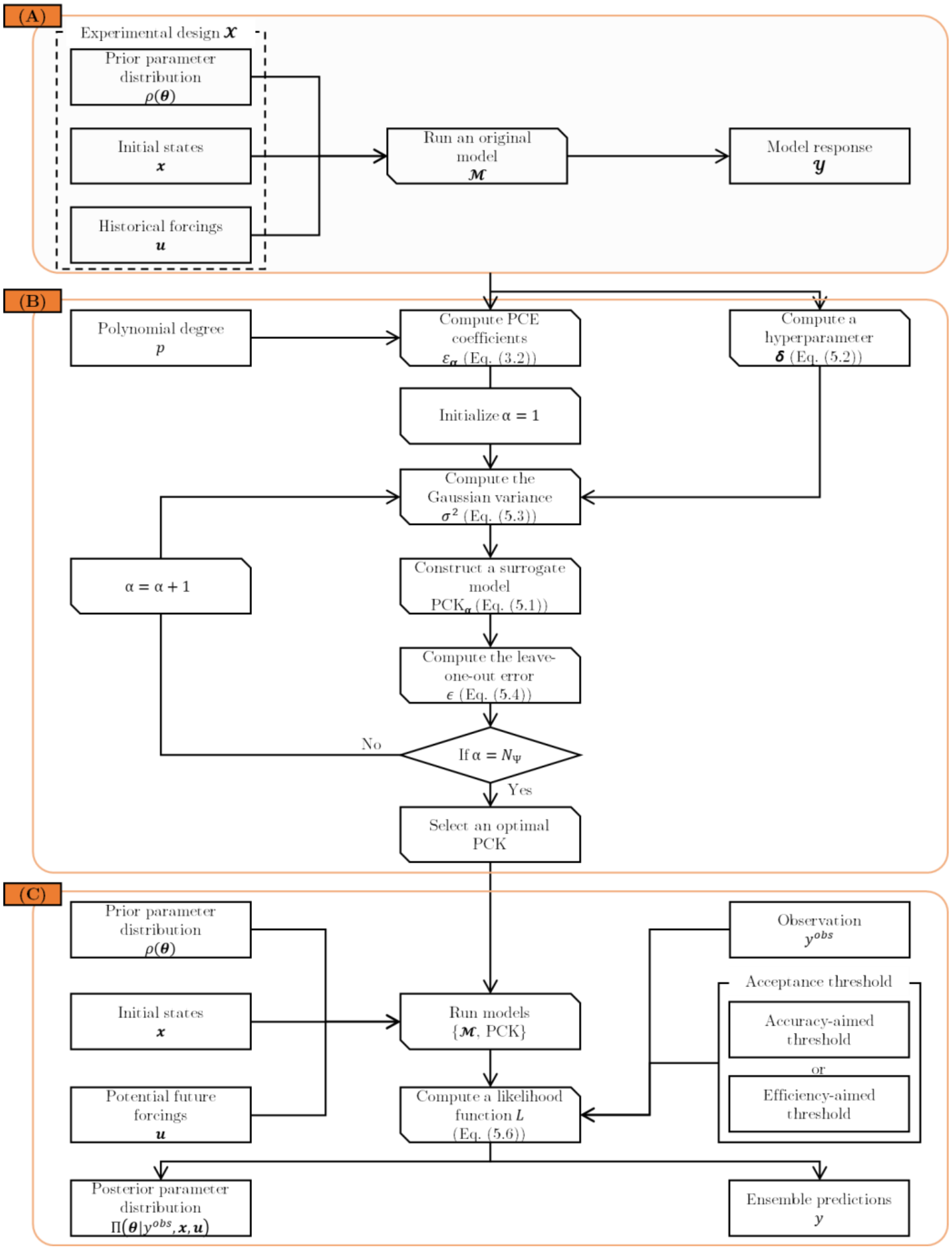

(A)
Experimental design $\boldsymbol{\mathcal{X}}$
Prior parameter distribution $\rho(\boldsymbol{\theta})$
Initial states $\boldsymbol{x}$
Historical forcings $\boldsymbol{u}$
Run an original model $\boldsymbol{\mathcal{M}}$
Model response $\boldsymbol{\mathcal{Y}}$
(B)
Polynomial degree p
Compute PCE coefficients ε_{α} (Eq. (3.2))
Compute a hyperparameter $\boldsymbol{\delta}$ (Eq. (5.2))
Initialize $\alpha = 1$
Compute the Gaussian variance σ^2 (Eq. (5.3))
$\alpha = \alpha + 1$
Construct a surrogate model PCK_{α} (Eq. (5.1))
Compute the leave-one-out error ϵ (Eq. (5.4))
No
If $\alpha = N_{\Psi}$
Yes
Select an optimal PCK
(C)
Prior parameter distribution $\rho(\boldsymbol{\theta})$
Initial states $\boldsymbol{x}$
Potential future forcings $\boldsymbol{u}$
Run models $\{\boldsymbol{\mathcal{M}}, \mathrm{PCK}\}$
Compute a likelihood function L (Eq. (5.6))
Observation y^{obs}
Acceptance threshold
Accuracy-aimed threshold
or
Efficiency-aimed threshold
Posterior parameter distribution $\Pi(\boldsymbol{\theta}|y^{obs}, \boldsymbol{x}, \boldsymbol{u})$
Ensemble predictions y

Figure 5.1. The workflow of PCK construction and its uncertainty quantification. Box (A) demonstrates the collection of the experimental design $\boldsymbol{\mathcal{X}}$ and model response $\boldsymbol{\mathcal{Y}}$. Wherein, $\boldsymbol{\mathcal{X}}$ includes the parameters $\boldsymbol{\theta}$ samped from their prior distributions $\rho(\boldsymbol{\theta})$ using LHS sampling, the model states $\boldsymbol{x}$, and the forcings $\boldsymbol{u}$ generated from historical events. $\boldsymbol{\mathcal{Y}}$ is attained by propagating $\boldsymbol{\mathcal{X}}$ through an original hydrological model $\boldsymbol{\mathcal{M}}$. Box (B) refers to the optimization process of constructing PCK. Box (C) describes the use of constructed PCK to make an inference (i.e., GLUE) to obtain ensemble streamflow y and posterior parameters $\Pi(\boldsymbol{\theta}|y^{obs}, \boldsymbol{x}, \boldsymbol{u})$ based on the likelihood function L, two types of acceptance thresholds ('accuracy-aimed' or 'efficiency-aimed' thresholds), and observed streamflow y^{obs}.

5.2.4 Performance metrics

To investigate how accurately the three surrogate models (PCE, OK, and PCK) mimic the original model for the experimental design during the training process, one often uses the leave-one-out error (ϵ) in Eq. (5.4). A smaller ϵ indicates a more accurate emulator.

For ensemble predictions, assessment through deterministic and probabilistic measures is necessary. The likelihood function L adopted in GLUE is also exploited as a deterministic measure. For the probabilistic measures, the $CRPS$ and $Spread$ are selected.

Other than the metrics above, a new "performance score" (PS) is proposed that can evaluate the overall performance of surrogate models ($\boldsymbol{\mathcal{M}}^{\mathrm{su}}$) by weighting the accuracy and efficiency of the surrogate model as compared to the original model.

$$PS(\boldsymbol{\mathcal{M}}^{\mathrm{su}}) = d\left(\frac{L(\boldsymbol{\mathcal{M}}^{\mathrm{su}})}{L(\boldsymbol{\mathcal{M}})}\right) \times d\left(\frac{GLUE(\boldsymbol{\mathcal{M}}^{\mathrm{su}})}{GLUE(\boldsymbol{\mathcal{M}})}\right) \tag{5.7}$$

$$d\left(\frac{L(\boldsymbol{\mathcal{M}}^{\mathrm{su}})}{L(\boldsymbol{\mathcal{M}})}\right) = \sqrt{\sum_{l=1}^{N_E}[L(\boldsymbol{\mathcal{M}}^{\mathrm{su}})_l - L(\boldsymbol{\mathcal{M}})_l]^2} \tag{5.8}$$

$$d\left(\frac{GLUE(\boldsymbol{\mathcal{M}}^{\mathrm{su}})}{GLUE(\boldsymbol{\mathcal{M}})}\right) = \frac{RT(\boldsymbol{\mathcal{M}}^{\mathrm{su}}) \times N_{runs}(\boldsymbol{\mathcal{M}}^{\mathrm{su}})}{RT(\boldsymbol{\mathcal{M}}) \times N_{runs}(\boldsymbol{\mathcal{M}})} \tag{5.9}$$

where $L(\boldsymbol{\mathcal{M}}^{\mathrm{su}})$ and $L(\boldsymbol{\mathcal{M}})$ signify the values of the likelihood function for the behavioral runs of the surrogate and original models, respectively; $GLUE(\boldsymbol{\mathcal{M}}^{\mathrm{su}})$ and $GLUE(\boldsymbol{\mathcal{M}})$ are the total runtimes to attain the behavioral runs of the surrogate and original models by GLUE; $RT(\boldsymbol{\mathcal{M}}^{\mathrm{su}})$ and $RT(\boldsymbol{\mathcal{M}})$ denote the runtimes needed for a single run for the surrogate and original models, respectively; and $N_{runs}(\boldsymbol{\mathcal{M}}^{\mathrm{su}})$ and $N_{runs}(\boldsymbol{\mathcal{M}})$ respectively signify the number of prior runs required to acquire a predefined number of behavioral runs that satisfy the condition of cutoff threshold. The first term, $d\left(\frac{L(\boldsymbol{\mathcal{M}}^{\mathrm{su}})}{L(\boldsymbol{\mathcal{M}})}\right)$ represents the accuracy of the performance, defined as the

Euclidean distance between the two ensemble sets, $L(\boldsymbol{\mathcal{M}}^{\text{su}})$ and $L(\boldsymbol{\mathcal{M}})$ in Eq. (5.8). A smaller $d\left(\frac{L(\boldsymbol{\mathcal{M}}^{\text{su}})}{L(\boldsymbol{\mathcal{M}})}\right)$ indicates that the surrogate model has the equivalent accuracy as the original model and effectively replaced the original model. On the other hand, the second term, $d\left(\frac{GLUE(\boldsymbol{\mathcal{M}}^{\text{su}})}{GLUE(\boldsymbol{\mathcal{M}})}\right)$, represents the efficiency performance, defined as the relative difference in the total runtimes needed for uncertainty quantification. If this term is small, it means that the efficiency of a surrogate model is very good. Therefore, the range of possible values of PS combining those two terms is from 0 to infinity. When PS is close to 0, it means that the chosen surrogate model has accuracy similar to the original model and finishes the uncertainty quantification in a very short time. When PS approaches infinity, both the accuracy and efficiency of the surrogate model are very low.

5.3 Experimental setup

The Thu Bon river watershed with eight flood events and the NAM model are selected to conduct all experiments. The detailed information about the study area and NAM was presented in Chapter II (see Sections 2.2.1.3 and 2.2.2). This study highlights the robustness of PCK in its ability to capture the original model sufficiently, even with small experimental designs, and also to predict extreme flood events that are very different from the training events, compared to two popular surrogate models of PCE and OK. For more descriptions of OK, refer to Appendix D. All experiments conducted for the three surrogate models are listed as follows.

5.3.1 Setup for constructing the surrogate models

To construct the experimental design with the size N, we initialized the ensemble of the states as zero, specified the ensemble of the parameters with the values sampled from the Uniform (prior) distribution, and generated random rainfall values over a bounded interval regarding the ensemble of forcings (i.e., hourly rainfall). In order to extract the rainfall data for training, it was assumed that the range of possible rainfall was 0 to 20 mm/hr, and the interval of rainfall used for later application and verification was 0 to 34.1 mm/hr. The number 34.1 mm/hr was the largest value among the rainfall events (Table 2.5) that occurred in the past in this study area. The reason for assuming 20 mm/hr was to use a value much smaller than the actual maximum rainfall in order to emphasize the predictability of the surrogate models for events beyond the training data space. The experimental design was established as N = 1,000, which was considered a reasonable size to adequately construct a surrogate model for NAM [*Tran et al.*, 2020]. For the polynomial degree (p) of PCE and PCK, we chose 3, the most preferred value in prior studies [*Fan et al.*, 2016; *Wang et al.*, 2017; *Hu et al.*, 2019a; *Tran and Kim*, 2019; *Tran et al.*, 2020].

5.3.2 Setup for the parameter sensitivity analysis

Once the surrogate models were constructed, a sensitivity analysis (SA) of the nine parameters was performed to investigate how similar the sensitive behaviors of the surrogate models and NAM were. A Sobol' sensitivity analysis was selected since it has been extensively employed as one of the most effective and attractive methods [*Sobol'*, 2001; *Saltelli*, 2002b; *Sudret*, 2008]. Our SA results were analyzed based on the main (total-order) index (S_{Total}) of Sobol' (see Section 2.2.1.4), calculated with 22,000 random runs as suggested in *Tran and Kim* [2019]. Their parameter sets were randomly generated by LHS. These SA experiments were done for the four models (NAM, PCE, OK, and PCK) for the eight selected rainfall events.

5.3.3 Setup for the parameter inference via GLUE

GLUE was applied to infer the parameter uncertainties for the eight flood events, and the two types of cutoff thresholds were adopted. The first threshold was the "accuracy-aimed threshold" that can control the accuracy, and a value of 0.1 was specified as the threshold for the likelihood function L defined in Eq. (5.6). This threshold corresponds to a combination of accuracies of about 0.8, 5%, and 5% for NSE, PE, and VE, respectively [*Tran and Kim*, 2019]. The implementation of GLUE can be stopped when 1,000 behavioral sets are attained, or intentionally stopped when the number of random runs reaches 100 million, even if 1,000 behavioral sets are not obtained. The second threshold is the "efficiency-aimed threshold" to control the efficiency, and for a total of 100,000 random runs generated by LHS, the acceptance rate of the top 1% was set as the threshold (i.e., among 100,000 runs, the 1,000 runs with higher accuracy were selected).

5.4 Results

5.4.1 Training error for constructing surrogate models

Using the same experimental design with the same size (N = 1,000) assembled in Section 5.3.1, three surrogate models of PCE, OK, and PCK were constructed, and the leave-one-out error (ϵ) was computed for six QoIs (i.e., streamflow and the five model states), as reported in Fig. 5.2. Quantitative inspection of this figure indicates that PCK always has a smaller ϵ value, which outperforms both PCE and OK in capturing the NAM behavior, while PCE and OK have almost identical performance except for the two QoIs of U and OF. Specifically, the ϵ of PCK is always less than 0.01 for all the QoIs, and is about two to seven times smaller than that of PCE and OK (Fig. 5.2). Thus, PCK was particularly effective in accurately estimating important QoIs such as y (the primary output of interest), U, and OF (two model states that have a significant impact on runoff, especially in the flood season). For example, the ϵ values computed for y are 0.006, 0.0252, and 0.0248, and those for U are 0.004, 0.0194, and 0.0121 for PCK, PCE, and OK, respectively; for OF, the PCK ϵ value was about 4.4 and 7.2 times smaller compared to those of PCE and OK, respectively.

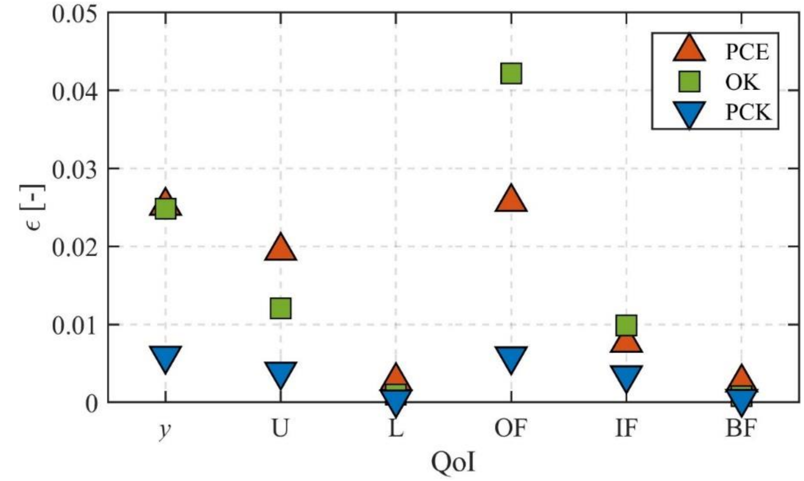


Figure 5.2. Training error (i.e., the leave-one-out error (ϵ) in Eq. (9)) of three surrogate models (PCE, OK, and PCK) for six quantity of interests (QoIs).

5.4.2 Comparisons of parameter sensitivity of surrogate and original models

The comparison results for the sensitivities of the nine parameters with respect to the four models are displayed in Figs. 5.3 and E.1. A parameter with a large value of the Sobol' main index (S_{Total}) indicates that it is relatively sensitive to L. Overall, several of the parameters (CQOF, CK12, and Lm) are the most sensitive parameters, but the more extreme the event, the more absolute the influence of one parameter (CQOF) (see S_{Total} of CQOF for the smallest event, Event 2 and the largest event, Event 8 in Fig. 5.3a-b).

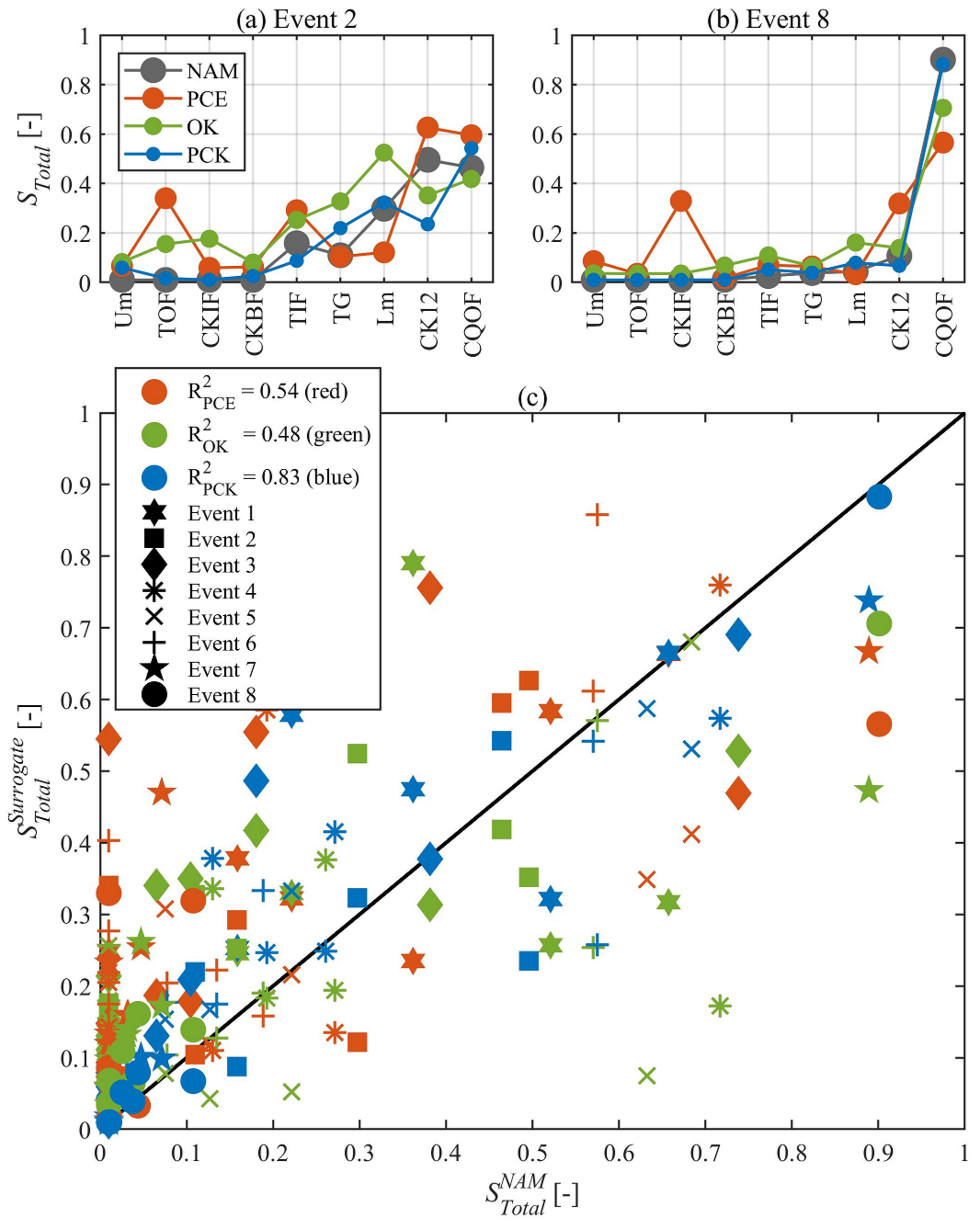


Figure 5.3. Main sensitivity indices (S_{Total}) of nine parameters for four models (NAM, PCE, OK, and PCK) based on the variance of L for (a) Event 2 and (b) Event 8. (c) demonstrates the 1:1 comparisons of the main indices between three surrogate models and NAM for nine parameters over eight test events.

Rather than comparing the relative sensitivities among the parameters, comparing the sensitivities between the three surrogate models and NAM showed that the sensitivity results of PCK were more consistent with those of NAM than with those of PCE and OK, especially for the three sensitive parameters Lm, CK12, and CQOF. The similarity of these parameter sensitivities was further confirmed by calculating R^2 for a 1:1 comparison of the main indices between the surrogates and NAM for the eight rainfall events (Fig. 5.3c). Among the three surrogate models, the sensitivity similarity between PCK and NAM was the highest (R^2 = 0.83), whereas PCE and OK were somewhat different from the sensitivity of the original model (R^2 = 0.54 and 0.48, respectively). In summary, the sensitivity analysis demonstrated that PCK had parametric characteristics and behaviors comparable to NAM.

5.4.3 Predictability skills of the surrogate models

Using the constructed surrogate models and the two types of cutoff thresholds designed in Section 5.3.3, we quantified the uncertainty of the flood prediction by GLUE. A total of eight selected flood events were used for this experiment (Table 2.5). First, the results of GLUE using the accuracy-aimed threshold of 0.1 are shown in Figs. 5.4 and 5.5. Figure 5.4 shows the hydrographs predicted by NAM and the three surrogate models (PCE, OK, and PCK), with a 95% confidence interval quantified from 1,000 behavioral runs of GLUE. For the small-to-medium events (Events 1 to 6), the NAM and the three surrogate models have very narrow uncertainty ranges and provide good results close to the observations. However, for the extreme flood events (Events 7 and 8), of the three surrogate models, only PCK provides satisfactory results (i.e., close to the observations and NAM). We further clarified these latter results by comparing the quantitative magnitudes of the various accuracy metrics in Fig. 5.5. The values of L, $\overline{CRPS}$, and $Spread$ show that PCK outperforms PCE and OK in characterizing the extreme flows (Events 7 and 8). All the values of L for PCK are smaller than 0.1 and are almost equal to those of NAM; its values of $\overline{CRPS}$ and $Spread$ are equivalent to those of NAM.

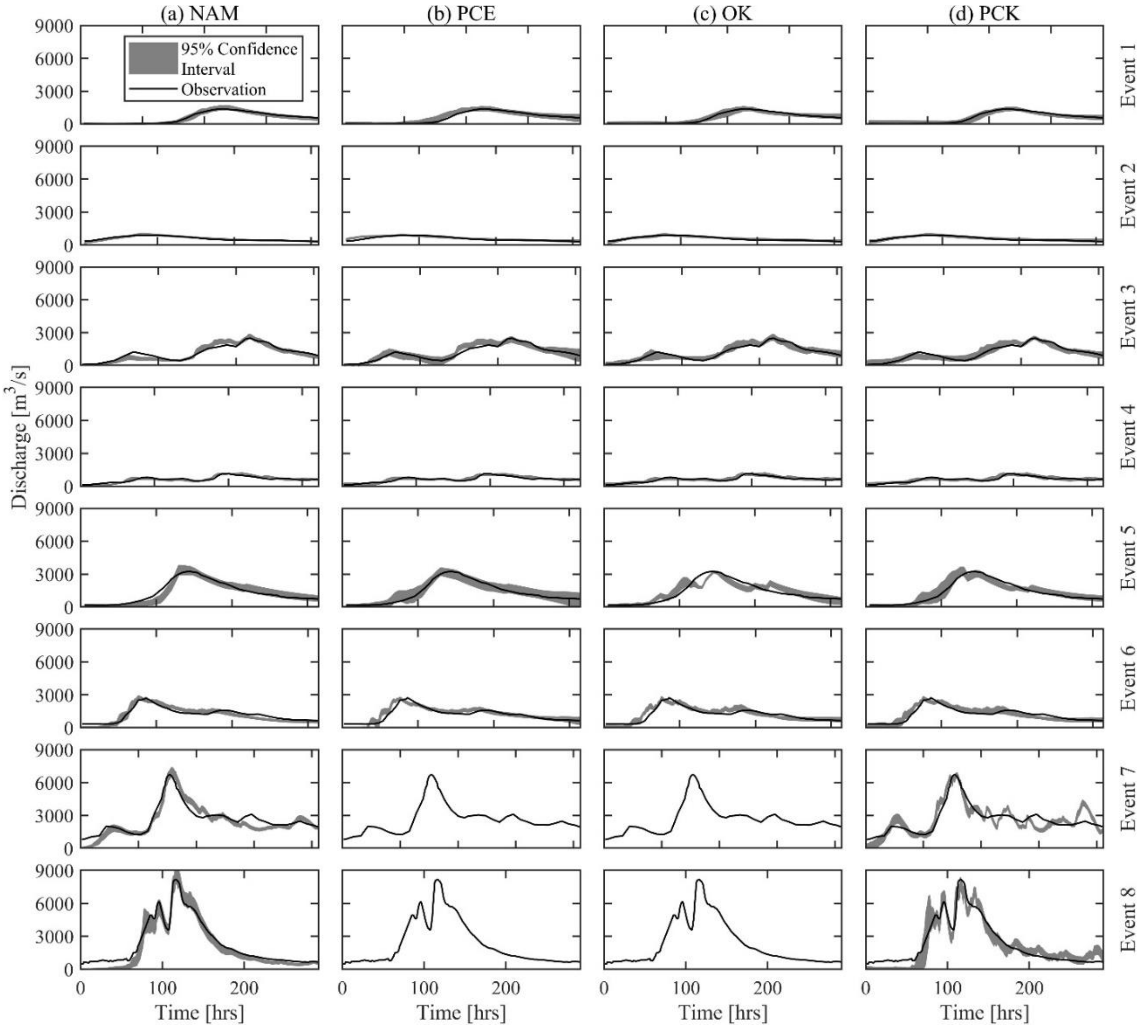


Figure 5.4. Streamflow observed and predicted from four models (NAM, PCE, OK, and PCK) for eight test events. Uncertainty of their predictions is quantified with a 95% confidence interval of the 1,000 behavioral ensemble members through GLUE. The accuracy-aimed threhold of 0.1 was used.

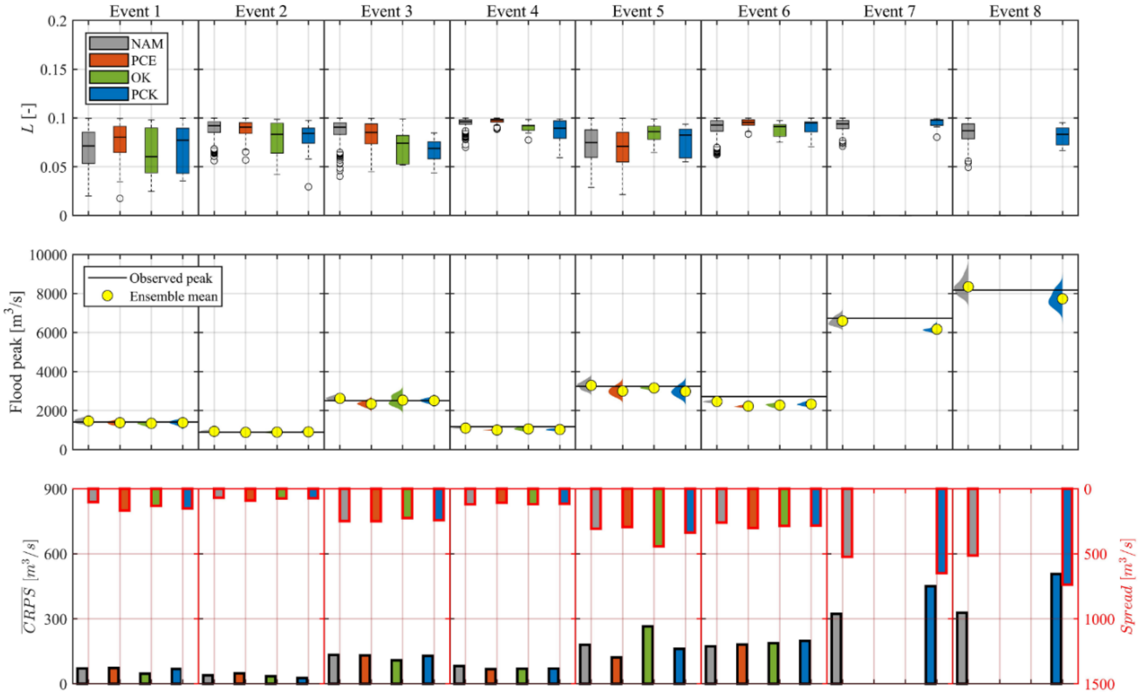


Figure 5.5. Accuracy metric comparisons of three surrogate models with NAM for the 1,000 ensemble members from GLUE with the accuracy-aimed threhold of 0.1 over eight test events: (first row) L, (second row) probability distribution function (PDF) of flood peak, (third row – left axis) $\overline{CRPS}$, and (the third row – right axis) $Spread$. The boxplots in the first row demonstrate the median (central mark), the 25th and 75th percentiles (the edges of the box), and the maximum and minimum (the upper and lower whiskers) except for outliers (dot symbols). The PDFs in the second row are made by kernel density estimation over 1,000 ensemble.

Then we quantified the results of GLUE using an efficiency-aimed threshold of 1% out of a total of 100,000 runs, shown in Figs. 5.6 and 5.7. For the two extreme events, only PCK gave satisfactory results that agreed with the observations, although its range of uncertainty was wide compared to NAM. For the rest of the normal-sized events, the simulated results of the surrogate models were all similar in terms of accuracy and uncertainty range (Fig. 5.6). The values of L, $\overline{CRPS}$, and $Spread$ reported in Fig. 5.7 further confirm that PCK made more accurate predictions for all the events compared to the other surrogate models, PCE and OK. The values of all the PCK metrics are relatively similar to those of NAM and some have better values. On the other hand, the values for PCE and OK were greater (giving worse results) than those for NAM and PCK.

Specifically, the values of $\bar{L}$, $\overline{CRPS}$, and $Spread$ for PCE were about 2.8, 1.4, and 1.6 times larger than those for NAM, respectively, at for Event 7, and about 3.5, 2.2, and 2.6 times larger for Event 8. These values for OK were about 3, 1.6, and 1.8 times larger than NAM for Event 7, and about 3.9, 2.5, and 2.8 times larger for Event 8. In comparing the flood peaks for the two extreme events, PCE and OK could not predict the observed peaks at all, while PCK did have some predictive ability.

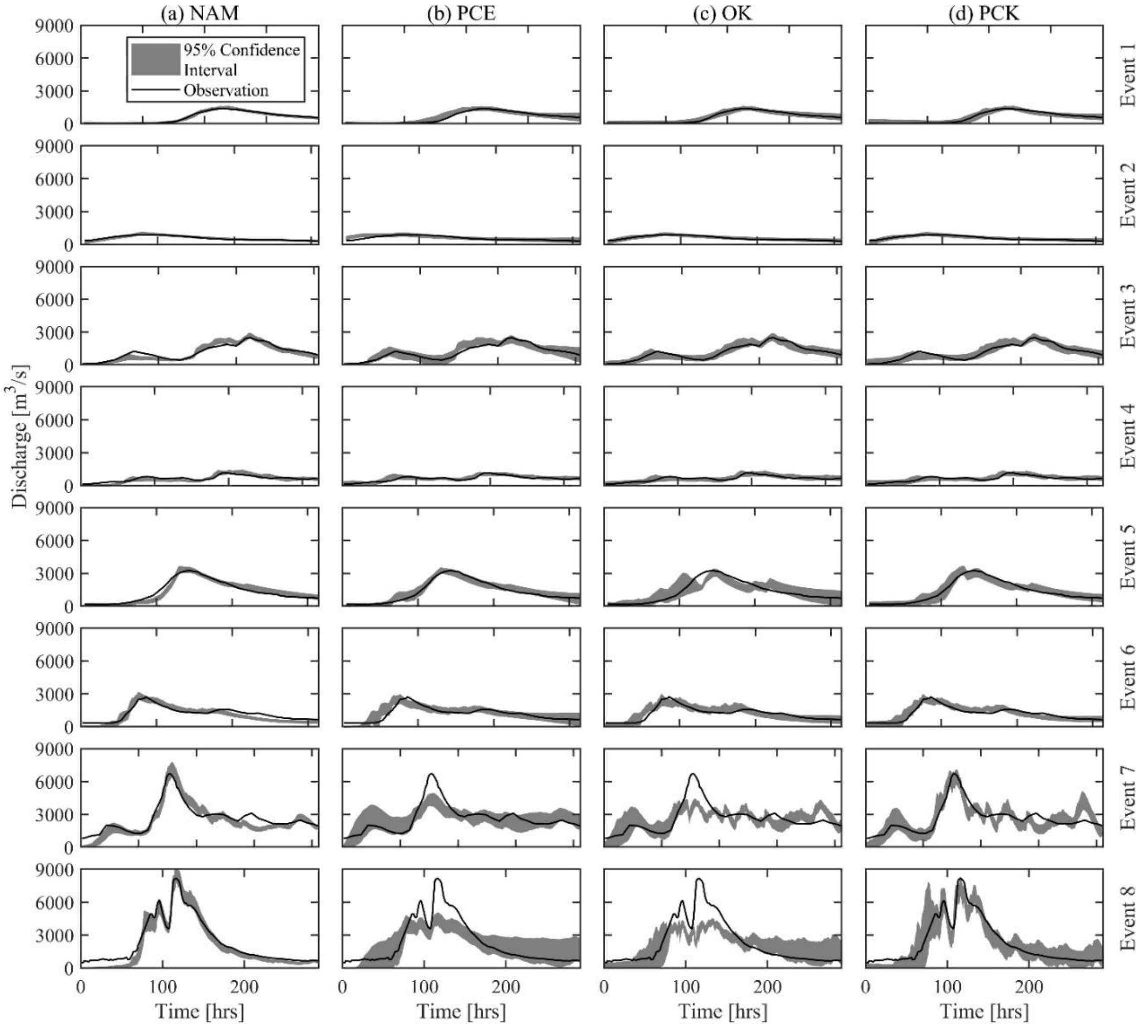


Figure 5.6. Streamflow observed and predicted from four models (NAM, PCE, OK, and PCK) for eight test events. Uncertainty of their predictions is quantified with a 95% confidence interval of the 1,000 behavioral ensemble members through GLUE. The efficiency-aimed threhold of 1% was used over 100,000 random runs.

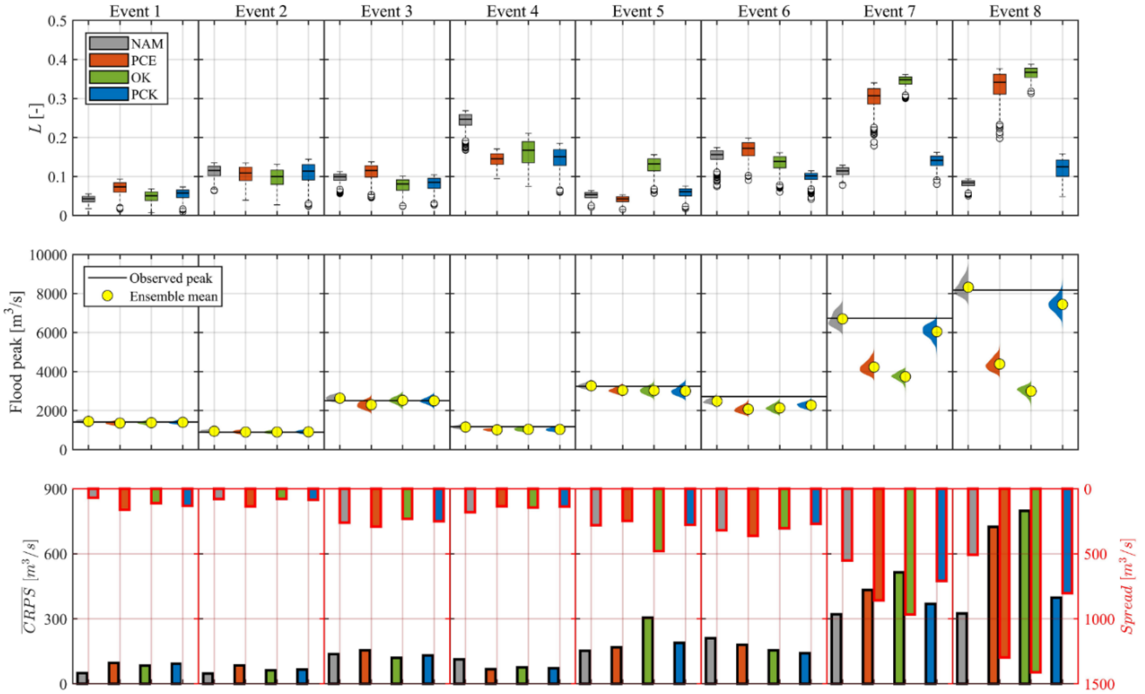


Figure 5.7. Accuracy metric comparisons of three surrogate models with NAM for the 1,000 ensemble members from GLUE with the efficiency-aimed threhold of 1% over eight test events: (first row) L, (second row) probability distribution function (PDF) of flood peak, (third row – left axis) $\overline{CRPS}$, and (the third row – right axis) $Spread$. The boxplots in the first row demonstrate the median (central mark), the 25th and 75th percentiles (the edges of the box), and the maximum and minimum (the upper and lower whiskers) except for outliers (dot symbols). The PDFs in the second row are made by kernel density estimation over 1,000 ensemble.

5.4.4 Performance score (PS) of the surrogate models to acceptance thresholds

Our evaluations of both the accuracy and the efficiency of the behavior sets were carried out simultaneously. We used the performance score (PS) given in in Eq. (5.7) for this purpose, and the changes in PS according to the different levels of criteria are shown in Figs. 5.8 and 5.9. First, Fig. 5.8 shows the change in PS with respect to the accuracy-aimed threshold. For a fixed threshold, both the denominator and numerator of the first term of PS in Eq. (5.7) have comparable values, that is, $d\left(\frac{L(\mathcal{M}^{\mathrm{su}})}{L(\mathcal{M})}\right) = \mathcal{O}(0)$, so the difference in PS values was greatly affected by the second term,

$d\left(\frac{GLUE(\boldsymbol{\mathcal{M}}^{\text{su}})}{GLUE(\boldsymbol{\mathcal{M}})}\right)$. Overall, PCK showed smaller PS values (i.e., had better performance) than PCE and OK; also, the stricter the accuracy criterion that was applied, the greater this performance difference was (especially in the extreme events, Events 7 and 8). This was because only PCK quickly provided behavior sets that satisfied the small acceptance thresholds. In the cases of PCE and OK, the time for uncertainty quantification was much longer than that of PCK, and we could not find any behavior sets that satisfied the thresholds of < 0.2 (for PCE) and < 0.4 (for OK).

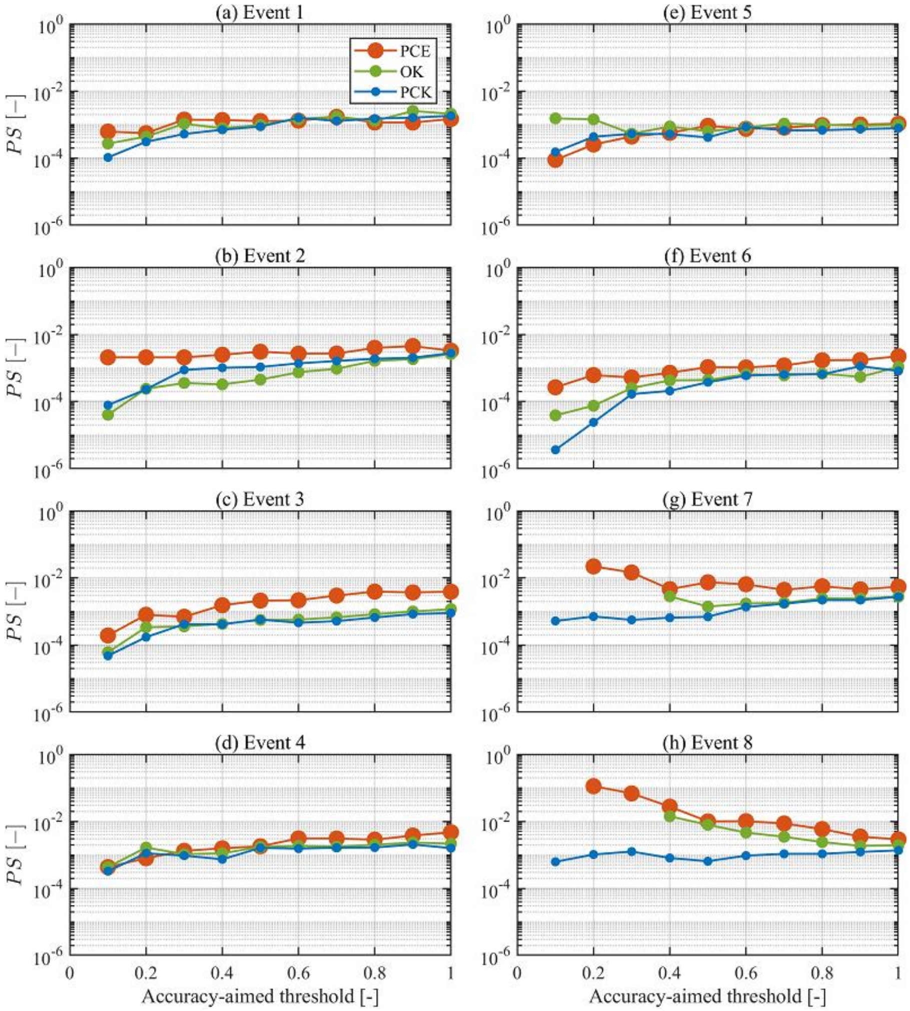


Figure 5.8. Performance score (PS) in Eq. (5.7) of surrogate models to the accuracy-aimed thresholds over eight test events. Note that in Events 7 and 8, some PS values of PCE and OK were not drawn because we could not attain any behavior runs from 100 million random runs.

Figure 5.9 shows the change in PS with respect to the efficiency-aimed fixed threshold. Both the denominator and numerator of the second term of the PS have similar values, that is, $d\left(\frac{GLUE(\boldsymbol{\mathcal{M}}^{\mathrm{su}})}{GLUE(\boldsymbol{\mathcal{M}})}\right) = \mathcal{O}(0)$, so the difference in PS values was greatly affected by the first term, $d\left(\frac{L(\boldsymbol{\mathcal{M}}^{\mathrm{su}})}{L(\boldsymbol{\mathcal{M}})}\right)$. Therefore, the performance difference depends on the difference in predictability for the behavior sets between the models. As before, PCK showed smaller PS values (i.e., had better performance) than PCE and OK, and this trend was especially pronounced in the extreme events (see Fig. 5.9g-h).

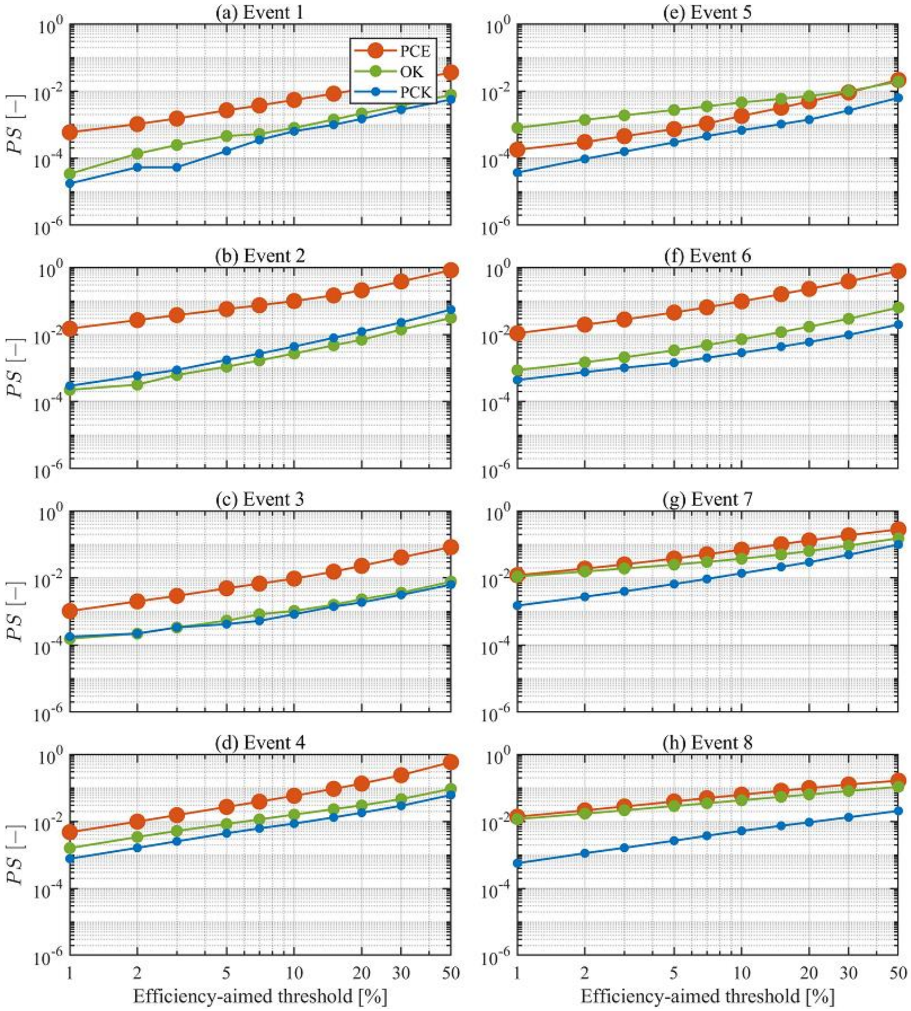


Figure 5.9. Performance score (PS) in Eq. (5.7) of surrogate models to the efficiency-aimed thresholds over eight test events.

5.5 Discussions

5.5.1 How can PCK accurately diagnose extreme events beyond its training data space?

Theoretically, a surrogate model has the predictive power of its original model, provided that the amount of data required to construct it is sufficient. Therefore, when past events are repeated countless times and serve as "big data," a surrogate model can become a great tool to replace the original model. However, in the absence of such sufficient data (for a variety of reasons), a surrogate model will not be properly constructed. This is indeed problematic if one attempts to predict extreme events outside the training data space.

The results in Section 5.4 demonstrated the promise of PCK to perform better than conventional surrogate models in predicting unknown extreme events. That is, the potential data space that PCK could simulate was much wider than that of PCE and OK, even though all three models were trained with the same experimental design (training set). This extraordinary capability of PCK can be explained by its ability to grasp real trends with a set of orthogonal polynomials $\Psi_{\alpha}(\boldsymbol{X})$ associated with PCE coefficients varying between 0 and 10,429 m^3/s, instead of using a trend as a fixed value (e.g., 3,216 m^3/s in OK). It is further explained by an ability to broaden the predictable data space in combination with a stochastic kriging variance. Adding this variance stochastically can increase the overall accuracy of model simulations compared to single PCE when predicting points located in a sparse or rare design data space [*Bichon et al.*, 2011; *Echard et al.*, 2011; *Schöbi et al.*, 2014]. By coupling both the ability of PCE to accurately capture global behaviors and the ability of OK to secure more prediction margins locally, the possibility of effectively addressing the extrapolation tasks, which is difficult with existing data-driven models, is confirmed.

5.5.2 A practical compromise between accuracy and efficiency in uncertainty quantification

From the simulation results presented in Sections 5.4.3 and 5.4.4, it is apparent that both forms of the cutoff threshold have a significant impact on the accuracy and efficiency of the GLUE products. Each has its pros and cons, so it is difficult to discern which method is better from a practical standpoint. If one has enough computational resources and it is more important to attain accurate outcomes, it is better to select the accuracy-aimed threshold. It can consistently guarantee the accuracy of the simulation, although it can lead to a computational burden because a substantial number of model runs (up to billions) is entailed (Fig. 5.10b). Therefore, this method is impractical when applied to expensive computational models [*Iorgulescu et al.*, 2007; *Tran et al.*, 2020]. One should always mind the possibility that GLUE will fail to converge (Fig. 5.8) if one desires too accurate results (i.e., if the threshold selection is not appropriate).

On the other hand, if one is concerned about non-convergence, or it is important to finalize the quantification of uncertainty within an allocated time, use of the efficiency-aimed threshold is recommended. This method can control computational efficiency with a predefined number of random runs, but may provide merely ordinary performance if the predefined number is insufficient (see the results of PCE and OK for Events 7 and 8 in Figs. 7 and 8). Note that if one arbitrarily selects a small efficiency-aimed threshold to yield more accurate behavioral results, the disadvantage of the increased computational burden may be greater than the advantage of improved accuracy. This is because the accuracy improvement of the posterior simulations is not linear with the change in the efficiency-aimed thresholds. For example, reducing the threshold by a factor of 2 from 2% to 1% only reduces L averaged over the ensemble by about 1.13, 1.10, 1.05, and 1.21 times for NAM, PCE, OK, and PCK, respectively (Figs. 5.10c and E.2). In summary, the choice of the acceptance threshold needed to attain the behavioral set of a model depends on the availability of computational resources and the degree of accuracy desired, and must be estimated a priori, like the determination of hyperparameters.

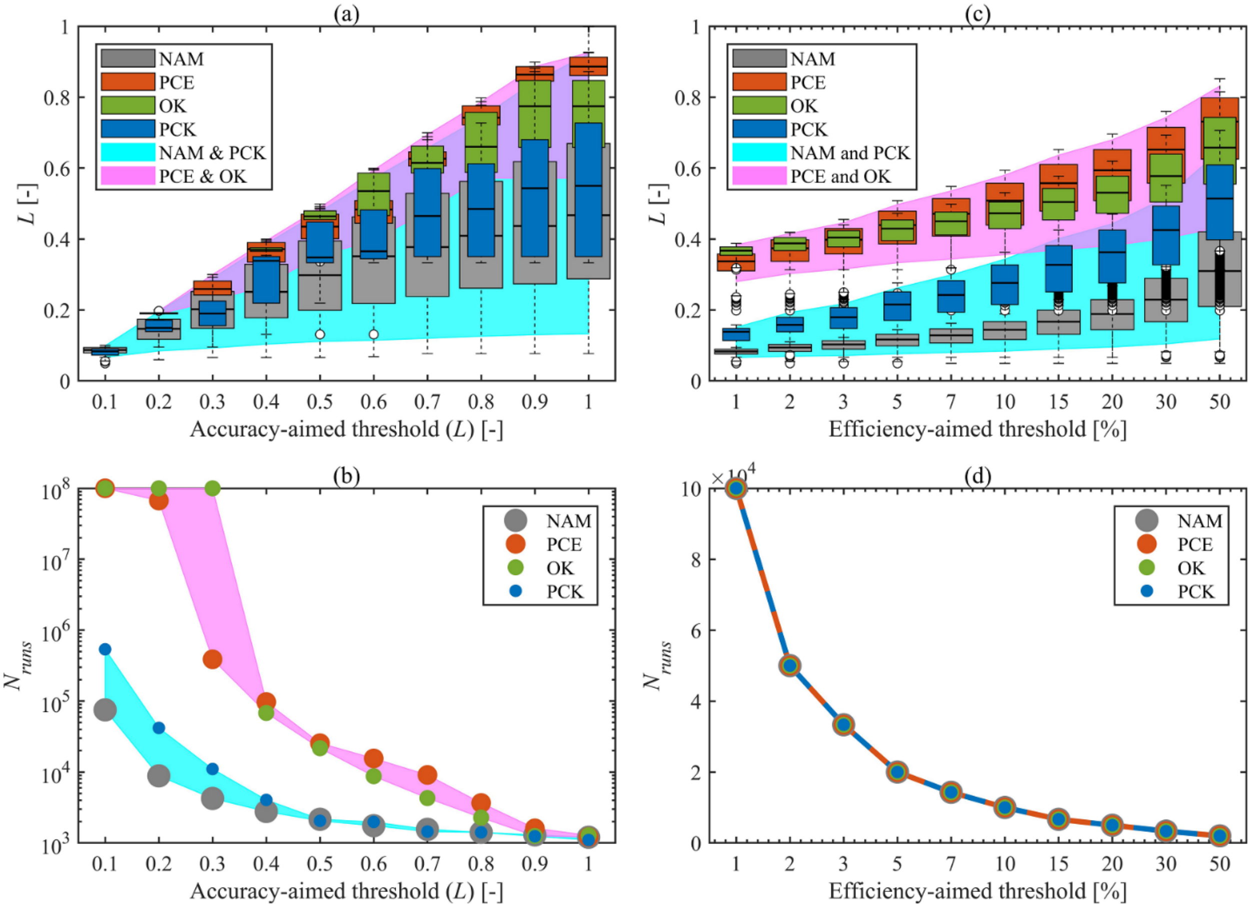


Figure 5.10. The effects of two types of acceptance thresholds (accuracy and efficiency-aimed thresholds) on accuracy (L) and efficiency (N_{runs}) performance for four models (NAM, PCE, OK, and PCK) over Event 8. The shaded areas in subplots (a) and (c) are drawn to highlight the difference between (cyan) NAM & PCK and (magenta) PCE & OK, representing the 90% confidence bands of 2,000 values of L. The boxplots demonstrate the median (central mark), the 25th and 75th percentiles (the edges of the box), and the maximum and minimum (the upper and lower whiskers) except for outliers (dot symbols) of 1,000 values of L. Subplots (b) and (d) denote the number of random runs (N_{runs}) required in the implementation of GLUE for each model.

5.5.3 Essentials of PS and superiority of PCK in uncertainty quantification

Can it be said that this surrogate model has excellent performance if its calculation speed is faster than that of its original model, or if its results are comparable with the original model? In the process of quantifying uncertainty through inverse modeling, simply comparing and evaluating computation speed and errors may fail to characterize a good surrogate model.

First, comparing the time taken for simple iterative tasks as well as to obtain the results of uncertainty quantification, it can be seen that the performances of the surrogate models considered here are significantly different. In this study, for simple iterations the computational speed was improved by about 500 times by PCE and 600 times by OK and PCK compared to NAM. This is because the CPU runtimes required for each single execution were approximately 2.5×10^{-4} sec for PCE to 2.0×10^{-4} sec for OK and PCK and about 0.12 sec for NAM. In the case of a single run or simple repeated runs, such an improvement stands on its own merit. However, when uncertainty needs to be quantified through inverse modeling (not limited to GLUE), the effort required to retrieve the behavioral set does not always offset and this speedup does not always follow. This is because the number of random runs (N_{runs}) required to obtain 1,000 behavior sets varies greatly, depending on the models or events used (Table 5.1). For example, for Event 4, the surrogate models have more random runs, while for Event 6, the original model does. In addition, the surrogate models PCE and OK, which were considered not able to adequately capture the original model, failed to achieve a single behavioral set even after 100 million random runs (see Events 7 and 8). In such events where uncertainty quantification is time-consuming, using the surrogate models PCE and OK to improve efficiency is of no benefit at all.

Table 5.1. The number of random model runs needed for obtaining 1,000 behavior sets when the accuracy-aimed threshold of 0.1 is used

Event	**NAM**	**PCE**	**OK**	**PCK**
1	27,992	77,883	49,901	24,662
2	484,687	852,775	149,766	405,614
3	213,861	363,550	111,483	78,862
4	261,666	2,303,546	2,490,675	1,045,565
5	30,936	27,051	498,553	50,583
6	6,057,083	4,621,272	3,078,314	264,375
7	631,209	–	–	9,386,783
8	75,237	–	–	535,150

'–' denotes no behavior set was obtained for 100 million random runs

Second, when determining an appropriate size for an experimental design by a conventional approach using a relative error, the performance of surrogate models may be misjudged. In general, the relative error of a surrogate model (ϵ in this work) decreases as the size of the experimental design increases, and the error tends not to decrease beyond a certain size (Fig. 5.11a). A surrogate model with a sufficiently small ϵ and guaranteed accuracy should be generated based on the experimental design of a certain size or larger (i.e., the elbows in Fig. 5.11a) [*Schöbi et al.*, 2017; *Tran and Kim*, 2021a]. But, securing the size of the design $\boldsymbol{\mathcal{X}}$ also increases the time ($RT_{\boldsymbol{\mathcal{X}}}$) required to construct the training set (from the experimental design $\boldsymbol{\mathcal{X}}$ to the model response $\boldsymbol{\mathcal{Y}}$). Moreover, the computational runtime (RT) of PCK and OK tends to be longer as N is larger (Fig. 5.11b), because it takes a substantial amount of time to compute the Gaussian variance in Eq. (5.3) for PCK and Eq. (D.4) for OK [*Razavi et al.,* 2012b; *Vigsnes et al.,* 2017]. This leads to a sharp drop in the performance of PCK, as evidenced by the increase of PS in Fig. 5.11c.

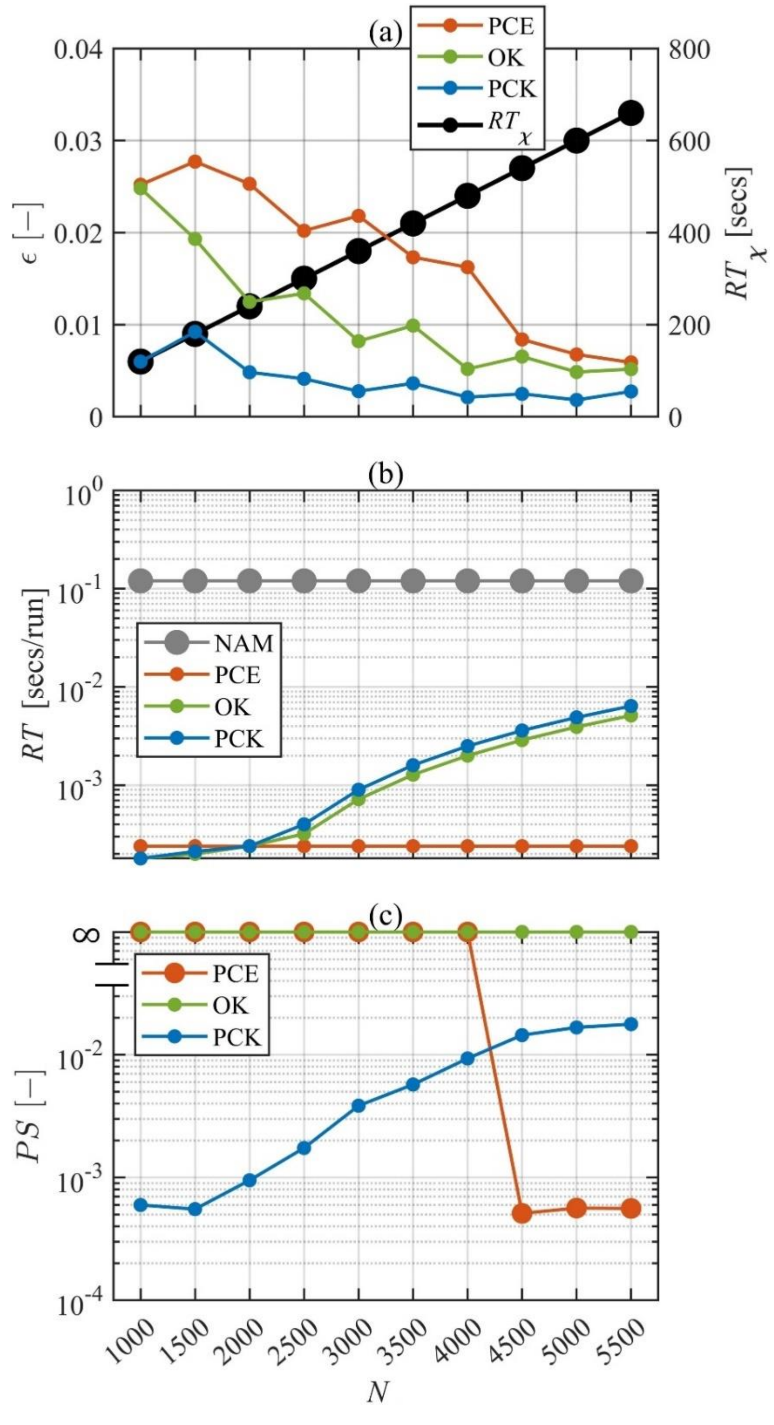


Figure 5.11. The effects of experimental design size (N) for Event 8 on (a) the leave-one-out error (ϵ) for QoI of y for three surrogate models and the runtime $RT_{\boldsymbol{\chi}}$ needed for generating experimental design of N; (b) the runtime RT per single run of NAM and three surrogate models; and (c) the performance score (PS) of three surrogate models with the accuracy-aimed threshold of 0.1.

In this study, Fig. 5.11a shows that to construct PCE, OK, and PCK to be tolerant of sufficiently small errors (e.g., $\epsilon = \sim 0.01$), the sizes of their experimental designs (i.e., the elbows in Fig. 5.11a) need to be about 4,500, 4,000, and 4,000, respectively. This is an agreed-upon standard from a traditional point of view for constructing a surrogate model. However, if uncertainty quantification is involved, this criterion may need to be changed. That is, its performance (efficiency and accuracy) in the process of uncertainty quantification must be reflected in the performance evaluation of the surrogate model. It can be seen that the result of Fig. 5.11c using the PS proposed in this study is very different from that of Fig. 5.11a. An experimental design with a size of ~4,500 for PCE and ~1,000 for PCK is required to exhibit sufficiently high performance ($PS = 0.001$), which is very different from the results of Fig. 5.11a. Interestingly, in the case of PCK, the size of the experimental design required was drastically reduced from 4,000 to 1,000. This indicates that one can build high-performing surrogate models by leveraging a much more limited training data size (Fig. 5.12), highlighting the superiority of PCK in that it helps modelers dramatically save computational resources.

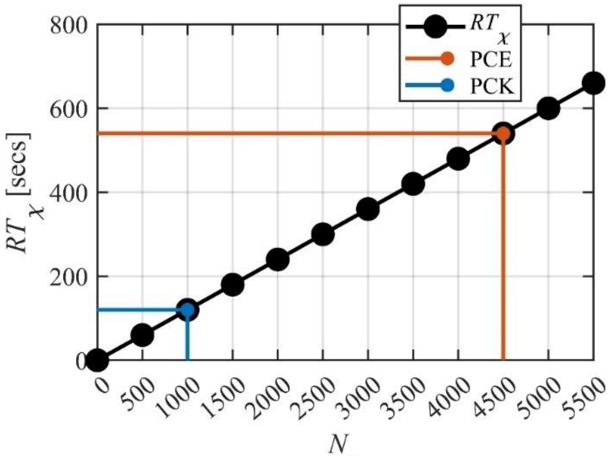


Figure 5.12. The runtime (RT_χ) needed for generating experimental design of N. RT_χ corresponds to N of PCE and PCK are indicated in red and blue.

5.5.4 Counsels for improving outlier performance of machine learning

Predicting outliers (extrapolation) seems to be a long-lasting challenge in applications of surrogate model as well as machine learning [*Kratzert et al.*, 2019; *Frame et al.,* 2021; *Nearing et al.*, 2021; *Tran et al.*, 2021; *Tran and Kim*, 2022]. In hypothesis, this would not be a challenge if the amount of data for training was sufficient and covers all possible, even low frequency-extreme events. However, in reality, it is difficult to collect such comprehensive observation data when targeting rare or exceptional phenomena. This study underlines three keys that have been carried out to improve the predictive power of extreme events that do not have enough data and deviate significantly from the training data. First, high-fidelity samples supervised by physical relationships as well as actual observations should be utilized to ensure sufficient learning when the number of training samples is small. Data generated by physics-based models or governing equations can improve the understanding of physical processes in machine learning models [*Ivanov et al.*, 2021]. The second suggestion is to enhance the extrapolation ability by extending the scope of the prediction space by additionally taking into account input noise and parameter (or learnable network weights) uncertainty [*Fang et al.*, 2020; *Abdar et al.*, 2021]. In this study, model parameters were considered as uncertain input vectors, and GLUE was used to quantify their uncertainty. The last suggestion is to build a hybrid model by combining a predictive model with a model with extrapolation capabilities. In this study, PCE, a global model that plays the role of trend, and kriging, an interpolation model that computes local changes, are combined. Instead of using PCE, other techniques that can increase extrapolation capabilities would be applied such as Richardson extrapolation [*Bach*, 2020], spectral mixing kernel [*Wilson et al.*, 2014], extrapolation algorithms [*Bakas*, 2019], sparse identification of nonlinear dynamics [*Champion et al.*, 2019], and generative models [*Hatakeyama-Sato and Oyaizu*, 2021]. In this work, we used PCK with the three aforementioned approaches to obtain satisfactory results for extreme values. Such a discovery will inspire novel designs of potentially more comprehensive hybrid models.

5.6 Conclusions

This Chapter presents a new surrogate model, PCK, that can not only efficiently quantify streamflow uncertainty, but also accurately predict even extreme events that deviate significantly from the trained data space. To enhance the extrapolation capability, this study underlines three keys: enriching the understanding of physical processes by establishing high-fidelity training samples supervised by physical relationships; broadening the scope of the prediction space by additionally taking into account input noise and parameter uncertainty; and creating a new hybrid model (e.g., PCK) that combines a local predictive model with a model with extrapolation capabilities. The advantages of PCK were confirmed by investigating how well the results of GLUE matched observations for eight testing flood events in the Thu Bon watershed compared to two well-known surrogate models, PCE and OK. The principal results of this study are summarized as follows.

First, with a relatively small-sized experimental design ($N_{\mathcal{X}} = 1{,}000$), PCK outperforms PCE and OK in mimicking the behavior of the original model with smaller leave-one-out error (ϵ) values for all QoIs. Also, from the sensitivity analysis of nine parameters, the sensitivity results of PCK were closer to those of NAM especially for the three most sensitive parameters (CQOF, CK12, and Lm) than were PCE and OK.

As a result of applying GLUE to the eight test events for the three surrogate models trained on the identical dataset (see the framework in Fig. 5.1), all of the surrogate models provided predictions equivalent to the original model for six smaller events that were similar to the training data space. However, for extreme Events 7 and 8, which differed significantly from the training experimental design, only PCK was found to accurately predict the hydrograph and flood peaks, regardless of the type of acceptance threshold (accuracy- or efficiency- aimed), while both PCE and OK failed.

The simulation results in Sections 5.4.3 and 5.4.4 confirmed that both types of acceptance thresholds had a significant impact on the performance of GLUE. Selecting the accuracy-aimed threshold can ensure the consistent accuracy of uncertain simulations, but can lead to a computational burden because of the substantial number of repeated runs needed. In contrast, the efficiency-aimed threshold can control the computational efficiency with a predefined number of random runs, but can only provide ordinary performance. Since each has its own pros and cons, from a practical point of view, the type and size of the threshold should be determined based on the availability of sufficient computational resources and the degree of accuracy needed.

The performance of a surrogate model cannot be said to be superior just because its calculation speed is faster than that of its original model or because its calculation results are comparable to the original's. In the process of quantifying uncertainty, the computation speed and outcomes of a surrogate model may deteriorate. Thus, merely comparing and evaluating computational speed and error in a traditional way can lead to a misjudgment in selecting a good surrogate model. In this study, we propose a new "performance score" (PS) that can measure the overall performance (including both accuracy and efficiency) of a surrogate model compared to its original model. This PS allows for assessing the actual achievement that arises in quantifying uncertainty, and helps to efficiently construct surrogate models and save computational budgets by limiting unnecessary increases in their experimental design sizes.

The combined surrogate model presented here can not only predict events that deviate significantly from the trained data space, but can greatly reduce the computational burden of uncertainty quantification. Such a discovery will ultimately inspire novel designs of potentially more comprehensive surrogate models.

CHAPTER VI

Research summary and future efforts

> "Remember to celebrate milestones as you prepare for the road ahead"
>
> - *(Mandela, N)*

6.1 Summary of research

Extreme floods occur more frequently than in the past due to global warming, and they have more profound socio-economic impacts [*Hirabayashi et al.*, 2013; *Winsemius et al.*, 2015]. Flood forecasting is an important component of flood risk management and mitigation but is subject to multiple uncertainties caused by meteorological inputs, initial states, model structures, and model parameters [*Beven*, 1989; *Ajami et al.*, 2007; *Moradkhani and Sorooshian*, 2008; *Mockler et al.*, 2016]. We have to reduce the uncertainties in some optimal fashions to get robust reliability of the flood predictions and to mitigate the flood damage. In recent years, numerous research efforts investigated the uncertainties in the tasks of flood prediction. However, at present we entirely lack comprehensive studies that can handle long-lasting challenges of computational burden, inaccuracy, and unreliable predictability in real-time ensemble flood forecasting with uncertainty quantification. This dissertation aims to gain comprehensive knowledge of building novel modeling frameworks for computationally efficient and accurate real-time ensemble flood forecasting with uncertainty quantification.

Chapter II presents a unified uncertainty quantification modeling framework that the GLUE framework was revisited and combined with a PCE surrogate model that is employed to offset the computational demands for numerous repeated calls of the model evaluation. It provides the benefits of an interpretable, probabilistic framework on which to make inferences about the drivers of model behavior, as well as the sensitivities of the model's output to the uncertain inputs. The central conclusions of Chapter II are resumed as follows: (1) the subjective aspects of GLUE (e.g., the cutoff threshold values of likelihood function) were investigated and reasonably optimized through two indices that represent for model accuracy and efficiency (i.e., AI and EI). Also, the number of ensemble behavioral sets was specified to maintain the sufficient range of uncertainty but to avoid any unnecessary computation. (2) The results computed using a PCE model with polynomial bases are as good as those given by the NAM and SFM, while the total amount of time required for making an ensemble in the PCE model are approximately 17 and 200 times faster, respectively. (3) Identification of the posterior parameter distributions from the calibration process helps to find the behavioral sets even faster. (5) The construction of a surrogate model becomes more advantageous with the use of sparse polynomial chaos expansion (SPCE) coupled with the least angle regression (LAR) method. Specifically, SPCE outperforms the full polynomial chaos expansion (FPCE) built by a well-known, ordinary least square regression (OLS) method with a more accurate surrogate model (i.e., smaller leave-one-out cross-validation error) and smaller size of experimental design.

Forecasting results should be provided within a predetermined time horizon and accurate enough to help the preparation and mitigation of flood damages. Many approaches have been focused on real-time problems of model accuracy, predictability, and efficiency with uncertainty quantification. Yet, there is still a lack of modeling framework that can comprehensively solve these aforementioned problems simultaneously [*Liu et al.*, 2012; *Cintra and Velho*, 2018]. Chapter III outlined one of the primary goals of this dissertation that is the development of a novel model framework that simultaneously improves accuracy, predictability, and computational efficiency for real-time ensemble flood forecasting. This framework provides a holistic, robust approach to accounting and understanding the uncertainties of hydrological parameters and vastly reducing the computational burden of ensemble simulations. It embraces the benefits of three modeling techniques integrated together for the first time: surrogate modeling, parameter inference, and data assimilation. The use of PCE surrogates significantly decreases computational time. Parameter inference (GLUE) allows for model faster convergence, reduced uncertainty, and superior accuracy of simulated results. EnKFs assimilate errors that occur during forecasting. To examine the applicability and effectiveness of the integrated framework, 18 approaches were developed according to how surrogate models are constructed, what type of parameter distributions are used as model inputs, and whether model parameters are updated during the data assimilation procedure.

The essential strengths of the modeling framework described in Chapter III are as follows, (1) PCE must be built over various forcing and flow conditions, and in contrast to previous studies, it does not need to be rebuilt at each time step. (2) Model parameter specification that relies on posterior information of parameters (so-called *Selected* specification) can significantly improve forecasting performance and reduce uncertainty bounds compared to *Random* specification using prior information of parameters. (3) No substantial differences in results exist between single and dual ensemble Kalman filters, but the latter better simulates flood peaks. The use of PCE effectively compensates for the computational load added by the parameter inference and data assimilation (up to ~80 times faster). Therefore, the presented approach contributes to a shift in modeling paradigm arguing that complex, high-fidelity hydrologic and hydraulic models should be increasingly adopted for real-time and ensemble flood forecasting.

Data assimilation plays an essential role in real-time forecasting but demands repetitive model evaluations given ensembles. To address this computational challenge, a novel, robust and efficient approach to surrogate data assimilation was presented in Chapter IV. Here we further exploited the power of surrogate approaches to form new surrogate filters by replacing the internal processes of the EnKFs with PCE surrogates. Eight types of surrogate filters, which can be characterized according to their different surrogate structures, building systems, and assimilating targets, are proposed and validated. To compensate for the potential shortcomings of the existing sequential experimental design (SED), an advanced optimization scheme, named sequential experimental design-polynomial degree (SED-PD), is also advised. Its dual optimization system resolves the issue of SED by which the value of the polynomial degree had to be selected ad-hoc or by trial and error; its multiple stopping criteria ensure convergence even when an accuracy metric does not monotonically decrease over iterations. A comprehensive investigation into how to configure a surrogate filter indicates that the new partial (replacing part of original filters) and invariant (valid for entire time periods) approaches are preferred in terms of accuracy and efficiency, which helps directly reduce the number of dimensions and bridge the gap between hindcasting and real-time forecasting. Of the eight filters, the Dual Invariant Partial filter performs best, with equivalent accuracy to Dual EnKF and about 500 times greater computational efficiency. Ultimately, this proposed surrogate filter will be a promising alternative tool for performing computationally-intensive data assimilation in high-dimensional problems.

Another crucial concern in surrogate applications is the capability of providing a reliable prognosis for outliers (or extremes) beyond the training data space, although they generally have excellent predictive power for regions within the training data. Especially in the context of floods, there will be a high probability of extreme events that, due to climate change, have not been experienced in the past. There is also the possibility that extreme events that deviate from recorded events will occur due to climate internal variability, even assuming that climate stationarity is maintained. Therefore, a common solution for ensuring the predictive power in the entire data space is to expand the data range of the design site to cover all possible cases. However, obtaining sufficient collections of extreme events for training is unfeasible, so one needs to develop an alternative solution to ensure predictability for the events beyond the data space. Therefore, for this purpose, Chapter V introduced a new surrogate model named polynomial chaos-kriging (PCK) that combines the advantages of two well-known surrogate models, polynomial chaos expansion (PCE) and kriging. This combination enabled streamflow prediction for extreme events that deviated significantly from the trained data space, and allowed for quantifying predictive uncertainty robustly and efficiently. The uncertainty quantification results to eight test flood events through a modeling framework that applies GLUE to surrogate models confirm that (1) PCK outperformed PCE and OK (ordinary kriging) in mimicking behaviors of the original model (i.e., smaller leave-one-out error and closer parameter sensitivity) with a smaller-sized training dataset; (2) three surrogate models trained on the identical dataset exhibited equivalent predictability with the original model for six smaller events similar to their training data space. However, for two extreme events, which differed significantly from the training set, only PCK was found to accurately predict the hydrograph and flood peaks, while both PCE and OK failed. Additionally, a new “performance score” is proposed here to assess the overall performance (including both accuracy and efficiency) of the surrogate models. This compensates for situations in which the performance of a surrogate model can be misjudged through individual indices of efficiency or accuracy in the process of uncertainty quantification. Our findings will ultimately inspire novel designs toward a potentially more comprehensive surrogate model

Furthermore, in Chapter V, the effects of the acceptance threshold types on the model accuracy and efficiency were investigated and discussed. Specifically, selecting the accuracy-aimed threshold can ensure the consistent accuracy of uncertain simulations, but can lead to a computational burden because of the substantial number of repeated runs needed. In contrast, the efficiency-aimed threshold can control the computational efficiency with a predefined number of random runs, but can only provide ordinary performance. Since two types of acceptance thresholds, defined here as “accuracy-aimed” or “efficiency-aimed” threshold, have their own pros and cons, the type and size of the threshold should be determined depending on the availability of computational resources and the degree of accuracy needed.

6.2 Critical assumptions, limitations and future studies

Any modeling work inherently contains a number of assumptions or limitations. Critical assumptions underlying this dissertation are given here.

- Rainfall uncertainty: The rainfall uncertainty of future flood events has not really been accounted for in Chapters III and IV. Although in Chapter IV, the uncertainty of rainfall was assumed following a log-normal error distribution with a relative error of 25% for the observed rainfall, forecasted rainfall is actually worth noting and plays a vital role in accurately diagnosing flood forecasts of developed modeling framework. So to extend the framework to general applications, forecasted rainfall from Numerical Weather Prediction (NWP) models need to be carefully considered.

- Model structure uncertainty: This framework only accounts for parameter/input uncertainty, but no model structure uncertainty. Every hydrologic model will have its own sets of assumptions and limitations. Although simplification of complex physical processes of rainfall-runoff improves computational efficiency, it also lowers the accuracy of the representation of the real-world in the model as a trade-off. The quantification of model structural error is necessary and an active area of future research.
- Distribution of uncertain parameters: Many distribution types (e.g., uniform, Gaussian, Gamma, Beta) could be chosen as the prior distribution for uncertain parameters. In this dissertation, the uniform distribution was selected for all experiments as it was preferred in prior studies. However, it is still remarkable. The importance of prior distribution was highlighted that it has a significant influence on the quantification of parameter uncertainty as well as determination of the posterior parameter [*Mandel and Beezley*, 2009; *Zhang et al.*, 2018a]. So one needs to be aware of this configuration during experiment setup.
- High-dimensional problems: Modeling frameworks in this dissertation were applied to conceptual rainfall-runoff models of relatively low dimension. As the discussion in Section 3.6.5, these frameworks can in theory be applied to higher-dimensional systems (e.g., distributed physical-based models), but it would be worthwhile to explore their applicability and transferability in further studies.

Appendices

A Supplementary Materials for Chapter II

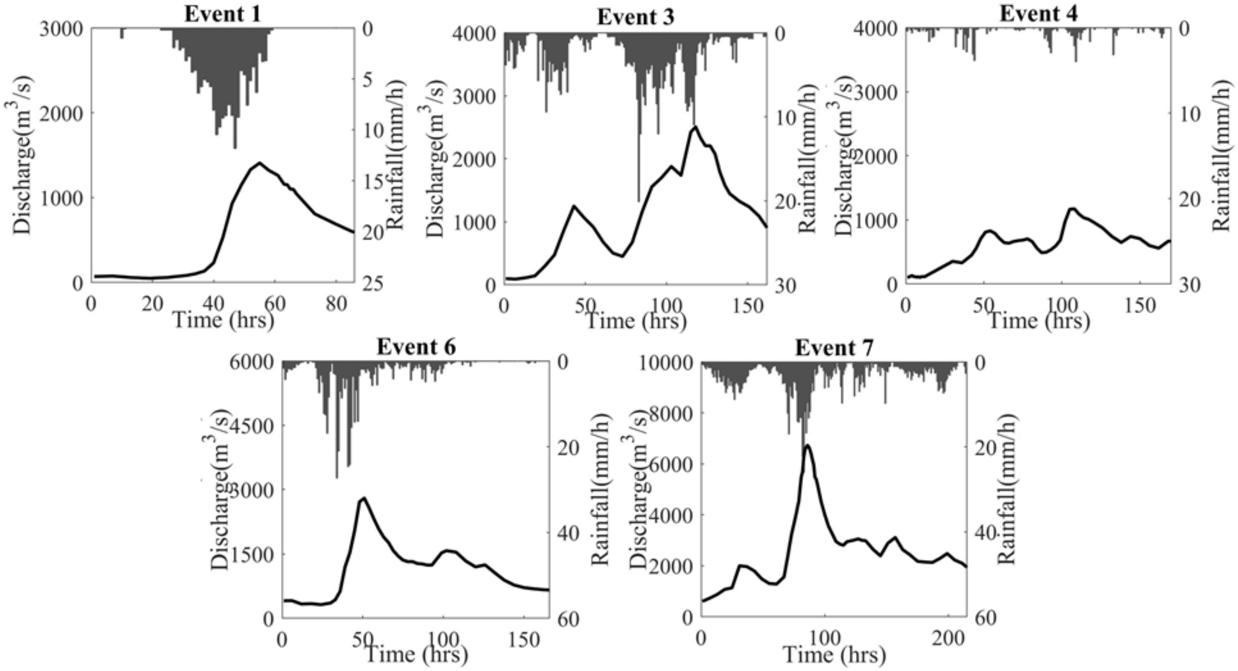


Figure A.1. Same as Fig. 2.4 except for the remaining events.

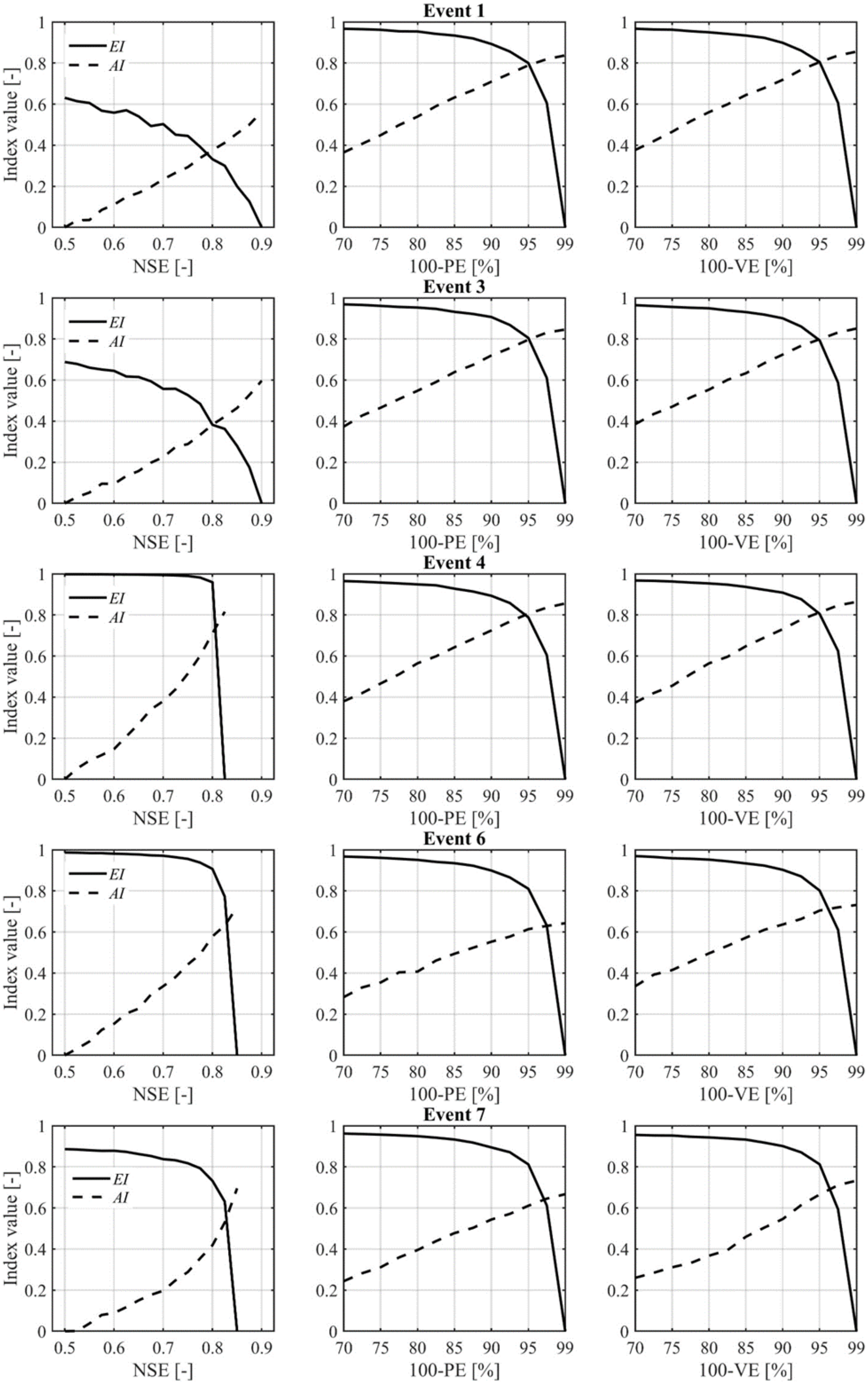


Figure A.2. Same as Fig. 2.5 except for the remaining events.

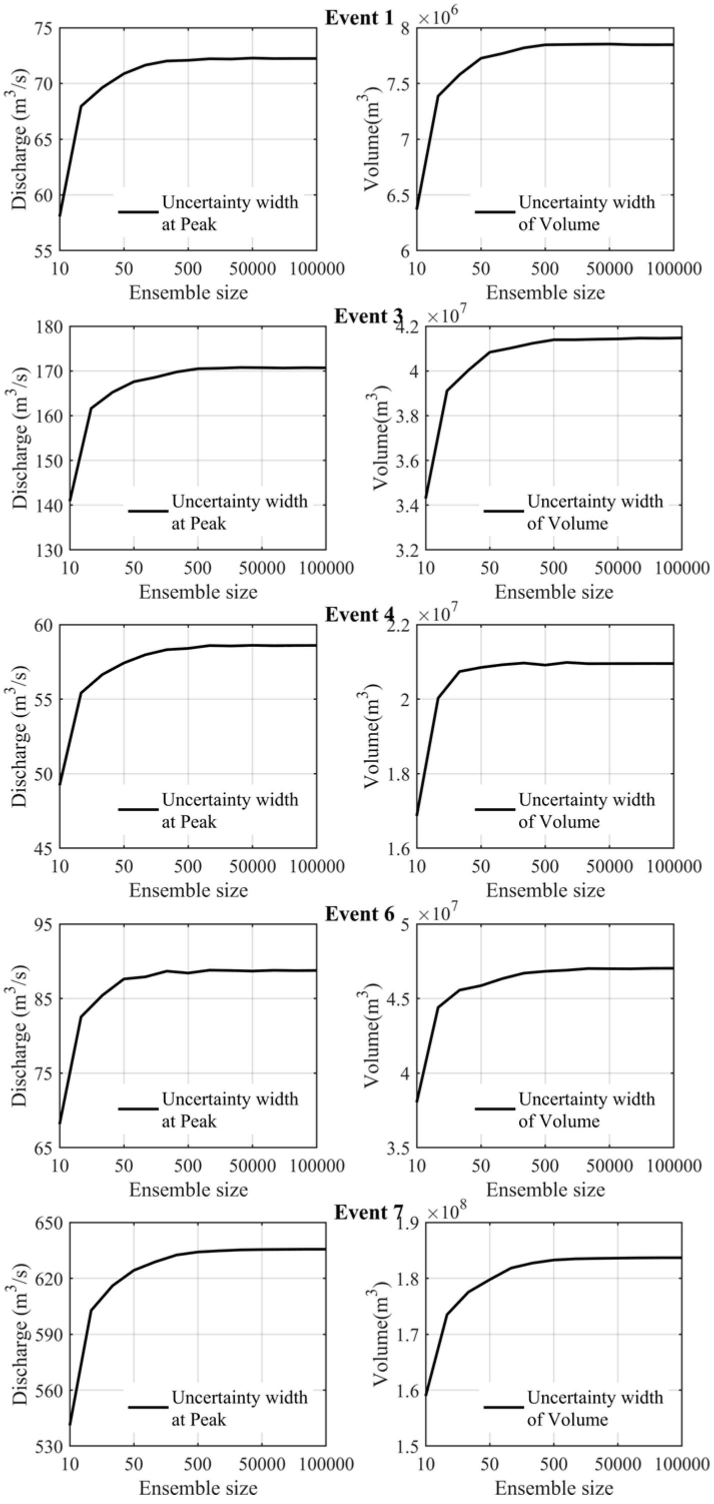


Figure A.3. Same as Fig. 2.6 except for the remaining events.

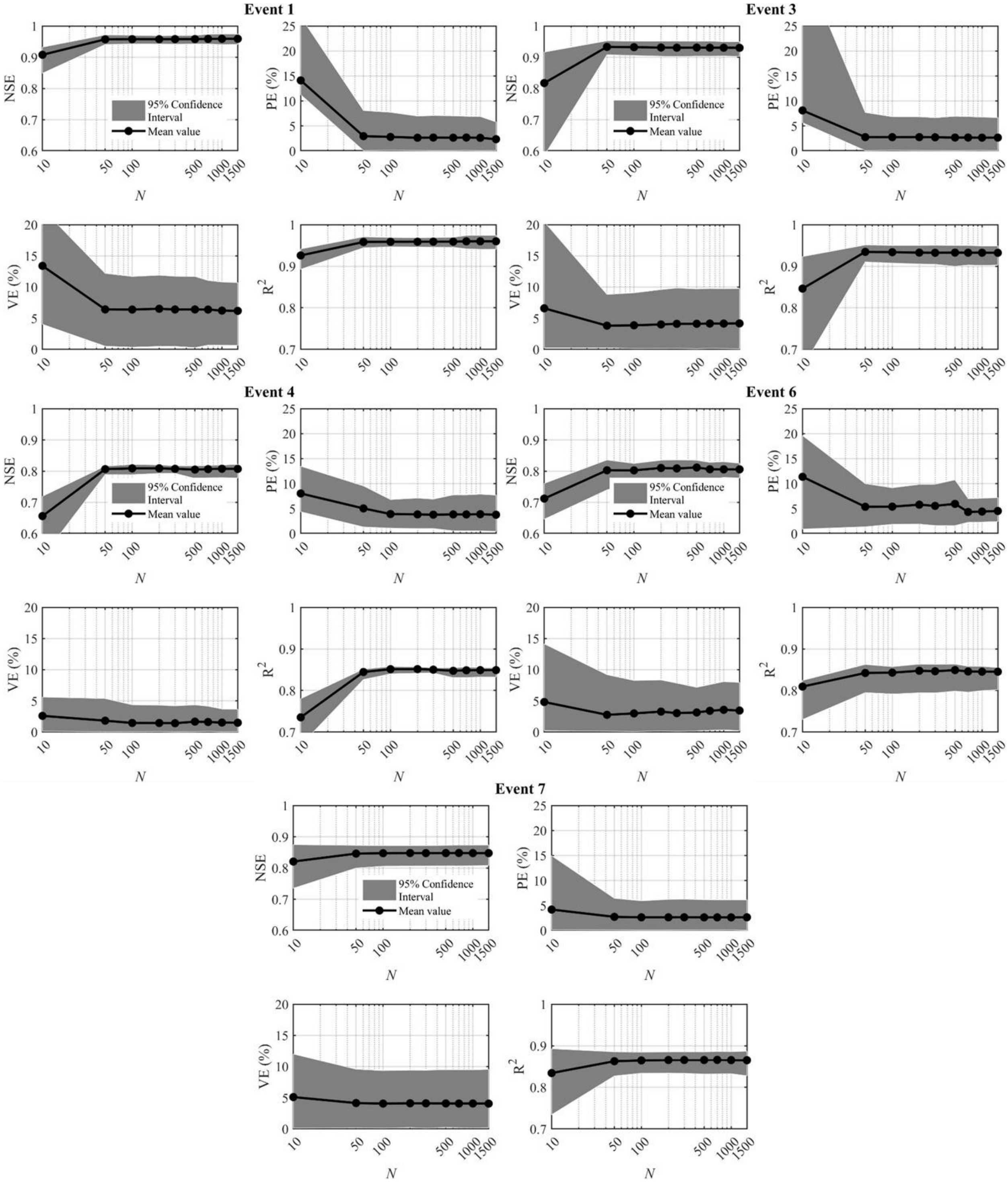


Figure A.4. Same as Fig. 2.7 except for the remaining events.

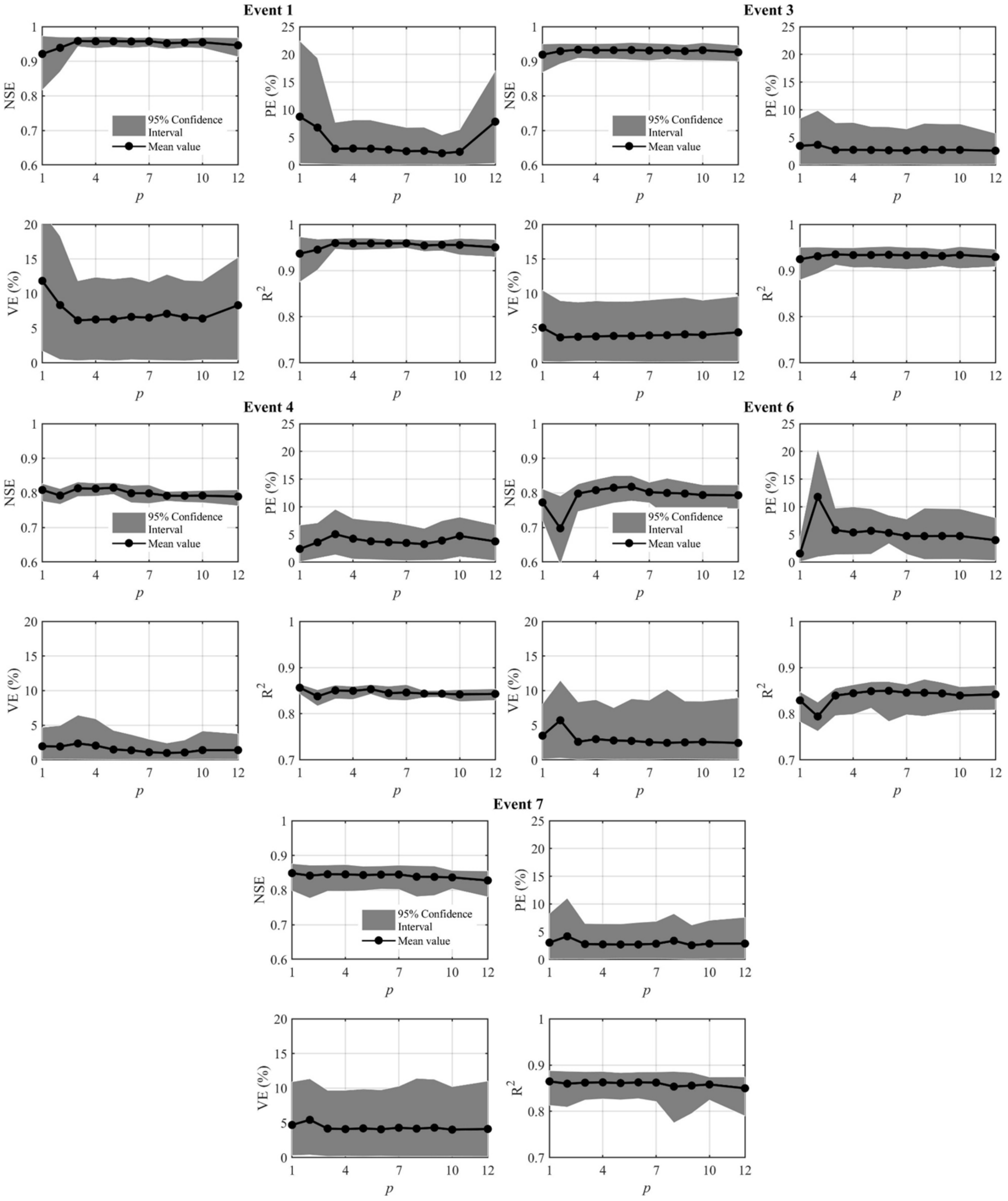


Figure A.5. Same as Fig. 2.8 except for the remaining events.

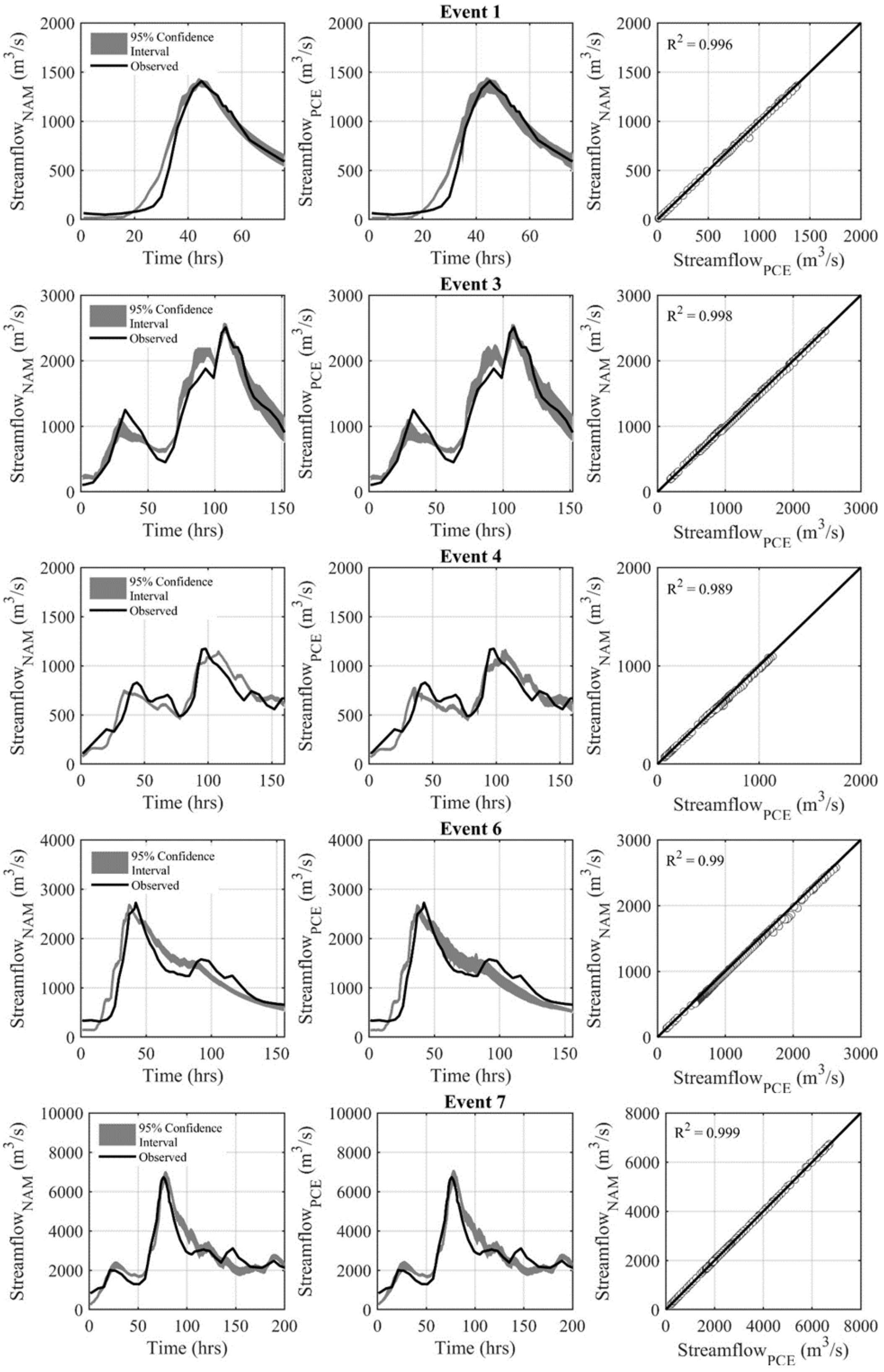


Figure A.6. Same as Fig. 2.9 except for the remaining events.

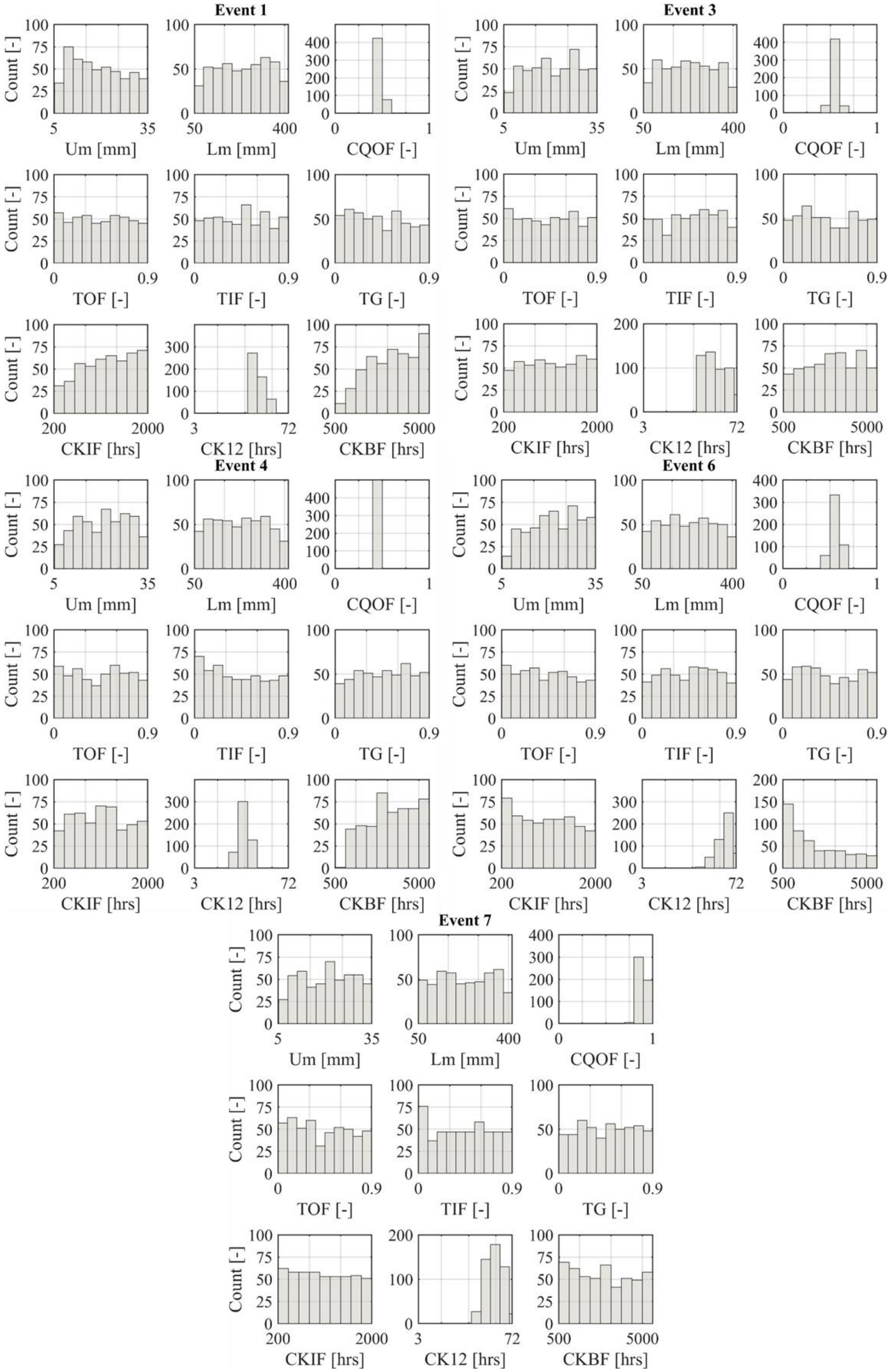


Figure A.7. Same as Fig. 2.10 except for the remaining events.

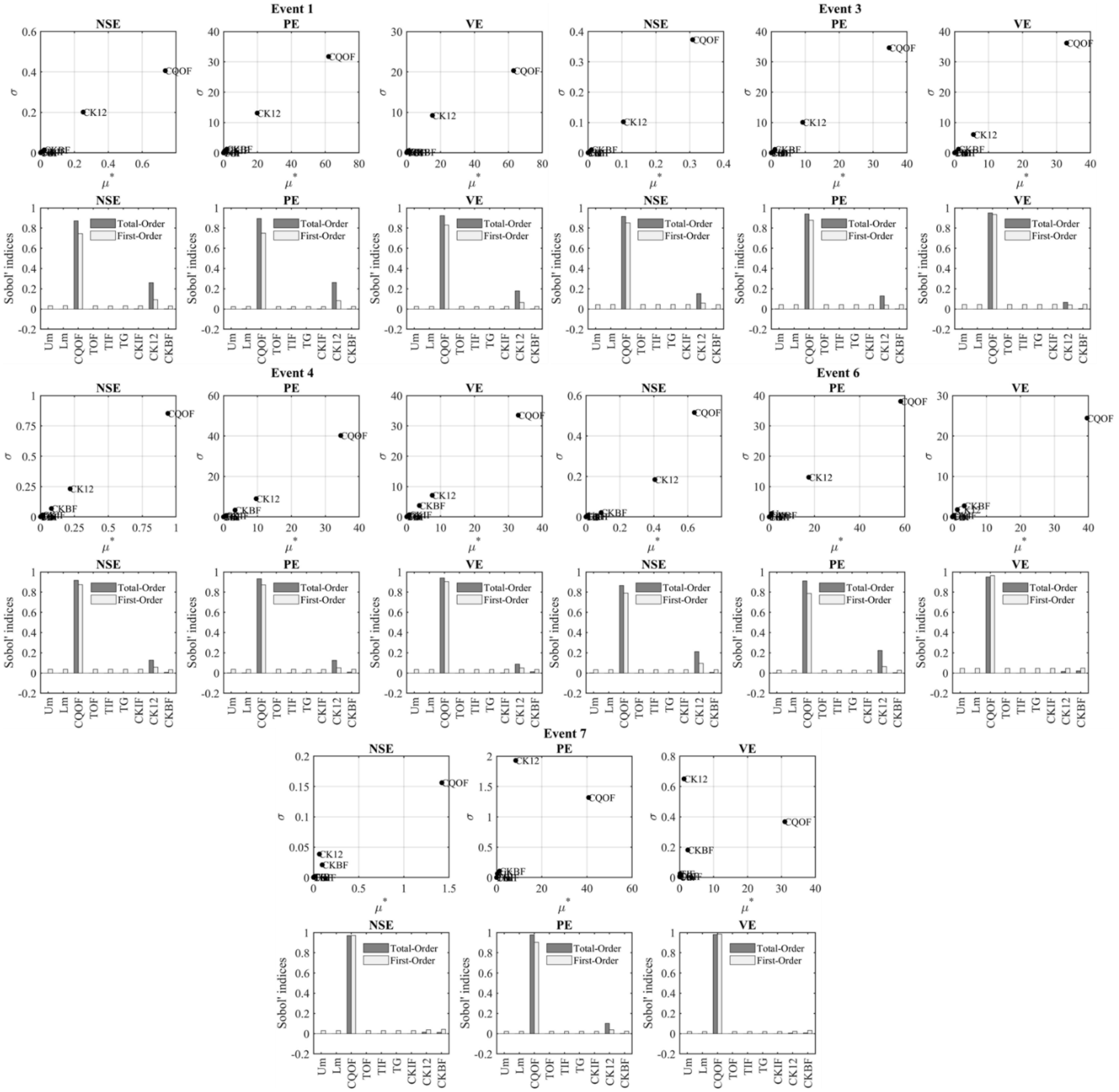


Figure A.8. Same as Fig. 2.11 except for the remaining events.

B Comparisons between time-invariant and time-variant PCE models

To compare the accuracy and efficiency between time-invariant and time-variant PCE models, a time-variant PCE (called hereafter PCE-III) is constructed with the experiment design, N of 1000 and polynomial degree, p of 3 (similar to PCE-I).

The total runtime (*TRT*) of PCE-I (as well as PCE-II and NAM) was written in Eq. (3.38) of the Chapter III. It computes the total runtime over the entire period including the warm-up, calibration, and forecasting periods. Eq. (3.38) is

$$\begin{aligned} TRT = \left(RT_{w+c,\boldsymbol{Model}} \times fac_{\boldsymbol{Model}} + RT_{f,\boldsymbol{Model},DA}\right) \times n \\ + RT_{build,\boldsymbol{Model}} \end{aligned} \tag{B.1}$$

Since the time-variant PCE should be newly built at every computational time step, the above equation can be rewritten to consider the runtime of building $\boldsymbol{Model}$ ($RT_{f,build,\boldsymbol{Model},t}$) and the runtime of performing DA ($RT_{f,\boldsymbol{Model},DA,t}$) at each time step for an ensemble member over the forecasting period:

$$\begin{aligned} TRT = \left(RT_{w+c,\boldsymbol{Model}} \times fac_{\boldsymbol{Model}} + RT_{f,\boldsymbol{Model},DA,t} \times \mathrm{T}\right) \times n \\ + (RT_{w+c,build,\boldsymbol{Model}} + RT_{f,build,\boldsymbol{Model},t} \times \mathrm{T}) \end{aligned} \tag{B.2}$$

This Eq. (B.2) is expressed as a linear form of an independent variable, n given T. If n is determined, this equation can be also expressed as a linear form of T.

$$\begin{aligned} TRT = \left(RT_{f,build,\boldsymbol{Model},t} + RT_{f,\boldsymbol{Model},DA,t} \times n\right) \times \mathrm{T} \\ +(RT_{w+c,\boldsymbol{Model}} \times fac_{\boldsymbol{Model}} \times n + RT_{w+c,build,\boldsymbol{Model}}) \end{aligned} \tag{B.3}$$

Note that Eq. (B.1) to (B.3) are all identical. For PCE-I and PCE-III, respectively, results of $RT_{f,build,\boldsymbol{Model},t}$ are 0 and 3.1 secs; results of $RT_{f,\boldsymbol{Model},DA,t}$ are 1.47e-4 and 1.26e-4 secs. Such a fast computation time is because it is a polynomial calculation; $RT_{w+c,\boldsymbol{Model}}$ is the runtime required to run one simulation of $\boldsymbol{Model}$ over the warm-up and calibration periods and these are 0.0508 and 0.0436 secs; results of $fac_{\boldsymbol{Model}}$ are 27 and 76; $RT_{w+c,build,\boldsymbol{Model}}$ is the runtime required to build $\boldsymbol{Model}$ over the warm-up and calibration periods and these are 541.2 and 1069.5 secs.

First, using Eq. (B.2), TRT is compared for PCE-I vs. PCE-III (A19 to A24) with respect to varying n given the original T of 40. The left plots of Figure B.1 show that for both *Random* and *Selected*, PCE-I is more efficient than PCE-III and the larger the number of ensemble, the better, e.g., approximately 2-3 times faster.

Second, using Eq. (B.3), TRT is compared for PCE-I vs. PCE-III (A19 to A24) with respect to varying T given the ensemble size of 1000. The right plots of Figure B.1 show that PCE-I is also more efficient than PCE-III and the larger the number of time steps, the greater the efficiency (note that the slope refers to the efficiency of PCE-I by about 22, 11, and 8 times for ‘None’, ‘EnKF’, and ‘Dual EnKF’, respectively).

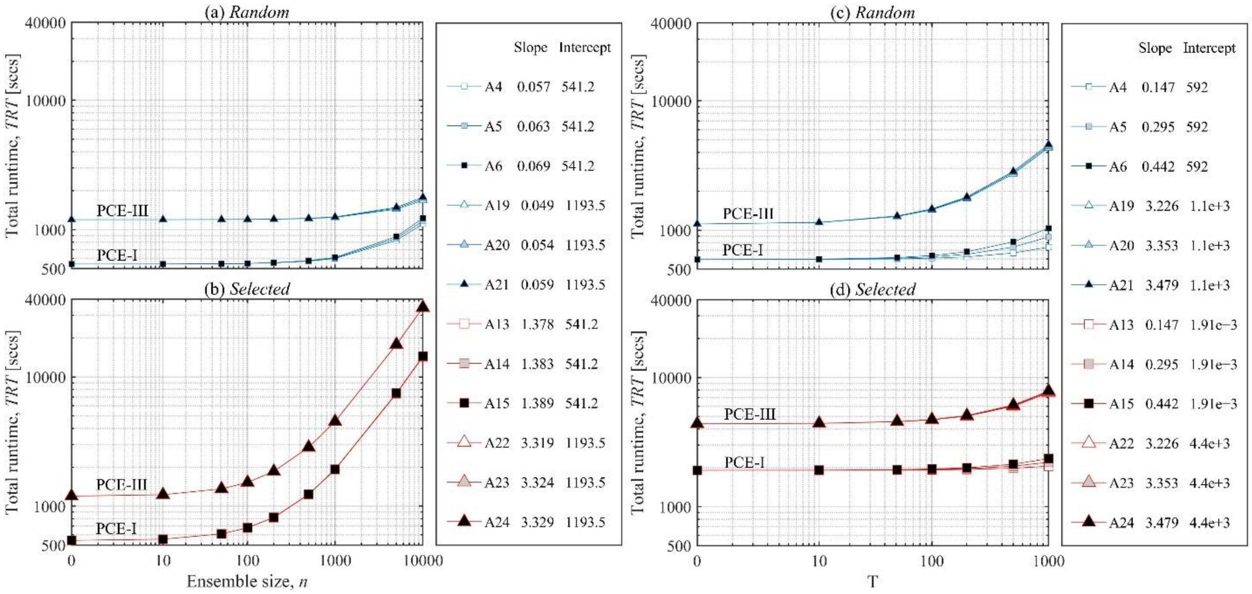


Figure B.1. The total runtime (TRT) with (left) varying the ensemble sizes (n) and (right) varying the number of time steps over the forecasting period (T)

To compare the accuracy between PCE-I and PCE-III, Figs. 3.10 to 3.12 in the manuscript is re-drawn for PCE-III in Figs. B.2 to B.4. The latter figures show that the results of PCE-III are also satisfactory enough, but the results of PCE-I are slightly better than those of PCE-III.

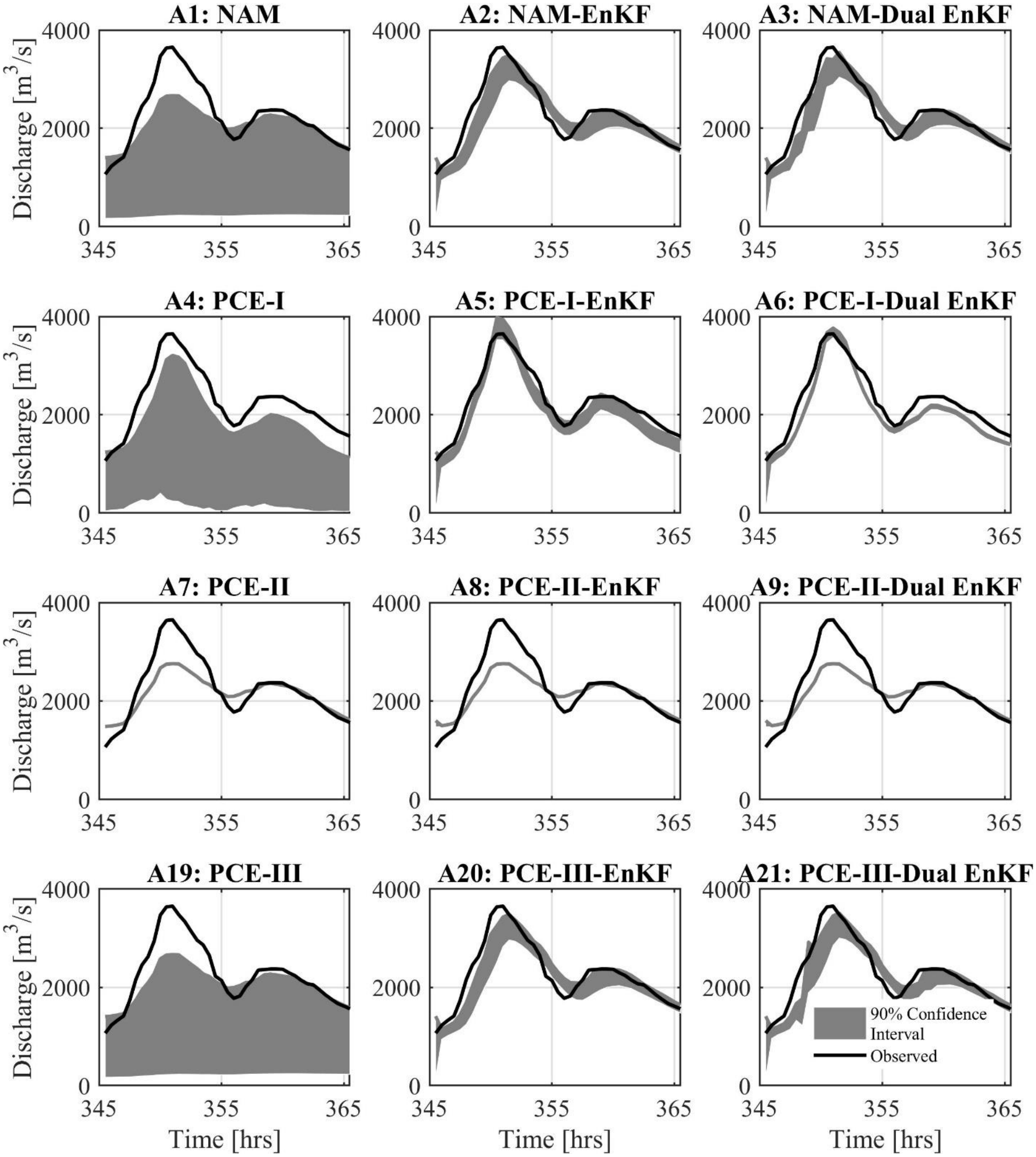


Figure B.2. Hydrographs over the forecasting period using NAM, PCE-I, and PCE-III, with a 90 % confidence interval of 500 *Random* model runs.

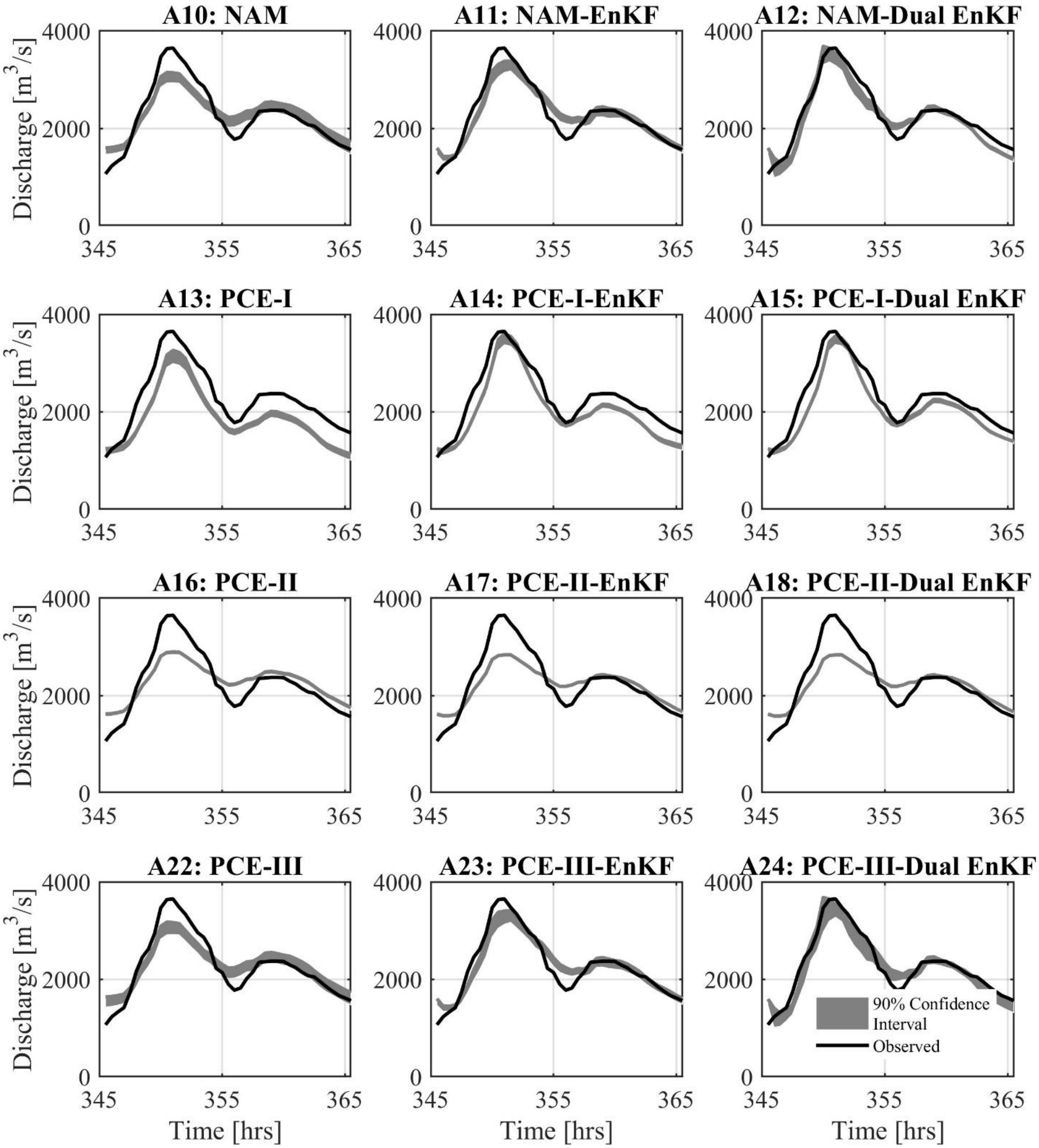


Figure B.3. Hydrographs over the forecasting period using NAM, PCE-I, and PCE-III, with a 90 % confidence interval of 500 *Selected* model runs.

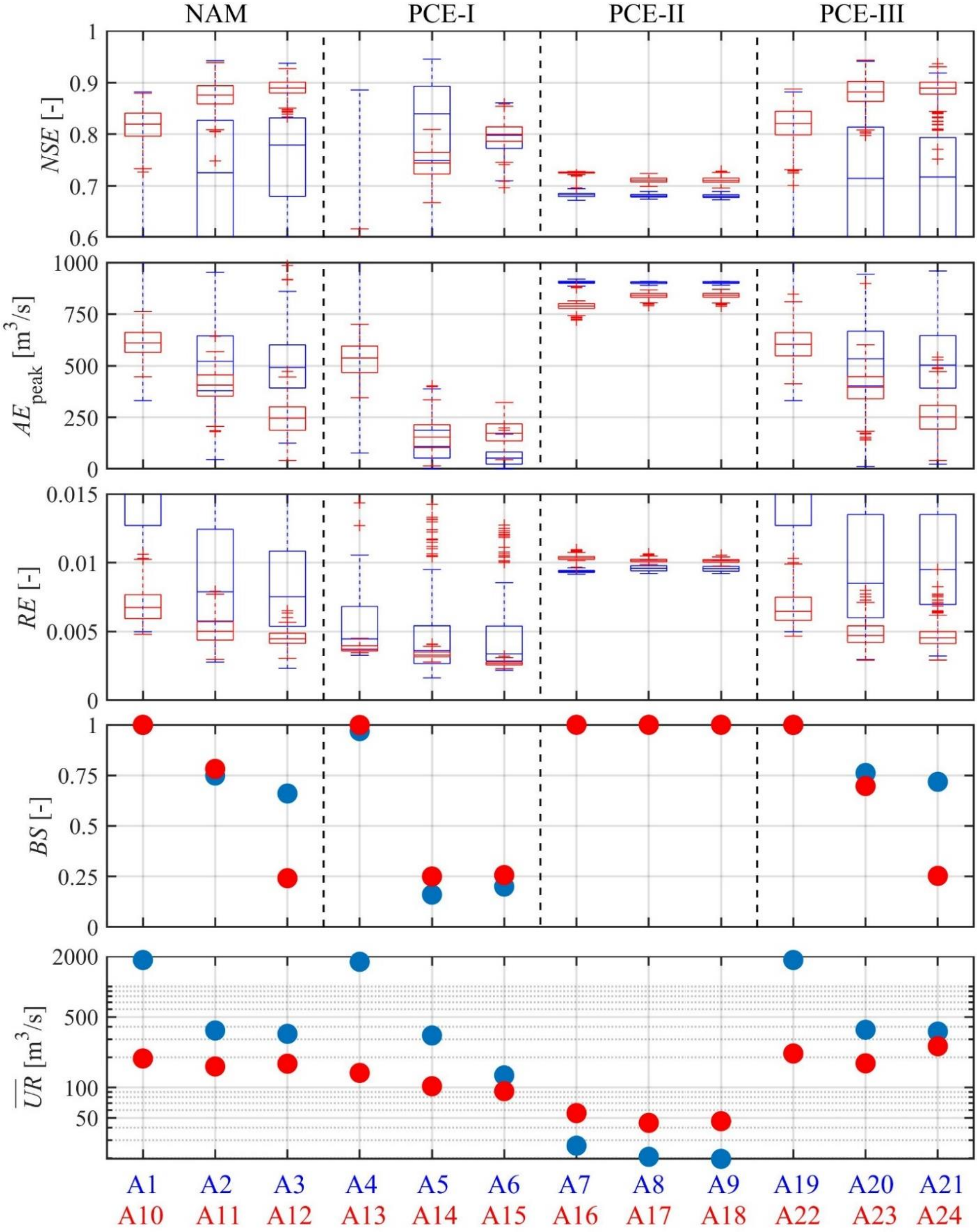


Figure B.4. The performance metrics reflecting accuracy and predictability for the forecasting period using NAM, PCE-I, and PCE-III.

C Supplementary Materials for Chapter IV

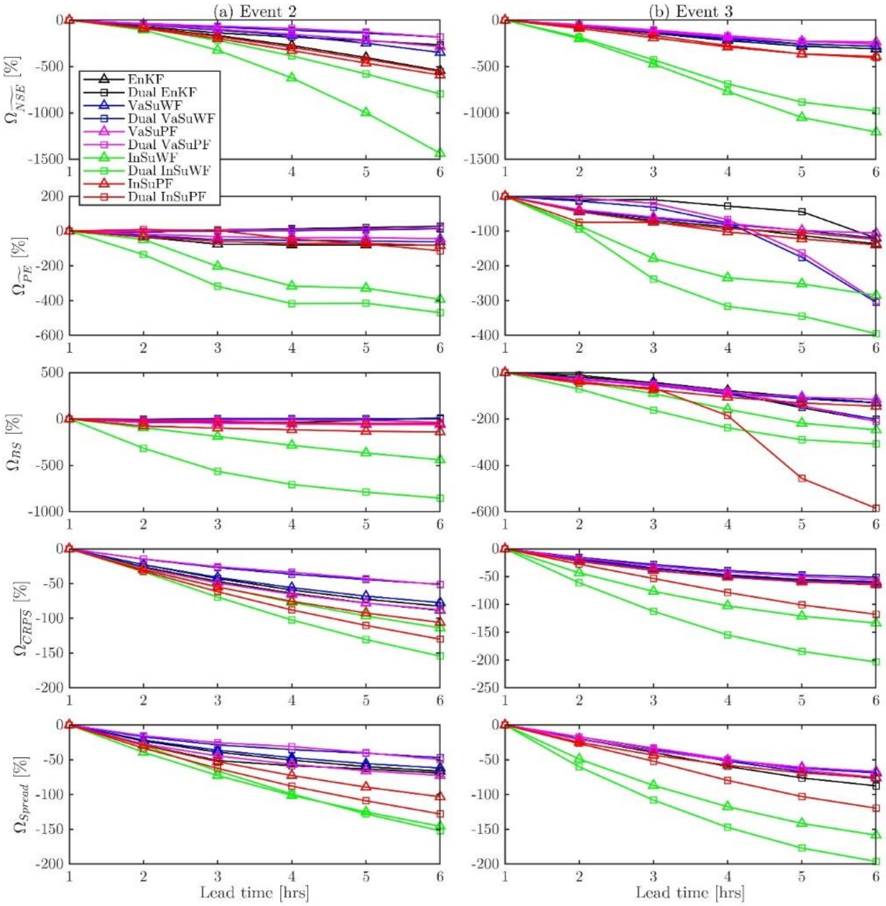


Figure C.1. The comparison of a relative "difference" metric (Ω) in Eq. (C.1) for the five evaluation metrics (five subplots in column) with different lead times from 1 to 6 hours (in x – axis) over (a) Event 2 and (b) Event 3. These values in each subplot were computed based on the metric value at lead time (LT) of 1 hour. The negative values of Ω indicate the degree of performance deterioration as compared to the value at LT = 1.

$$\Omega = \frac{|\mathrm{Metric}(\mathrm{SuFs}, LT = 1) - \mathrm{Metric}_{\mathrm{ideal}}| - |\mathrm{Metric}(\mathrm{SuFs}, LT) - \mathrm{Metric}_{\mathrm{ideal}}|}{|\mathrm{Metric}(\mathrm{SuFs}, LT = 1) - \mathrm{Metric}_{\mathrm{ideal}}|} \times 100 \qquad \text{(C.1)}$$

where $\mathrm{Metric}(\mathrm{SuFs}, LT = 1)$ denote the value of metric at lead time of 1 hour.

Table C.1. Three paired comparisons of a relative 'difference' metric (Δ) in Eq. (4.37) for the five evaluation metrics (in *x*-axis) over (a) Event 2 and (b) Event 3. These comparing pairs are based on three standards of surrogate filter construction described in Section 4.2.2. The positive (negative) values of Δ indicate that the prediction results of Partial, Invariant, and Dual filters are more (less) accurate than those computed by Whole, Variant, and Single filters, respectively. This table matches Fig. 4.11 over Events 2 and 3.

Paired comparison		**Event 1**					**Event 2**					**Event 3**				
		$\widetilde{NSE}$	$\widetilde{PE}$	BS	$\overline{CRPS}$	$Spread$	$\widetilde{NSE}$	$\widetilde{PE}$	BS	$\overline{CRPS}$	$Spread$	$\widetilde{NSE}$	$\widetilde{PE}$	BS	$\overline{CRPS}$	$Spread$
Whole vs. Partial	VaSuFs	-38	-22	-41	9	2	-31	-59	3	3	4	-39	8	-9	1	-2
	Dual VaSuFs	43	-106	-89	1	-4	-160	-117	-115	-40	-40	-66	-2	22	-1	1
	InSuFs	-87	-343	-103	-6	-15	-163	-112	-26	-21	-19	-203	-52	-17	-41	-44
	Dual InSuFs	77	13	55	13	28	31	-34	-59	19	29	-55	21	87	-9	10
Variant vs. Invariant	SuWFs	38	77	49	-5	14	59	38	35	17	30	60	45	3	26	24
	Dual SuWFs	-99	-62	-164	-66	-76	-83	-14	13	-27	-25	-9	-68	-30	27	16
	SuPFs	16	18	26	-23	0	17	18	16	-3	13	13	9	-3	-6	-7
	Dual SuPFs	19	32	37	-46	-21	52	30	36	27	37	-2	-31	79	21	24
Single vs. Dual	VaSuWFs	68	80	76	59	62	54	48	34	54	50	55	69	30	42	31
	VaSuPFs	87	65	68	56	60	9	29	-46	33	27	47	65	50	40	32
	InSuWFs	-2	-46	-25	36	22	-104	4	11	29	11	-22	5	5	42	23
	InSuPFs	87	71	72	48	51	47	40	-12	52	47	38	50	90	56	52

D Universal kriging

Universal kriging (UK) is a stochastic interpolation algorithm which assumes that the model output $\boldsymbol{\mathcal{M}}(\boldsymbol{X})$ is a realization of an underlying Gaussian process [*Santner et al.*, 2003]:

$$\boldsymbol{Y} = \boldsymbol{\mathcal{M}}(\boldsymbol{X}) \approx \mathrm{UK}(\boldsymbol{X}) = \boldsymbol{\beta}^\top f(\boldsymbol{X}) + \sigma^2 Z(\boldsymbol{X}) \tag{D.1}$$

where $\boldsymbol{\beta}^\top f(\boldsymbol{X})$ is the trend (or the mean value of the Gaussian process); σ^2 is a Gaussian process variance (or kriging variance) and can be estimated using the empirical linear unbiased estimator as Eq. (D.4); $Z(\boldsymbol{X})$ is a stationary Gaussian process with zero-mean and unit-variance. The trend is composed of the predefined sets of function $f_i(\boldsymbol{X})$ and kriging coefficients β_i with $i = 1, \dots, P_{UK}$, where P_{UK} denotes the number of kriging coefficients.

$$\boldsymbol{\beta}^\top f(\boldsymbol{X}) = \sum_{i=1}^{P_{UK}} \beta_i f_i(\boldsymbol{X}) \tag{D.2}$$

The Gaussian process $Z(\boldsymbol{X})$ is characterized by an autocorrelation function (ACF) between two arbitrary input samples $\boldsymbol{X}$ and $\boldsymbol{X}'$, i.e., $R(\boldsymbol{X}, \boldsymbol{X}') = R(|\boldsymbol{X} - \boldsymbol{X}'|; \boldsymbol{\delta})$, where $\boldsymbol{\delta}$ is a priori unknown correlation parameter that can be obtained by the leave-one-out cross-validation in this study [*Bachoc*, 2013] (see Eq. (5.2)).

Three types of Kriging are defined with different assumptions of the trend term:

(1) Universal kriging: the trend term is a linear combination of P_{UK} predefined functions $f_i(\boldsymbol{X})$ as Eq. (D.1), where β_i, $i = 1, \dots, P_{UK}$, are unknown to be estimated.

(2) Simple Kriging: the trend is a known constant with $P_{UK} = 1$, $f_1(\boldsymbol{X}) = 1$, thus $\boldsymbol{\beta}^\top f(\boldsymbol{X}) = \beta$, where β is known.

(3) Ordinary Kriging: the trend is unknown constant with $P_{UK} = 1$, $f_1(\boldsymbol{X}) = 1$, thus $\boldsymbol{\beta}^{\top} f(\boldsymbol{X}) = \beta_1$, where β is a unknown and can be estimated using an empirical linear unbiased estimator [*Schöbi et al.*, 2015]:

$$\beta_1(\widehat{\boldsymbol{\delta}}) = \left(F^{\top}\mathbf{R}^{-1}F\right)^{-1}F\mathbf{R}^{-1}\boldsymbol{\mathcal{Y}} \tag{D.3}$$

where F is the correlation matrix of kriging, $F_k = f_1\left(\mathcal{X}^{(k)}\right)$, $k = 1, \dots, N$.

Given the trend part, the Gaussian process variance σ^2 can be estimated using an empirical linear unbiased estimator as:

$$\sigma^2(\widehat{\boldsymbol{\delta}}) = \frac{1}{N}(\boldsymbol{\mathcal{Y}} - F\boldsymbol{\beta})^{\mathrm{T}}\mathbf{R}(\widehat{\boldsymbol{\delta}})^{-1}(\boldsymbol{\mathcal{Y}} - F\boldsymbol{\beta}) \tag{D.4}$$

E Supplementary Materials for Chapter V

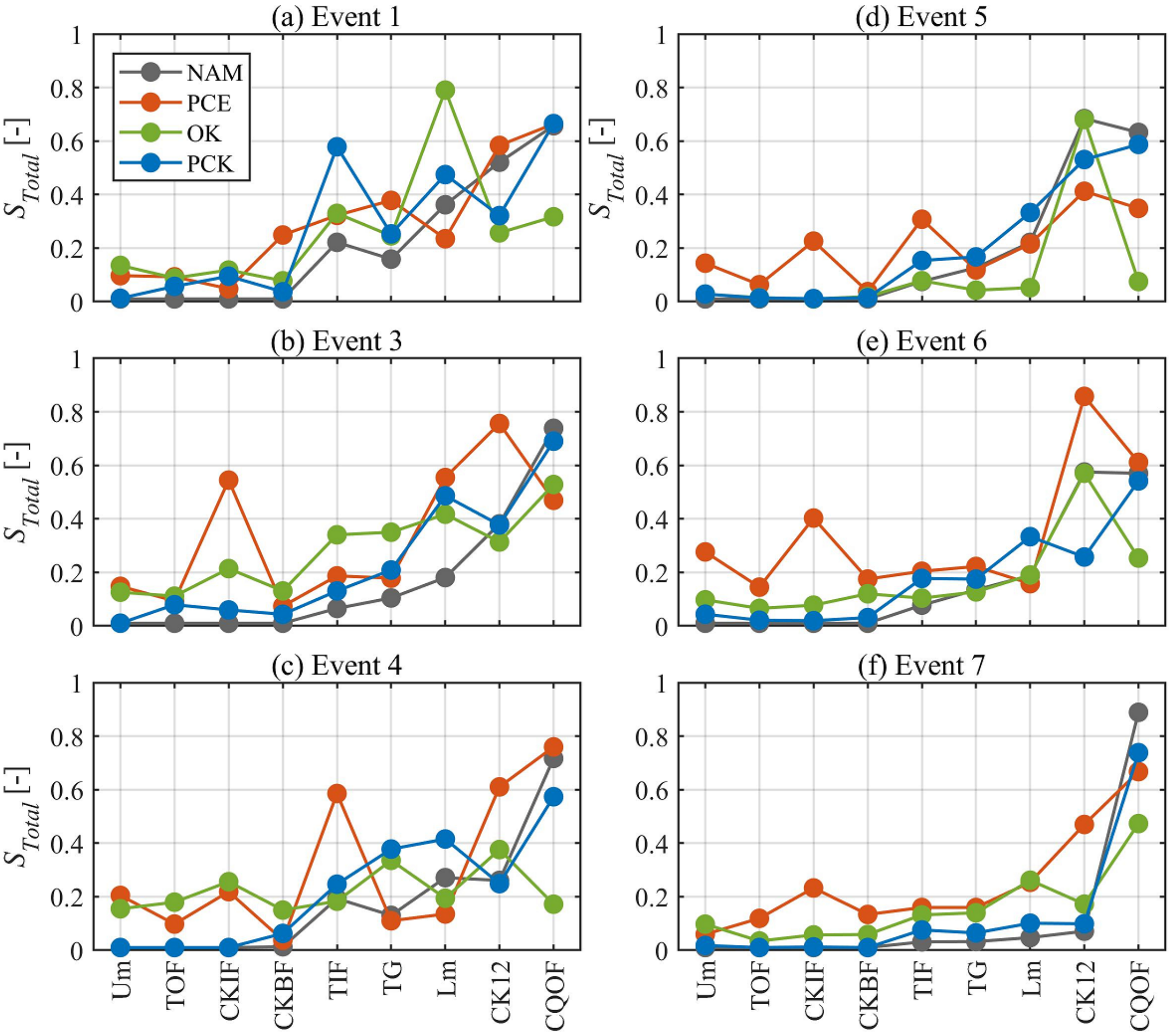


Figure E.1. Main sensitivity indices (S_{Total}) of nine parameters for four models (NAM, PCE, OK, and PCK) based on the variance of L over six flood events

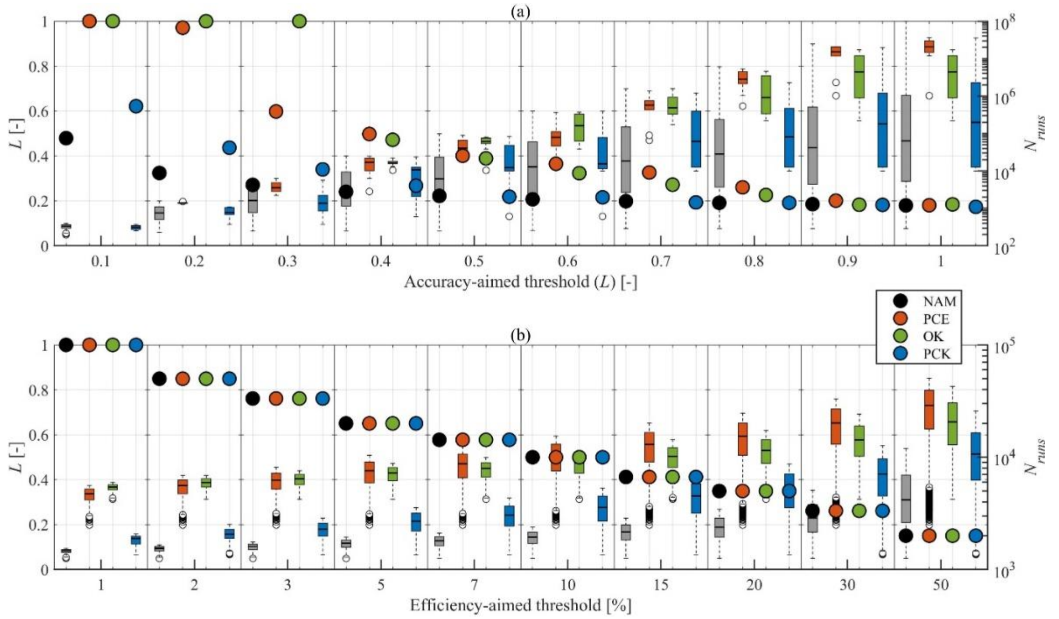


Figure E.2. The effects of two types of acceptance thresholds, (a) accuracy-aimed threshold and (b) efficiency-aimed threshold on the accuracy (the likelihood function, L) and the efficiency (the number of random runs, N_{runs}) for the four models. Each boxplot on the left axis demonstrates the median (central mark), the 25th and 75th percentiles (the edges of the box), and the maximum and minimum (the upper and lower whiskers) except for outliers (dot symbols) of 1,000 L values. The circles on the right axis represent N_{runs} required to attain 1,000 behavior runs.

Acronyms and Notations

Symbol	Description
AI	Accuracy index
BS	The Brier score
$CRPS$	The continuously ranked probability score
Dual InSuPF	Dual Invariant Surrogate Partial filter
Dual InSuWF	Dual Invariant Surrogate Whole filter
Dual VaSuPF	Dual Variant Surrogate Partial filter
Dual VaSuWF	Dual Variant Surrogate Whole filter
EI	Efficiency index
EnKF	Ensemble Kalman filter
InPCE	Invariant PCE
InSuF	Invariant Surrogate Filter
InSuPF	Single Invariant Surrogate Partial filter
InSuWF	Single Invariant Surrogate Whole filter
LAR	Least angle regression
LOO	Leave-one-out cross-validation error
LT	Lead time
NAM	Nedbør–Afstrømnings model
NRR	The Normalized RMSE Ratio
NSE	Nash-Sutcliffe efficiency
OK	Ordinary kriging
OLS	Ordinary least square regression
PCE	Polynomial chaos expansion
PCK	Polynomial chaos-kriging
PE	Peak error
PS	Performance score
QoI	Quantity of interest
SED-PD	Sequential experimental design-polynomial degree

SFM	Storage function model
SuF	Surrogate filter
SuPF	Surrogate Partial Filter
SuWF	Surrogate Whole Filter
UK	Universal kriging
UR	Uncertainty range
VaPCE	Variant PCE
VaSuF	Variant Surrogate Filter
VaSuPF	Single Variant Surrogate Partial filter
VaSuWF	Single Variant Surrogate Whole filter
VE	Volume error
$\boldsymbol{\alpha}$	Multi-indices in Eq. (2.2)
$D(y)$	Total variance of model output y in Eq. (2.15)
$D_{\tilde{a}}$	The variance averaged of model output y in Eq. (2.17)
E	Ensemble error matrix in EnKF
ε	PCE coefficients in Eq. (2.2)
ϵ	Leave-one-out error
$\epsilon_{LOO}^{\mathrm{QoI}}$	Leave-one-out cross-validation error for a QoI
ϵ_{th}^{lower}	Lower threshold of $\epsilon_{LOO}^{\mathrm{QoI}}$ used in Eq. (27)
ϵ_{th}^{upper}	Upper threshold of $\epsilon_{LOO}^{\mathrm{QoI}}$ used in Eq. (28)
ϵ_{th}^{slope}	Slope threshold of $\epsilon_{LOO}^{\mathrm{QoI}}$ used in Eq. (29)
F	Cumulative distribution of streamflow used in Eq. (4.30)
$\boldsymbol{F}$	Information matrix in Eq. (2.8)
$f(\cdot)$	nonlinear propagator for model states in Eq. (3.6)
$fac_{\boldsymbol{Model}}$	Factor in Eq. (3.38)
$h(\cdot)$	nonlinear propagator for mode output in Eq. (3.7)
i	An index for ensemble member
id	Index corresponding to varying threshold values in Eqs. (2.13) and (2.14)
j	An index for input of PCE
K	Kalman gain
k	An index for experimental design

$\mathcal{h}_k$	The k-th diagonal term of the matrix $\boldsymbol{F}$ in Eq. (2.21)
L	Likelihood function
$\mathcal{L}$	The likelihood in Eq. (5.5)
l	An index for multi-indices $\boldsymbol{\alpha}$ used in Eq. (2.8)
$\boldsymbol{\mathcal{M}}$	A deterministic hydrological model
$\boldsymbol{\mathcal{M}}^{\mathrm{su}}$	Surrogate model
$\boldsymbol{\mathcal{M}}^{PCE}$	A PCE surrogate model
N	The number of experimental design
N^*	The number of increasing samples of experimental design
N_I	The number of hydrologic model inputs
N_P	The number of uncertain parameters $\boldsymbol{\theta}$
N_S	The number of hydrologic model states
N_X	The number of PCE inputs
N_Y	The number of PCE outputs
N_{Ψ}	The number of PCE coefficients
n	The number of ensemble members in data assimilation
n_m	The number of samples used to implement Morris method
n_s	The number of samples used to implement Sobol' indices
n_w	The number of model runs to obtain the n number of the behavioral set
o	observed probability in Eq. (3.36)
p	Polynomial degree
p^f	forecast probability in Eq. (3.36)
$\boldsymbol{Q}$	Error covariance matrix in EnKF
Q	The number of random model runs in Eq. (2.14)
Q_{bas}	The direct runoffs (flow rates) of the basin in SFM
Q_{chn}	The direct runoffs (flow rates) of the channel in SFM
Ra	Ratio of the time-averaged RMSE in Eq. (3.30)
R_e	Effective rainfall
$R\left(\boldsymbol{X}, \boldsymbol{X}'\right)$	The correlation matrix of two arbitrary input samples $\boldsymbol{X}$ and $\boldsymbol{X}'$
$RT_{w+c,\boldsymbol{Model}}$	run time to compute one simulation over the warm-up and calibration periods in Eq. (3.38)

$RT_{f,\boldsymbol{Model},DA}$	run time to compute one simulation over the forecasting period in Eq. (3.38)
$RT_{build,t}^{Va}$	Runtime needed to construct variant filter at each time t
RT_{build}^{In}	Runtime needed to construct invariant surrogate filter
$RT_{run,t}$	Runtime needed to implement a single run of the filter at each time t
RT_{t}	Runtime needed to implement n runs of the filter at each time t
$RT_{cum,t}$	Cumulative runtime needed to implement n
$RT_{opt,t}^{\mathrm{QoI}}$	Runtime needed to estimate PCE coefficients during the implement of SED-PD for each QoI of VaPCE
RT_{opt}^{QoI}	Runtime needed to estimate PCE coefficients during the implement of SED-PD for each QoI of InPCE
$RT_{\boldsymbol{\mathcal{X}},t}^{\mathrm{QoI}}$	Runtime needed to generate corresponding response $\boldsymbol{\mathcal{Y}}$ from given $\boldsymbol{\mathcal{X}}$ for each QoI of VaPCE
$RT_{\boldsymbol{\mathcal{X}}}^{\mathrm{QoI}}$	Runtime needed to generate corresponding response $\boldsymbol{\mathcal{Y}}$ from given $\boldsymbol{\mathcal{X}}$ for each QoI of InPCE
S_{a}	First-order Sobol' indices of a-th model parameter, $\boldsymbol{\theta}_a$
S_{ab}	Second-order Sobol' indices of a-th ($\boldsymbol{\theta}_a$) and b-th ($\boldsymbol{\theta}_b$) parameters
S_{bas}	The storage amounts of the basin in SFM
S_{chn}	The storage amounts of the channel in SFM
$S_{Total,a}$	Total-order Sobol' indices of a-th model parameter, $\boldsymbol{\theta}_a$
S_{Ψ}	The number of significant PCE coefficients
T	The total duration of a flood/rainfall event
TRT	total run time in Eq. (3.38)
t	An index for the computational time step
U	Temporal average of GLUE uncertainty in Eq. (2.12)
$\boldsymbol{u}$	Hydrologic model forcings
V	Total volume of hydrograph
w	Model errors used in Eq. (3.6)
$\boldsymbol{X}$	Input of PCE
$\boldsymbol{\mathcal{X}}$	Experimental design
$\boldsymbol{x}$	Hydrologic model states
$\boldsymbol{Y}$	Output of a deterministic hydrological model
$\boldsymbol{\mathcal{Y}}$	The corresponding model response by given $\boldsymbol{\mathcal{X}}$
y	Simulated streamflow

y^{obs}	Observed streamflow
$Z(\boldsymbol{X})$	The Gaussian process Eq. (5.1)
Δ	"Difference" metric in the paired comparison between surrogate filters used in Eq. (4.37)
$\boldsymbol{\theta}$	Hydrologic model parameters
μ	Mean value in Morris method
σ	Standard deviation value in Morris method
λ	Non-negative constant in Eq. (2.19)
η	Actual observation error
τ	Parameter noise used in Eq. (4.2)
σ^2	The variance (or kriging variance) of the Gaussian process Eq. (5.1)
$\boldsymbol{\delta}$	Hyperparameter of the autocorrelation function
Ψ	multivariate orthonormal polynomials
ρ	Prior distribution of parameters in Eq. (5.5)
Γ	"Difference" metric in the comparison of the performance of the filters with different lead times used in Eq. (4.38)
Π	Posterior distribution of parameters in Eq. (5.5)

Bibliography

Abbaszadeh, P., H. Moradkhani, and H. Yan (2018), Enhancing hydrologic data assimilation by evolutionary Particle Filter and Markov Chain Monte Carlo, *Advances in Water Resources*, *111*, 192-204, doi:10.1016/j.advwatres.2017.11.011.

Abdar, M., F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi (2021), A review of uncertainty quantification in deep learning: Techniques, applications and challenges, *Information Fusion*, *76*, 243-297, doi:10.1016/j.inffus.2021.05.008.

Ajami, N. K., Q. Duan, and S. Sorooshian (2007), An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction, *Water Resources Research*, *43*(1), doi:10.1029/2005wr004745.

Anderson, J. L. (2001), An Ensemble Adjustment Kalman Filter for Data Assimilation, *Monthly Weather Review*, *129*(12), 2884-2903, doi:10.1175/1520-0493(2001)129<2884:aeakff>2.0.co;2.

APFM (2013), Integrated flood management tools series: Flood Forecasting and Early Warning*Rep.*, Associated Programme on Flood Management.

Arulampalam, M. S., S. Maskell, N. Gordon, and T. Clapp (2002), A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking, *IEEE Transactions on Signal Processing*, *50*(2), 174-188, doi:10.1109/78.978374.

Asher, M. J., B. F. W. Croke, A. J. Jakeman, and L. J. M. Peeters (2015), A review of surrogate models and their application to groundwater modeling, *Water Resources Research*, *51*(8), 5957-5973, doi:10.1002/2015wr016967.

Bach, F. (2020), On the effectiveness of richardson extrapolation in machine learning, *arXiv preprint arXiv:2002.02835*.

Bachoc, F. (2013), Cross Validation and Maximum Likelihood estimations of hyper-parameters of Gaussian processes with model misspecification, *Computational Statistics & Data Analysis*, *66*, 55-69, doi:10.1016/j.csda.2013.03.016.

Bae, D.-H., and B.-J. Lee (2011), Development of Continuous Rainfall-Runoff Model for Flood Forecasting on the Large-Scale Basin, *Journal of Korea Water Resources Association*, *44*(1), 51-64, doi:10.3741/jkwra.2011.44.1.51.

Bai, Y., Z. Chen, J. Xie, and C. Li (2016), Daily reservoir inflow forecasting using multiscale deep feature learning with hybrid models, *Journal of Hydrology*, *532*, 193-206, doi:10.1016/j.jhydrol.2015.11.011.

Bakas, N. P. (2019), Numerical Solution for the Extrapolation Problem of Analytic Functions, *Research (Wash D C)*, *2019*, 3903187, doi:10.34133/2019/3903187.

Ballio, F., and A. Guadagnini (2004), Convergence assessment of numerical Monte Carlo simulations in groundwater hydrology, *Water Resources Research*, *40*(4), doi:10.1029/2003wr002876.

Bannister, R. N. (2017), A review of operational methods of variational and ensemble-variational data assimilation, *Quarterly Journal of the Royal Meteorological Society*, *143*(703), 607-633, doi:10.1002/qj.2982.

Bao, J., S. C. Sherwood, L. V. Alexander, and J. P. Evans (2017), Future increases in extreme precipitation exceed observed scaling rates, *Nature Climate Change*, *7*(2), 128-132, doi:10.1038/nclimate3201.

Baştuğ, E., A. Menafoglio, and T. Okhulkova (2013), Polynomial Chaos Expansion for an efficient uncertainty and sensitivity analysis of complex numerical models, 3153-3161, doi:10.1201/b15938-477.

Baú, D. A., and A. S. Mayer (2006), Stochastic management of pump-and-treat strategies using surrogate functions, *Advances in Water Resources*, *29*(12), 1901-1917, doi:10.1016/j.advwatres.2006.01.008.

Bazargan, H., M. Christie, A. H. Elsheikh, and M. Ahmadi (2015), Surrogate accelerated sampling of reservoir models with complex structures using sparse polynomial chaos expansion, *Advances in Water Resources*, *86*, 385-399, doi:10.1016/j.advwatres.2015.09.009.

Benke, K. K., K. E. Lowell, and A. J. Hamilton (2008), Parameter uncertainty, sensitivity analysis and prediction error in a water-balance hydrological model, *Mathematical and Computer Modelling*, *47*(11-12), 1134-1149, doi:10.1016/j.mcm.2007.05.017.

Berveiller, M., B. Sudret, and M. Lemaire (2006), Stochastic finite elements: a non intrusive approach by regression, *Eur. J. Comput. Mech. , 15*.

Beven, K. (1989), Changing ideas in hydrology — The case of physically-based models, *Journal of Hydrology*, *105*(1-2), 157-172, doi:10.1016/0022-1694(89)90101-7.

Beven, K. (2000), Uniqueness of place and non-uniqueness of models in assessing predictive uncertainty, paper presented at Computational methods in water resources - Volume 2 - Computational methods,surface water systems and hydrology.

Beven, K. (2006), A manifesto for the equifinality thesis, *Journal of Hydrology*, *320*(1-2), 18-36, doi:10.1016/j.jhydrol.2005.07.007.

Beven, K. (2012), Rainfall-Runoff Modelling: The Primer, doi:10.1002/9781119951001.

Beven, K., and A. Binley (1992), The future of distributed models: Model calibration and uncertainty prediction, *Hydrological Processes*, *6*(3), 279-298, doi:10.1002/hyp.3360060305.

Beven, K., and A. Binley (2014), GLUE: 20 years on, *Hydrological Processes*, *28*(24), 5897-5918, doi:10.1002/hyp.10082.

Beven, K., and J. Freer (2001), Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, *Journal of Hydrology*, *249*(1-4), 11-29, doi:10.1016/s0022-1694(01)00421-8.

Beven, K. J., S. Almeida, W. P. Aspinall, P. D. Bates, S. Blazkova, E. Borgomeo, J. Freer, K. Goda, J. W. Hall, J. C. Phillips, M. Simpson, P. J. Smith, D. B. Stephenson, T. Wagener, M.

Watson, and K. L. Wilkins (2018), Epistemic uncertainties and natural hazard risk assessment – Part 1: A review of different natural hazard areas, *Natural Hazards and Earth System Sciences*, *18*(10), 2741-2768, doi:10.5194/nhess-18-2741-2018.

Bichon, B. J., J. M. McFarland, and S. Mahadevan (2011), Efficient surrogate models for reliability analysis of systems with multiple failure modes, *Reliability Engineering & System Safety*, *96*(10), 1386-1395, doi:10.1016/j.ress.2011.05.008.

Blasone, R.-S., H. Madsen, and D. Rosbjerg (2008a), Uncertainty assessment of integrated distributed hydrological models using GLUE with Markov chain Monte Carlo sampling, *Journal of Hydrology*, *353*(1-2), 18-32, doi:10.1016/j.jhydrol.2007.12.026.

Blasone, R.-S., J. A. Vrugt, H. Madsen, D. Rosbjerg, B. A. Robinson, and G. A. Zyvoloski (2008b), Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov Chain Monte Carlo sampling, *Advances in Water Resources*, *31*(4), 630-648, doi:10.1016/j.advwatres.2007.12.003.

Blatman, G., and B. Sudret (2008), Sparse polynomial chaos expansions and adaptive stochastic finite elements using a regression approach, *Comptes Rendus Mécanique*, *336*(6), 518-523, doi:10.1016/j.crme.2008.02.013.

Blatman, G., and B. Sudret (2010), An adaptive algorithm to build up sparse polynomial chaos expansions for stochastic finite element analysis, *Probabilistic Engineering Mechanics*, *25*(2), 183-197, doi:10.1016/j.probengmech.2009.10.003.

Blatman, G., and B. Sudret (2011), Adaptive sparse polynomial chaos expansion based on least angle regression, *Journal of Computational Physics*, *230*(6), 2345-2367, doi:10.1016/j.jcp.2010.12.021.

Bloschl, G., A. Kiss, A. Viglione, M. Barriendos, O. Bohm, R. Brazdil, D. Coeur, G. Demaree, M. C. Llasat, N. Macdonald, D. Retso, L. Roald, P. Schmocker-Fackel, I. Amorim, M. Belinova, G. Benito, C. Bertolin, D. Camuffo, D. Cornel, R. Doktor, L. Elleder, S. Enzi, J. C. Garcia, R. Glaser, J. Hall, K. Haslinger, M. Hofstatter, J. Komma, D. Limanowka, D. Lun, A. Panin, J. Parajka, H. Petric, F. S. Rodrigo, C. Rohr, J. Schonbein, L. Schulte, L. P. Silva, W. H. J. Toonen, P. Valent, J. Waser, and O. Wetter (2020), Current European flood-rich period exceptional compared with past 500 years, *Nature*, *583*(7817), 560-566, doi:10.1038/s41586-020-2478-3.

Bogner, K., and F. Pappenberger (2011), Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system, *Water Resources Research*, *47*(7), doi:10.1029/2010wr009137.

Bowden, G. J., H. R. Maier, and G. C. Dandy (2012), Real-time deployment of artificial neural network forecasting models: Understanding the range of applicability, *Water Resources Research*, *48*(10), doi:10.1029/2012wr011984.

Brier, G. W. (1950), Verification of Forecasts Expressed in Terms of Probability, *Monthly Weather Review*, *78*(1), 1-3, doi:10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2.

Burgers, G., P. Jan van Leeuwen, and G. Evensen (1998), Analysis Scheme in the Ensemble Kalman Filter, *Monthly Weather Review*, *126*(6), 1719-1724, doi:10.1175/1520-0493(1998)126<1719:asitek>2.0.co;2.

Butts, M. B., J. T. Payne, M. Kristensen, and H. Madsen (2004), An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation, *Journal of Hydrology*, *298*(1-4), 242-266, doi:10.1016/j.jhydrol.2004.03.042.

Caflisch, R. E. (1998), Monte carlo and quasi-monte carlo methods, *Acta numerica*, *7*, 1-49.

Cameron, D., K. Beven, J. Tawn, and P. Naden (2000), Flood frequency estimation by continuous simulation (with likelihood based uncertainty estimation), *Hydrology and Earth System Sciences Discussions*, *4*(1), 23-34.

Campolongo, F., J. Cariboni, and A. Saltelli (2007), An effective screening design for sensitivity analysis of large models, *Environmental Modelling & Software*, *22*(10), 1509-1518, doi:10.1016/j.envsoft.2006.10.004.

Champion, K., B. Lusch, J. N. Kutz, and S. L. Brunton (2019), Data-driven discovery of coordinates and governing equations, *Proc Natl Acad Sci U S A*, *116*(45), 22445-22451, doi:10.1073/pnas.1906995116.

Chen, H., D. Yang, Y. Hong, J. J. Gourley, and Y. Zhang (2013), Hydrological data assimilation with the Ensemble Square-Root-Filter: Use of streamflow observations to update model states for real-time flash flood forecasting, *Advances in Water Resources*, *59*, 209-220, doi:10.1016/j.advwatres.2013.06.010.

Cheng, R., and Y. Jin (2015), A competitive swarm optimizer for large scale optimization, *IEEE Trans Cybern*, *45*(2), 191-204, doi:10.1109/TCYB.2014.2322602.

Choi, H. T., and K. Beven (2007), Multi-period and multi-criteria model conditioning to reduce prediction uncertainty in an application of TOPMODEL within the GLUE framework, *Journal of Hydrology*, *332*(3-4), 316-336, doi:10.1016/j.jhydrol.2006.07.012.

Christelis, V., and A. G. Hughes (2018), Metamodel-assisted analysis of an integrated model composition: An example using linked surface water – groundwater models, *Environmental Modelling & Software*, *107*, 298-306, doi:10.1016/j.envsoft.2018.05.004.

Christensen, S. (2004), A synthetic groundwater modelling study of the accuracy of GLUE uncertainty intervals, *Hydrology Research*, *35*(1), 45-59.

Cintra, R. S., and H. F. d. C. Velho (2018), Data Assimilation by Artificial Neural Networks for an Atmospheric General Circulation Model, in *Advanced Applications for Artificial Neural Networks*, edited, doi:10.5772/intechopen.70791.

Ciriello, V., V. Di Federico, M. Riva, F. Cadini, J. De Sanctis, E. Zio, and A. Guadagnini (2012), Polynomial chaos expansion for global sensitivity analysis applied to a model of radionuclide migration in a randomly heterogeneous aquifer, *Stochastic Environmental Research and Risk Assessment*, *27*(4), 945-954, doi:10.1007/s00477-012-0616-7.

Clark, M. P., D. E. Rupp, R. A. Woods, X. Zheng, R. P. Ibbitt, A. G. Slater, J. Schmidt, and M. J. Uddstrom (2008), Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model, *Advances in Water Resources*, *31*(10), 1309-1324, doi:10.1016/j.advwatres.2008.06.005.

Cloke, H. L., and F. Pappenberger (2009), Ensemble flood forecasting: A review, *Journal of Hydrology*, *375*(3-4), 613-626, doi:10.1016/j.jhydrol.2009.06.005.

Cortesi, A. F., G. Jannoun, and P. M. Congedo (2019), Kriging-sparse Polynomial Dimensional Decomposition surrogate model with adaptive refinement, *Journal of Computational Physics*, *380*, 212-242, doi:10.1016/j.jcp.2018.10.051.

CRED-UNISDR (2015), The human cost of weather-related disasters 1995-2015*Rep.*

Crestaux, T., O. Le Maıˆtre, and J.-M. Martinez (2009), Polynomial chaos expansion for sensitivity analysis, *Reliability Engineering & System Safety*, *94*(7), 1161-1172, doi:10.1016/j.ress.2008.10.008.

Cukier, R. I., C. M. Fortuin, K. E. Shuler, A. G. Petschek, and J. H. Schaibly (1973), Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory, *The Journal of Chemical Physics*, *59*(8), 3873-3878, doi:10.1063/1.1680571.

Davis, P. J., and P. Rabinowitz (2007), *Methods of numerical integration*, Courier Corporation.

Davison, B., V. Fortin, A. Pietroniro, M. K. Yau, and R. Leconte (2017), Parameter-state ensemble data assimilation using Approximate Bayesian Computing for short-term hydrological prediction, *Hydrology and Earth System Sciences Discussions*, 1-38, doi:10.5194/hess-2017-482.

DeChant, C. M., and H. Moradkhani (2012), Examining the effectiveness and robustness of sequential data assimilation methods for quantification of uncertainty in hydrologic forecasting, *Water Resources Research*, *48*(4), doi:10.1029/2011wr011011.

DeChant, C. M., and H. Moradkhani (2014), Toward a reliable prediction of seasonal forecast uncertainty: Addressing model and initial condition uncertainty with ensemble data assimilation and Sequential Bayesian Combination, *Journal of Hydrology*, *519*, 2967-2977, doi:10.1016/j.jhydrol.2014.05.045.

DHI (2014), *DHI Mike 11: A Modelling System for Rivers and Channels, Reference Manual*, Danish Hydraulic Institute (DHI) Water & Environment: Hørsholm, Denmark.

Diaz, P., A. Doostan, and J. Hampton (2018), Sparse polynomial chaos expansions via compressed sensing and D-optimal design, *Computer Methods in Applied Mechanics and Engineering*, *336*, 640-666, doi:10.1016/j.cma.2018.03.020.

Doi, M. V., and J. Kim (2020), Projections on climate internal variability and climatological mean at fine scales over South Korea, *Stochastic Environmental Research and Risk Assessment*, *34*(7), 1037-1058, doi:10.1007/s00477-020-01807-y.

Doi, M. V., and J. Kim (2021), Addressing Climate Internal Variability on Future Intensity-Duration-Frequency Curves at Fine Scales across South Korea, *Water*, *13*(20), 2828.

Donat, M. G., A. L. Lowry, L. V. Alexander, P. A. O'Gorman, and N. Maher (2016), More extreme precipitation in the world's dry and wet regions, *Nature Climate Change*, *6*(5), 508-513, doi:10.1038/nclimate2941.

Dottori, F., W. Szewczyk, J.-C. Ciscar, F. Zhao, L. Alfieri, Y. Hirabayashi, A. Bianchi, I. Mongelli, K. Frieler, R. A. Betts, and L. Feyen (2018), Increased human and economic losses from river flooding with anthropogenic warming, *Nature Climate Change*, *8*(9), 781-786, doi:10.1038/s41558-018-0257-z.

Du, X., and L. Leifsson (2019), Efficient uncertainty propagation for MAPOD via polynomial chaos-based Kriging, *Engineering Computations*, *ahead-of-print*(ahead-of-print), doi:10.1108/ec-04-2019-0157.

Duan, Q., S. Sorooshian, and V. Gupta (1992), Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resources Research*, *28*(4), 1015-1031, doi:10.1029/91wr02985.

Dubreuil, S., N. Bartoli, C. Gogu, T. Lefebvre, and J. M. Colomer (2018), Extreme value oriented random field discretization based on an hybrid polynomial chaos expansion — Kriging approach, *Computer Methods in Applied Mechanics and Engineering*, *332*, 540-571, doi:10.1016/j.cma.2018.01.009.

Dwelle, M. C., J. Kim, K. Sargsyan, and V. Y. Ivanov (2019), Streamflow, stomata, and soil pits: sources of inference for complex models with fast, robust uncertainty quantification, *Advances in Water Resources*, doi:10.1016/j.advwatres.2019.01.002.

Echard, B., N. Gayton, and M. Lemaire (2011), AK-MCS: An active learning reliability method combining Kriging and Monte Carlo Simulation, *Structural Safety*, *33*(2), 145-154, doi:10.1016/j.strusafe.2011.01.002.

Echeverribar, I., M. Morales-Hernández, P. Brufau, and P. García-Navarro (2019), 2D numerical simulation of unsteady flows for large scale floods prediction in real time, *Advances in Water Resources*, *134*, 103444, doi:10.1016/j.advwatres.2019.103444.

Elsheikh, A. H., I. Hoteit, and M. F. Wheeler (2014), Efficient Bayesian inference of subsurface flow models using nested sampling and sparse polynomial chaos surrogates, *Computer Methods in Applied Mechanics and Engineering*, *269*, 515-537, doi:10.1016/j.cma.2013.11.001.

Evensen, G. (1994), Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *Journal of Geophysical Research*, *99*(C5), 10143, doi:10.1029/94jc00572.

Evensen, G. (2003), The Ensemble Kalman Filter: theoretical formulation and practical implementation, *Ocean Dynamics*, *53*(4), 343-367, doi:10.1007/s10236-003-0036-9.

Faber, B. A., and J. R. Stedinger (2001), Reservoir optimization using sampling SDP with ensemble streamflow prediction (ESP) forecasts, *Journal of Hydrology, 249*(1-4), 113-133, doi:10.1016/s0022-1694(01)00419-x.

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

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, *Environmental Modelling & Software, 86,* 30-49, doi:10.1016/j.envsoft.2016.09.012.

Fang, K., D. Kifer, K. Lawson, and C. Shen (2020), Evaluating the potential and challenges of an uncertainty quantification method for long short-term memory models for soil moisture predictions, *Water Resources Research*, doi:10.1029/2020wr028095.

Fatichi, S., V. Y. Ivanov, A. Paschalis, N. Peleg, P. Molnar, S. Rimkus, J. Kim, P. Burlando, and E. Caporali (2016a), Uncertainty partition challenges the predictability of vital details of climate change, *Earth's Future*, *4*(5), 240-251, doi:doi:10.1002/2015EF000336.

Fatichi, S., G. G. Katul, V. Y. Ivanov, C. Pappas, A. Paschalis, A. Consolo, J. Kim, and P. Burlando (2015), Abiotic and biotic controls of soil moisture spatiotemporal variability and the occurrence of hysteresis, *Water Resources Research*, *51*(5), 3505-3524, doi:10.1002/2014wr016102.

Fatichi, S., E. R. Vivoni, F. L. Ogden, V. Y. Ivanov, B. Mirus, D. Gochis, C. W. Downer, M. Camporese, J. H. Davison, B. Ebel, N. Jones, J. Kim, G. Mascaro, R. Niswonger, P. Restrepo, R. Rigon, C. Shen, M. Sulis, and D. Tarboton (2016b), An overview of current applications, challenges, and future trends in distributed process-based models in hydrology, *Journal of Hydrology*, *537*, 45-60, doi:10.1016/j.jhydrol.2016.03.026.

Flood, I., and N. Kartam (1994), Neural Networks in Civil Engineering. I: Principles and Understanding, *Journal of Computing in Civil Engineering*, *8*(2), 131-148, doi:10.1061/(ASCE)0887-3801(1994)8:2(131).

Fortin, V., M. Abaza, F. Anctil, and R. Turcotte (2014), Why Should Ensemble Spread Match the RMSE of the Ensemble Mean?, *Journal of Hydrometeorology*, *15*(4), 1708-1713, doi:10.1175/jhm-d-14-0008.1.

Frame, J., F. Kratzert, D. Klotz, M. Gauch, G. Shelev, O. Gilon, L. M. Qualls, H. V. Gupta, and G. S. Nearing (2021), Deep learning rainfall-runoff predictions of extreme events, doi:10.5194/hess-2021-423.

Franz, K. J., and T. S. Hogue (2011), Evaluating uncertainty estimates in hydrologic models: borrowing measures from the forecast verification community, *Hydrology and Earth System Sciences*, *15*(11), 3367-3382, doi:10.5194/hess-15-3367-2011.

Freer, J., K. Beven, and B. Ambroise (1996), Bayesian Estimation of Uncertainty in Runoff Prediction and the Value of Data: An Application of the GLUE Approach, *Water Resources Research*, *32*(7), 2161-2173, doi:10.1029/95wr03723.

Freni, G., G. Mannina, and G. Viviani (2008), Uncertainty in urban stormwater quality modelling: the effect of acceptability threshold in the GLUE methodology, *Water Research*, *42*(8-9), 2061-2072, doi:10.1016/j.watres.2007.12.014.

Freni, G., G. Mannina, and G. Viviani (2009a), Identifiability analysis for receiving water body quality modelling, *Environmental Modelling & Software*, *24*(1), 54-62, doi:10.1016/j.envsoft.2008.04.013.

Freni, G., G. Mannina, and G. Viviani (2009b), Uncertainty in urban stormwater quality modelling: the influence of likelihood measure formulation in the GLUE methodology, *Sci Total Environ*, *408*(1), 138-145, doi:10.1016/j.scitotenv.2009.09.029.

Fu, G., Z. Kapelan, and P. Reed (2012), Reducing the Complexity of Multiobjective Water Distribution System Optimization through Global Sensitivity Analysis, *Journal of Water Resources Planning and Management*, *138*(3), 196-207, doi:10.1061/(asce)wr.1943-5452.0000171.

Fujita, T., D. J. Stensrud, and D. C. Dowell (2007), Surface Data Assimilation Using an Ensemble Kalman Filter Approach with Initial Condition and Model Physics Uncertainties, *Monthly Weather Review*, *135*(5), 1846-1868, doi:10.1175/mwr3391.1.

Gerstner, T. M. G. (1998), Numerical integration using sparse grids, *Numer. Algorithms*, *18*.

Ghanem, R. G., and P. D. Spanos (1991), *Stochastic Finite Elements: a Spectral Approach*, Springer, Verlag New York, doi:10.1007/978-1-4612-3094-6.

Gharamti, M. E., I. Hoteit, and J. Valstar (2013), Dual states estimation of a subsurface flow-transport coupled model using ensemble Kalman filtering, *Advances in Water Resources*, *60*, 75-88, doi:10.1016/j.advwatres.2013.07.011.

Ghiocel, D. M., and R. G. Ghanem (2002), Stochastic Finite-Element Analysis of Seismic Soil–Structure Interaction, *Journal of Engineering Mechanics*, *128*(1), 66-77, doi:10.1061/(asce)0733-9399(2002)128:1(66).

Giannakis, D., and A. J. Majda (2012), Quantifying the Predictive Skill in Long-Range Forecasting. Part II: Model Error in Coarse-Grained Markov Models with Application to Ocean-Circulation Regimes, *Journal of Climate*, *25*(6), 1814-1826, doi:10.1175/jcli-d-11-00110.1.

Giraldo, F. X., and M. Restelli (2008), A study of spectral element and discontinuous Galerkin methods for the Navier–Stokes equations in nonhydrostatic mesoscale atmospheric modeling: Equation sets and test cases, *Journal of Computational Physics*, *227*(8), 3849-3877, doi:10.1016/j.jcp.2007.12.009.

Glenis, V., V. Kutija, and C. G. Kilsby (2018), A fully hydrodynamic urban flood modelling system representing buildings, green space and interventions, *Environmental Modelling & Software*, *109*, 272-292, doi:10.1016/j.envsoft.2018.07.018.

Gneiting, T., and A. E. Raftery (2007), Strictly Proper Scoring Rules, Prediction, and Estimation, *Journal of the American Statistical Association*, *102*(477), 359-378, doi:10.1198/016214506000001437.

Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez (2009), Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *Journal of Hydrology*, *377*(1-2), 80-91, doi:10.1016/j.jhydrol.2009.08.003.

Gupta, H. V., S. Sorooshian, and P. O. Yapo (1998), Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, *Water Resources Research*, *34*(4), 751-763, doi:10.1029/97wr03495.

Hampton, J., and A. Doostan (2015), Compressive sampling of polynomial chaos expansions: Convergence analysis and sampling strategies, *Journal of Computational Physics*, *280*, 363-386, doi:10.1016/j.jcp.2014.09.019.

Han, D., T. Kwong, and S. Li (2007), Uncertainties in real-time flood forecasting with neural networks, *Hydrological Processes*, *21*(2), 223-228, doi:10.1002/hyp.6184.

Hatakeyama-Sato, K., and K. Oyaizu (2021), Generative Models for Extrapolation Prediction in Materials Informatics, *ACS Omega*, *6*(22), 14566-14574, doi:10.1021/acsomega.1c01716.

He, M., T. S. Hogue, S. A. Margulis, and K. J. Franz (2012), An integrated uncertainty and ensemble-based data assimilation approach for improved operational streamflow predictions, *Hydrol. Earth Syst. Sci.*, *16*(3), 815-831, doi:10.5194/hess-16-815-2012.

Helton, J. C., and F. J. Davis (2003), Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems, *Reliability Engineering & System Safety*, *81*(1), 23-69, doi:https://doi.org/10.1016/S0951-8320(03)00058-9.

Heo, K.-Y., K.-J. Ha, K.-S. Yun, S.-S. Lee, H.-J. Kim, and B. Wang (2014), Methods for uncertainty assessment of climate models and model predictions over East Asia, *International Journal of Climatology*, *34*(2), 377-390, doi:10.1002/joc.3692.

Herman, J. D., J. B. Kollat, P. M. Reed, and T. Wagener (2013), Technical Note: Method of Morris effectively reduces the computational demands of global sensitivity analysis for distributed watershed models, *Hydrology and Earth System Sciences*, *17*(7), 2893-2903, doi:10.5194/hess-17-2893-2013.

Hirabayashi, Y., R. Mahendran, S. Koirala, L. Konoshima, D. Yamazaki, S. Watanabe, H. Kim, and S. Kanae (2013), Global flood risk under climate change, *Nature Climate Change*, *3*(9), 816-821, doi:10.1038/nclimate1911.

Hossain, F., and E. N. Anagnostou (2005), Assessment of a stochastic interpolation based parameter sampling scheme for efficient uncertainty analyses of hydrologic models, *Computers & Geosciences*, *31*(4), 497-512, doi:10.1016/j.cageo.2004.11.001.

Hosseiny, H., F. Nazari, V. Smith, and C. Nataraj (2020), A Framework for Modeling Flood Depth Using a Hybrid of Hydraulics and Machine Learning, *Sci Rep*, *10*(1), 8222, doi:10.1038/s41598-020-65232-5.

Houtekamer, P. L., B. He, and H. L. Mitchell (2014), Parallel Implementation of an Ensemble Kalman Filter, *Monthly Weather Review*, *142*(3), 1163-1182, doi:10.1175/mwr-d-13-00011.1.

Houtekamer, P. L., and F. Zhang (2016), Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation, *Monthly Weather Review*, *144*(12), 4489-4532, doi:10.1175/mwr-d-15-0440.1.

Hu, C., and B. D. Youn (2010), Adaptive-sparse polynomial chaos expansion for reliability analysis and design of complex engineering systems, *Structural and Multidisciplinary Optimization*, *43*(3), 419-442, doi:10.1007/s00158-010-0568-9.

Hu, J., S. Chen, A. Behrangi, and H. Yuan (2019a), Parametric uncertainty assessment in hydrological modeling using the generalized polynomial chaos expansion, *Journal of Hydrology*, *579*, 124158, doi:10.1016/j.jhydrol.2019.124158.

Hu, R., F. Fang, C. C. Pain, and I. M. Navon (2019b), Rapid spatio-temporal flood prediction and uncertainty quantification using a deep learning method, *Journal of Hydrology*, *575*, 911-920, doi:10.1016/j.jhydrol.2019.05.087.

Iorgulescu, I., K. J. Beven, and A. Musy (2007), Flow, mixing, and displacement in using a data-based hydrochemical model to predict conservative tracer data, *Water Resources Research*, *43*(3), doi:10.1029/2005wr004019.

Ivanov, V. Y., S. Fatichi, G. D. Jenerette, J. F. Espeleta, P. A. Troch, and T. E. Huxman (2010), Hysteresis of soil moisture spatial heterogeneity and the “homogenizing” effect of vegetation, *Water Resources Research*, *46*(9), doi:10.1029/2009wr008611.

Ivanov, V. Y., D. Xu, M. C. Dwelle, K. Sargsyan, D. B. Wright, N. Katopodes, J. Kim, V. N. Tran, A. Warnock, S. Fatichi, P. Burlando, E. Caporali, P. Restrepo, B. F. Sanders, M. M. Chaney,

A. M. B. Nunes, F. Nardi, E. R. Vivoni, E. Istanbulluoglu, G. Bisht, and R. L. Bras (2021), Breaking Down the Computational Barriers to Real-Time Urban Flood Forecasting, *Geophysical Research Letters*, doi:10.1029/2021gl093585.

Jiang, S., Y. Zheng, and D. Solomatine (2020), Improving AI System Awareness of Geoscience Knowledge: Symbiotic Integration of Physical Approaches and Deep Learning, *Geophysical Research Letters*, *47*(13), doi:10.1029/2020gl088229.

Jiang, Y., C. Liu, X. Li, L. Liu, and H. Wang (2015), Rainfall-runoff modeling, parameter estimation and sensitivity analysis in a semiarid catchment, *Environmental Modelling & Software*, *67*, 72-88, doi:10.1016/j.envsoft.2015.01.008.

Jin, X., C.-Y. Xu, Q. Zhang, and V. P. Singh (2010), Parameter and modeling uncertainty simulated by GLUE and a formal Bayesian method for a conceptual hydrological model, *Journal of Hydrology*, *383*(3-4), 147-155, doi:10.1016/j.jhydrol.2009.12.028.

Karhunen, K. (1946), Zur spektraltheorie stochastischer prozesse, *Annales Academiae Scientiarum Fennicae. Mathematica-Physica*, *34. 1946*, 1-7.

Kavetski, D., G. Kuczera, and S. W. Franks (2006), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, *Water Resources Research*, *42*(3), doi:10.1029/2005wr004368.

Keating, E. H., J. Doherty, J. A. Vrugt, and Q. Kang (2010), Optimization and uncertainty assessment of strongly nonlinear groundwater models with high parameter dimensionality, *Water Resources Research*, *46*(10), doi:10.1029/2009wr008584.

Kennedy, J., and R. Eberhart (1995), Particle swarm optimization, *4*, 1942-1948, doi:10.1109/icnn.1995.488968.

Kersaudy, P., B. Sudret, N. Varsier, O. Picon, and J. Wiart (2015), A new surrogate modeling technique combining Kriging and polynomial chaos expansions – Application to uncertainty analysis in computational dosimetry, *Journal of Computational Physics*, *286*, 103-117, doi:10.1016/j.jcp.2015.01.034.

Kim, B., S. Y. Choi, and K.-Y. Han (2019a), Integrated Real-Time Flood Forecasting and Inundation Analysis in Small–Medium Streams, *Water*, *11*(5), 919, doi:10.3390/w11050919.

Kim, J., M. C. Dwelle, S. K. Kampf, S. Fatichi, and V. Y. Ivanov (2016a), On the non-uniqueness of the hydro-geomorphic responses in a zero-order catchment with respect to soil moisture, *Advances in Water Resources*, *92*, 73-89, doi:10.1016/j.advwatres.2016.03.019.

Kim, J., and V. Y. Ivanov (2014), On the nonuniqueness of sediment yield at the catchment scale: The effects of soil antecedent conditions and surface shield, *Water Resources Research*, *50*(2), 1025-1045, doi:10.1002/2013wr014580.

Kim, J., and V. Y. Ivanov (2015), A holistic, multi-scale dynamic downscaling framework for climate impact assessments and challenges of addressing finer-scale watershed dynamics, *Journal of Hydrology*, *522*, 645-660, doi:10.1016/j.jhydrol.2015.01.025.

Kim, J., V. Y. Ivanov, and S. Fatichi (2015), Climate change and uncertainty assessment over a hydroclimatic transect of Michigan, *Stochastic Environmental Research and Risk Assessment*, *30*(3), 923-944, doi:10.1007/s00477-015-1097-2.

Kim, J., V. Y. Ivanov, and S. Fatichi (2016b), Environmental stochasticity controls soil erosion variability, *Sci Rep*, *6*, 22065, doi:10.1038/srep22065.

Kim, J., V. Y. Ivanov, and S. Fatichi (2016c), Soil erosion assessment-Mind the gap, *Geophysical Research Letters*, *43*(24), 12,446-412,456, doi:10.1002/2016gl071480.

Kim, J., V. Y. Ivanov, and N. D. Katopodes (2012a), Hydraulic resistance to overland flow on surfaces with partially submerged vegetation, *Water Resources Research*, *48*(10), doi:10.1029/2012wr012047.

Kim, J., V. Y. Ivanov, and N. D. Katopodes (2013), Modeling erosion and sedimentation coupled with hydrological and overland flow processes at the watershed scale, *Water Resources Research*, *49*(9), 5134-5154, doi:10.1002/wrcr.20373.

Kim, J., J. Lee, D. Kim, and B. Kang (2019b), The role of rainfall spatial variability in estimating areal reduction factors, *Journal of Hydrology*, *568*, 416-426, doi:10.1016/j.jhydrol.2018.11.014.

Kim, J., M. E. Tanveer, and D.-H. Bae (2018), Quantifying climate internal variability using an hourly ensemble generator over South Korea, *Stochastic Environmental Research and Risk Assessment*, *32*(11), 3037-3051, doi:10.1007/s00477-018-1607-0.

Kim, J., A. Warnock, V. Y. Ivanov, and N. D. Katopodes (2012b), Coupled modeling of hydrologic and hydrodynamic processes including overland and channel flow, *Advances in Water Resources*, *37*, 104-126, doi:10.1016/j.advwatres.2011.11.009.

Kimura, T. (1961), The Flood Runoff Analysis Method by the Storage Function Model, *The Public Works Research Institute, Ministry of Construction, Japan*.

Kitanidis, P. K., and R. L. Bras (1980), Real-time forecasting with a conceptual hydrologic model: 1. Analysis of uncertainty, *Water Resources Research*, *16*(6), 1025-1033, doi:10.1029/WR016i006p01025.

Kleeman, R. (2002), Measuring Dynamical Prediction Utility Using Relative Entropy, *Journal of the Atmospheric Sciences*, *59*(13), 2057-2072, doi:10.1175/1520-0469(2002)059<2057:mdpuur>2.0.co;2.

Kollet, S. J., R. M. Maxwell, C. S. Woodward, S. Smith, J. Vanderborght, H. Vereecken, and C. Simmer (2010), Proof of concept of regional scale hydrologic simulations at hydrologic resolution utilizing massively parallel computer resources, *Water Resources Research*, *46*(4), doi:10.1029/2009wr008730.

Konakli, K., and B. Sudret (2016), Polynomial meta-models with canonical low-rank approximations: Numerical insights and comparison to sparse polynomial chaos expansions, *Journal of Computational Physics*, *321*, 1144-1169, doi:10.1016/j.jcp.2016.06.005.

Kratzert, F., D. Klotz, M. Herrnegger, A. K. Sampson, S. Hochreiter, and G. S. Nearing (2019), Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning, *Water Resources Research*, *55*(12), 11344-11354, doi:10.1029/2019wr026065.

Kuczera, G., D. Kavetski, S. Franks, and M. Thyer (2006), Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters, *Journal of Hydrology*, *331*(1-2), 161-177, doi:10.1016/j.jhydrol.2006.05.010.

Kuczera, G., and E. Parent (1998), Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm, *Journal of Hydrology*, *211*(1-4), 69-85, doi:10.1016/s0022-1694(98)00198-x.

Kullback, S. (1997), *Information theory and statistics*, Courier Corporation.

Kullback, S., and R. A. Leibler (1951), On Information and Sufficiency, *The Annals of Mathematical Statistics*, *22*(1), 79-86, doi:10.1214/aoms/1177729694.

Lafaysse, M., B. Hingray, A. Mezghani, J. Gailhard, and L. Terray (2014), Internal variability and model uncertainty components in future hydrometeorological projections: The Alpine Durance basin, *Water Resources Research*, *50*(4), 3317-3341, doi:10.1002/2013wr014897.

Laloy, E., B. Rogiers, J. A. Vrugt, D. Mallants, and D. Jacques (2013), Efficient posterior exploration of a high-dimensional groundwater model from two-stage Markov chain Monte Carlo simulation and polynomial chaos expansion, *Water Resources Research*, *49*(5), 2664-2682, doi:10.1002/wrcr.20226.

Lataniotis, C., S. Marelli, and B. Sudret (2020), Extending Classical Surrogate Modeling to High Dimensions through Supervised Dimensionality Reduction: A Data-Driven Approach, *International Journal for Uncertainty Quantification*, *10*(1), 55-82, doi:10.1615/Int.J.UncertaintyQuantification.2020031935.

Le Maître, O. P., M. T. Reagan, H. N. Najm, R. G. Ghanem, and O. M. Knio (2002), A Stochastic Projection Method for Fluid Flow, *Journal of Computational Physics*, *181*(1), 9-44, doi:10.1006/jcph.2002.7104.

Leifsson, L., X. Du, and S. Koziel (2020), Efficient yield estimation of multiband patch antennas by polynomial chaos-based Kriging, *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, *33*(6), doi:10.1002/jnm.2722.

Levy, S., and D. M. Steinberg (2011), Computer experiments: a review, *AStA Advances in Statistical Analysis*, *94*(4), 311-324, doi:10.1007/s10182-010-0147-9.

Li, J., and D. Xiu (2008), On numerical properties of the ensemble Kalman filter for data assimilation, *Computer Methods in Applied Mechanics and Engineering*, *197*(43-44), 3574-3583, doi:10.1016/j.cma.2008.03.022.

Li, J., and D. Xiu (2009), A generalized polynomial chaos based ensemble Kalman filter with high accuracy, *Journal of Computational Physics*, *228*(15), 5454-5469, doi:10.1016/j.jcp.2009.04.029.

Li, L., J. Xia, C.-Y. Xu, and V. P. Singh (2010), Evaluation of the subjective factors of the GLUE method and comparison with the formal Bayesian method in uncertainty assessment of hydrological models, *Journal of Hydrology*, *390*(3-4), 210-221, doi:10.1016/j.jhydrol.2010.06.044.

Li, Y., D. Ryu, A. W. Western, and Q. J. Wang (2015), Assimilation of stream discharge for flood forecasting: Updating a semidistributed model with an integrated data assimilation scheme, *Water Resources Research*, *51*(5), 3238-3258, doi:10.1002/2014wr016667.

Li, Y., D. Ryu, A. W. Western, Q. J. Wang, D. E. Robertson, and W. T. Crow (2014), An integrated error parameter estimation and lag-aware data assimilation scheme for real-time flood forecasting, *Journal of Hydrology*, *519*, 2722-2736, doi:10.1016/j.jhydrol.2014.08.009.

Liu, B., Q. Zhang, and G. G. Gielen (2013), A Gaussian process surrogate model assisted evolutionary algorithm for medium scale expensive optimization problems, *IEEE Transactions on Evolutionary Computation*, *18*(2), 180-192.

Liu, H.-L., X. Chen, A.-M. Bao, and L. Wang (2007), Investigation of groundwater response to overland flow and topography using a coupled MIKE SHE/MIKE 11 modeling system for an arid watershed, *Journal of Hydrology*, *347*(3-4), 448-459, doi:10.1016/j.jhydrol.2007.09.053.

Liu, H., A. Thiboult, B. Tolson, F. Anctil, and J. Mai (2019), Efficient treatment of climate data uncertainty in ensemble Kalman filter (EnKF) based on an existing historical climate ensemble dataset, *Journal of Hydrology*, *568*, 985-996, doi:10.1016/j.jhydrol.2018.11.047.

Liu, Y., and H. V. Gupta (2007), Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, *Water Resources Research*, *43*(7), doi:10.1029/2006wr005756.

Liu, Y., A. H. Weerts, M. Clark, H. J. Hendricks Franssen, S. Kumar, H. Moradkhani, D. J. Seo, D. Schwanenberg, P. Smith, A. I. J. M. van Dijk, N. van Velzen, M. He, H. Lee, S. J. Noh, O. Rakovec, and P. Restrepo (2012), Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities, *Hydrology and Earth System Sciences*, *16*(10), 3863-3887, doi:10.5194/hess-16-3863-2012.

Lohani, A. K., N. K. Goel, and K. K. S. Bhatia (2014), Improving real time flood forecasting using fuzzy inference system, *Journal of Hydrology*, *509*, 25-41, doi:10.1016/j.jhydrol.2013.11.021.

Loos, S., C. M. Shin, J. Sumihar, K. Kim, J. Cho, and A. H. Weerts (2020), Ensemble data assimilation methods for improving river water quality forecasting accuracy, *Water Res*, *171*, 115343, doi:10.1016/j.watres.2019.115343.

Lüthen, N., S. Marelli, and B. Sudret (2020), Sparse polynomial chaos expansions: Literature survey and benchmark, *arXiv preprint arXiv:2002.01290*.

Madsen, H. (2000), Automatic calibration of a conceptual rainfall–runoff model using multiple objectives, *Journal of Hydrology*, *235*(3-4), 276-288, doi:10.1016/s0022-1694(00)00279-1.

Madsen, H., and C. Skotner (2005), Adaptive state updating in real-time river flow forecasting—a combined filtering and error forecasting procedure, *Journal of Hydrology*, *308*(1-4), 302-312, doi:10.1016/j.jhydrol.2004.10.030.

Makungo, R., J. O. Odiyo, J. G. Ndiritu, and B. Mwaka (2010), Rainfall–runoff modelling approach for ungauged catchments: A case study of Nzhelele River sub-quaternary catchment, *Physics and Chemistry of the Earth, Parts A/B/C*, *35*(13-14), 596-607, doi:10.1016/j.pce.2010.08.001.

Mandel, J., and J. D. Beezley (2009), An ensemble Kalman-particle predictor-corrector filter for non-Gaussian data assimilation, paper presented at International Conference on Computational Science, Springer.

Mantovan, P., and E. Todini (2006), Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, *Journal of Hydrology*, *330*(1-2), 368-381, doi:10.1016/j.jhydrol.2006.04.046.

Marrel, A., B. Iooss, F. Van Dorpe, and E. Volkova (2008), An efficient methodology for modeling complex computer codes with Gaussian processes, *Computational Statistics & Data Analysis*, *52*(10), 4731-4744, doi:10.1016/j.csda.2008.03.026.

Marshall, J., A. Adcroft, C. Hill, L. Perelman, and C. Heisey (1997), A finite-volume, incompressible Navier Stokes model for studies of the ocean on parallel computers, *Journal of Geophysical Research: Oceans*, *102*(C3), 5753-5766, doi:10.1029/96jc02775.

Matos, J. P., M. M. Portela, and A. J. Schleiss (2017), Towards Safer Data-Driven Forecasting of Extreme Streamflows, *Water Resources Management*, *32*(2), 701-720, doi:10.1007/s11269-017-1834-z.

Maxwell, R. M., F. K. Chow, and S. J. Kollet (2007), The groundwater–land-surface–atmosphere connection: Soil moisture effects on the atmospheric boundary layer in fully-coupled simulations, *Advances in Water Resources*, *30*(12), 2447-2466, doi:10.1016/j.advwatres.2007.05.018.

Maxwell, R. M., M. Putti, S. Meyerhoff, J.-O. Delfs, I. M. Ferguson, V. Ivanov, J. Kim, O. Kolditz, S. J. Kollet, M. Kumar, S. Lopez, J. Niu, C. Paniconi, Y.-J. Park, M. S. Phanikumar, C. Shen, E. A. Sudicky, and M. Sulis (2014), Surface-subsurface model intercomparison: A first set of benchmark results to diagnose integrated hydrology and feedbacks, *Water Resources Research*, *50*(2), 1531-1549, doi:10.1002/2013wr013725.

McKay, M. D., R. J. Beckman, and W. J. Conover (1979a), Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code, *Technometrics*, *21*(2), 239-245, doi:10.1080/00401706.1979.10489755.

McKay, M. D., R. J. Beckman, and W. J. Conover (1979b), A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code, *Technometrics*, *21*(2), 239, doi:10.2307/1268522.

McKenna, S. A., J. Doherty, and D. B. Hart (2003), Non-uniqueness of inverse transmissivity field calibration and predictive transport modeling, *Journal of Hydrology*, *281*(4), 265-280, doi:10.1016/s0022-1694(03)00194-x.

Mendoza, P. A., J. McPhee, and X. Vargas (2012), Uncertainty in flood forecasting: A distributed modeling approach in a sparse data catchment, *Water Resources Research*, *48*(9), doi:10.1029/2011wr011089.

Meng, J., and H. Li (2018), Uncertainty Quantification for Subsurface Flow and Transport: Coping With Nonlinearity/Irregularity via Polynomial Chaos Surrogate and Machine Learning, *Water Resources Research*, doi:10.1029/2018wr022676.

Miller, K. L., S. J. Berg, J. H. Davison, E. A. Sudicky, and P. A. Forsyth (2018), Efficient uncertainty quantification in fully-integrated surface and subsurface hydrologic simulations, *Advances in Water Resources*, *111*, 381-394, doi:10.1016/j.advwatres.2017.10.023.

Minns, A. W., and M. J. Hall (1996), Artificial neural networks as rainfall-runoff models, *Hydrological Sciences Journal*, *41*(3), 399-417, doi:10.1080/02626669609491511.

Mirzaei, M., Y. F. Huang, A. El-Shafie, and A. Shatirah (2015), Application of the generalized likelihood uncertainty estimation (GLUE) approach for assessing uncertainty in hydrological models: a review, *Stochastic Environmental Research and Risk Assessment*, *29*(5), 1265-1273, doi:10.1007/s00477-014-1000-6.

Mockler, E. M., K. P. Chun, G. Sapriza-Azuri, M. Bruen, and H. S. Wheater (2016), Assessing the relative importance of parameter and forcing uncertainty and their interactions in conceptual hydrological model simulations, *Advances in Water Resources*, *97*, 299-313, doi:10.1016/j.advwatres.2016.10.008.

Mohanty, S. (2015), Chapter 12 Metamodel-Based Fast AMS-SoC Design Methodologies, in *Nanoelectronic Mixed-Signal System Design*, edited, McGraw-Hill.

Mondal, A., and P. P. Mujumdar (2012), On the basin-scale detection and attribution of human-induced climate change in monsoon precipitation and streamflow, *Water Resources Research*, *48*(10), doi:10.1029/2011wr011468.

Montanari, A. (2005), Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations, *Water Resources Research*, *41*(8), doi:10.1029/2004wr003826.

Moradkhani, H. (2008), Hydrologic Remote Sensing and Land Surface Data Assimilation, *Sensors*, *8*(5), 2986-3004, doi:10.3390/s8052986.

Moradkhani, H., C. M. DeChant, and S. Sorooshian (2012), Evolution of ensemble data assimilation for uncertainty quantification using the particle filter-Markov chain Monte Carlo method, *Water Resources Research*, *48*(12), doi:10.1029/2012wr012144.

Moradkhani, H., K.-L. Hsu, H. Gupta, and S. Sorooshian (2005a), Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter, *Water Resources Research*, *41*(5), doi:10.1029/2004wr003604.

Moradkhani, H., K. L. Hsu, H. Gupta, and S. Sorooshian (2005b), Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter, *Water Resources Research*, *41*(5), 1-17, doi:10.1029/2004wr003604.

Moradkhani, H., and S. Sorooshian (2008), General Review of Rainfall-Runoff Modeling: Model Calibration, Data Assimilation, and Uncertainty Analysis, in *Hydrological Modelling and the Water Cycle: Coupling the Atmospheric and Hydrologic Models*, edited by S. Sorooshian, K.-L. Hsu, E. Coppola, B. Tomassetti, M. Verdecchia and G. Visconti, pp. 1-24, Springer, Berlin, doi:10.1007/978-3-540-77843-1_1.

Moradkhani, H., S. Sorooshian, H. V. Gupta, and P. R. Houser (2005c), Dual state–parameter estimation of hydrological models using ensemble Kalman filter, *Advances in Water Resources*, *28*(2), 135-147, doi:https://doi.org/10.1016/j.advwatres.2004.09.002.

Moriasi, D. N., J. G. Arnold, M. W. V. Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith (2007a), Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations, *Transactions of the ASABE*, *50*(3), 885-900, doi:10.13031/2013.23153.

Moriasi, D. N., J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith (2007b), Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations, *Transactions of the ASABE*, *50*(3), 885-900, doi:https://doi.org/10.13031/2013.23153.

Morris, M. D. (1991), Factorial Sampling Plans for Preliminary Computational Experiments, *Technometrics*, *33*(2), 161-174, doi:10.1080/00401706.1991.10484804.

Nagawkar, J., L. T. Leifsson, and X. Du (2020), Applications of Polynomial Chaos-Based Cokriging to Aerodynamic Design Optimization Benchmark Problems, doi:10.2514/6.2020-0542.

Neal, J. C., T. J. Fewtrell, P. D. Bates, and N. G. Wright (2010), A comparison of three parallelisation methods for 2D flood inundation models, *Environmental Modelling & Software*, *25*(4), 398-411, doi:10.1016/j.envsoft.2009.11.007.

Nearing, G. S., F. Kratzert, A. K. Sampson, C. S. Pelissier, D. Klotz, J. M. Frame, C. Prieto, and H. V. Gupta (2021), What Role Does Hydrological Science Play in the Age of Machine Learning?, *Water Resources Research*, *57*(3), doi:10.1029/2020wr028091.

Nelder, J. A., and R. Mead (1965), A Simplex Method for Function Minimization, *The Computer Journal*, *7*(4), 308-313, doi:10.1093/comjnl/7.4.308.

Neumann, M. B. (2012), Comparison of sensitivity analysis methods for pollutant degradation modelling: a case study from drinking water treatment, *Sci Total Environ*, *433*, 530-537, doi:10.1016/j.scitotenv.2012.06.026.

Nga, P. H., K. Takara, and N. H. Son (2015), Flood Hazard Impact Analysis in the Downstream of Vu Gia-Thu Bon River System, Quang Nam Province, Central Vietnam, *Journal of Japan Society of Civil Engineers, Ser. B1 (Hydraulic Engineering)*, *71*(4), I_157-I_162, doi:10.2208/jscejhe.71.I_157.

Nielsen, S. A., and E. Hansen (1973), Numerical simulation of the rainfall-runoffprocess on a daily basis, *Hydrology Research*, *4*(3), 171-190.

Nikiema, O., and R. Laprise (2011), Budget study of the internal variability in ensemble simulations of the Canadian Regional Climate Model at the seasonal scale, *Journal of Geophysical Research*, *116*(D16), doi:10.1029/2011jd015841.

O’Brien, R. J., B. D. Misstear, L. W. Gill, J. L. Deakin, and R. Flynn (2013), Developing an integrated hydrograph separation and lumped modelling approach to quantifying hydrological pathways in Irish river catchments, *Journal of Hydrology*, *486*, 259-270, doi:10.1016/j.jhydrol.2013.01.034.

Office, H. R. F. C. (2012), Improvement of flood prediction system by applying stochastic technique*Rep.*, Ministry of Land, Transport and Maritime Affairs, South Korea.

Oladyshkin, S., and W. Nowak (2012), Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion, *Reliability Engineering & System Safety*, *106*, 179-190, doi:10.1016/j.ress.2012.05.002.

Paprotny, D., A. Sebastian, O. Morales-Napoles, and S. N. Jonkman (2018), Trends in flood losses in Europe over the past 150 years, *Nat Commun*, *9*(1), 1985, doi:10.1038/s41467-018-04253-1.

Park, M., D. Kim, J. Kwak, and H. Kim (2014), Evaluation of Parameter Characteristics of a Storage Function Model, *Journal of Hydrologic Engineering*, *19*(2), 308-318, doi:10.1061/(asce)he.1943-5584.0000678.

Pathiraja, S., H. Moradkhani, L. Marshall, A. Sharma, and G. Geenens (2018), Data-Driven Model Uncertainty Estimation in Hydrologic Data Assimilation, *Water Resources Research*, *54*(2), 1252-1280, doi:10.1002/2018wr022627.

Pokhrel, P., H. V. Gupta, and T. Wagener (2008), A spatial regularization approach to parameter estimation for a distributed watershed model, *Water Resources Research*, *44*(12), doi:10.1029/2007wr006615.

Prein, A. F., R. M. Rasmussen, K. Ikeda, C. Liu, M. P. Clark, and G. J. Holland (2016), The future intensification of hourly precipitation extremes, *Nature Climate Change*, *7*(1), 48-52, doi:10.1038/nclimate3168.

Rajabi, M. M. (2019), Review and comparison of two meta-model-based uncertainty propagation analysis methods in groundwater applications: polynomial chaos expansion and Gaussian process emulation, *Stochastic Environmental Research and Risk Assessment*, doi:10.1007/s00477-018-1637-7.

Razavi, S., B. A. Tolson, and D. H. Burn (2012a), Numerical assessment of metamodelling strategies in computationally intensive optimization, *Environmental Modelling & Software*, *34*, 67-86, doi:10.1016/j.envsoft.2011.09.010.

Razavi, S., B. A. Tolson, and D. H. Burn (2012b), Review of surrogate modeling in water resources, *Water Resources Research*, *48*(7), doi:10.1029/2011wr011527.

Reed, S., V. Koren, M. Smith, Z. Zhang, F. Moreda, D.-J. Seo, and a. Dmip Participants (2004), Overall distributed model intercomparison project results, *Journal of Hydrology*, *298*(1-4), 27-60, doi:10.1016/j.jhydrol.2004.03.031.

Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks (2010), Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, *Water Resources Research*, *46*(5), doi:10.1029/2009wr008328.

Ricciuto, D., K. Sargsyan, and P. Thornton (2018), The Impact of Parametric Uncertainties on Biogeochemistry in the E3SM Land Model, *Journal of Advances in Modeling Earth Systems*, *10*(2), 297-319, doi:10.1002/2017ms000962.

Romanowicz, R., K. Beven, and J. Tawn (1994), Evaluation of predictive uncertainty in non-linear hydrological models using a Bayesian approach, *Statistics for the Environment*, *2*.

Rosenzweig, B. R., P. Herreros Cantis, Y. Kim, A. Cohn, K. Grove, J. Brock, J. Yesuf, P. Mistry, C. Welty, T. McPhearson, J. Sauer, and H. Chang (2021), The Value of Urban Flood Modeling, *Earth's Future*, *9*(1), doi:10.1029/2020ef001739.

Saad, G., and R. Ghanem (2009), Characterization of reservoir simulation models using a polynomial chaos-based ensemble Kalman filter, *Water Resources Research*, *45*(4), doi:10.1029/2008wr007148.

Saltelli, A. (2002a), Making best use of model evaluations to compute sensitivity indices, *Computer Physics Communications*, *145*(2), 280-297, doi:10.1016/s0010-4655(02)00280-1.

Saltelli, A. (2002b), Sensitivity Analysis for Importance Assessment, *Risk Analysis*, *22*(3), 579-590, doi:10.1111/0272-4332.00040.

Saltelli, A., S. Tarantola, F. Campolongo, and M. Ratto (2004), *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*, Halsted Press.

Sanders, B. F., J. E. Schubert, K. A. Goodrich, D. Houston, D. L. Feldman, V. Basolo, A. Luke, D. Boudreau, B. Karlin, W. Cheung, S. Contreras, A. Reyes, A. Eguiarte, K. Serrano, M. Allaire, H. Moftakhari, A. AghaKouchak, and R. A. Matthew (2020), Collaborative

Modeling With Fine-Resolution Data Enhances Flood Awareness, Minimizes Differences in Flood Perception, and Produces Actionable Flood Maps, *Earth's Future*, *8*(1), doi:10.1029/2019ef001391.

Santner, T. J., B. J. Williams, and W. I. Notz (2003), *The Design and Analysis of Computer Experiments*, Springer, doi:10.1007/978-1-4757-3799-8.

Sargsyan, K., C. Safta, H. N. Najm, B. J. Debusschere, D. Ricciuto, and P. Thornton (2014), Dimensionality Reduction for Complex Models Via Bayesian Compressive Sensing, *International Journal for Uncertainty Quantification*, *4*(1), 63-93, doi:10.1615/Int.J.UncertaintyQuantification.2013006821.

Schöbi, R., P. Kersaudy, B. Sudret, and J. Wiart (2014), Combining polynomial chaos expansions and kriging.

Schobi, R., and B. Sudret (2014), PC-Kriging: A new metamodelling method combining Polynomial Chaos Expansions and Kriging, in *the 2nd International Symposium on Uncertainty Quantification and Stochastic Modeling*, edited, Rouen, France.

Schöbi, R., B. Sudret, and S. Marelli (2017), Rare Event Estimation Using Polynomial-Chaos Kriging, *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, *3*(2), D4016002, doi:10.1061/ajrua6.0000870.

Schöbi, R., B. Sudret, and J. Wiart (2015), Polynomial-Chaos-based Kriging, *International Journal for Uncertainty Quantification*, 171-193, doi:10.1615/Int.J.UncertaintyQuantification.2015012467.hal-01432195.

Sene, K. (2008), *Flood Warning, Forecasting and Emergency Response*, Springer Science & Business Media, doi:10.1007/978-3-540-77853-0.

Shen, Z. Y., L. Chen, and T. Chen (2012), Analysis of parameter uncertainty in hydrological and sediment modeling using GLUE method: a case study of SWAT model applied to Three Gorges Reservoir Region, China, *Hydrology and Earth System Sciences*, *16*(1), 121-132, doi:10.5194/hess-16-121-2012.

Shi, L., J. Yang, D. Zhang, and H. Li (2009), Probabilistic collocation method for unconfined flow in heterogeneous media, *Journal of Hydrology*, *365*(1-2), 4-10, doi:10.1016/j.jhydrol.2008.11.012.

Shukla, J., T. DelSole, M. Fennessy, J. Kinter, and D. Paolino (2006), Climate model fidelity and projections of climate change, *Geophysical Research Letters*, *33*(7), doi:10.1029/2005gl025579.

Si, W., W. Bao, and H. V. Gupta (2015), Updating real-time flood forecasts via the dynamic system response curve method, *Water Resources Research*, *51*(7), 5128-5144, doi:10.1002/2015wr017234.

Simpson, T. W., J. D. Poplinski, P. N. Koch, and J. K. Allen (2001), Metamodels for Computer-based Engineering Design: Survey and recommendations, *Engineering with Computers*, *17*(2), 129-150, doi:10.1007/pl00007198.

Slivinski, L., and C. Snyder (2016), Exploring Practical Estimates of the Ensemble Size Necessary for Particle Filters, *Monthly Weather Review*, *144*(3), 861-875, doi:10.1175/mwr-d-14-00303.1.

Smith, M. B., V. Koren, S. Reed, Z. Zhang, Y. Zhang, F. Moreda, Z. Cui, N. Mizukami, E. A. Anderson, and B. A. Cosgrove (2012), The distributed model intercomparison project – Phase 2: Motivation and design of the Oklahoma experiments, *Journal of Hydrology*, *418-419*, 3-16, doi:10.1016/j.jhydrol.2011.08.055.

Smith, M. B., D.-J. Seo, V. I. Koren, S. M. Reed, Z. Zhang, Q. Duan, F. Moreda, and S. Cong (2004), The distributed model intercomparison project (DMIP): motivation and experiment design, *Journal of Hydrology*, *298*(1-4), 4-26, doi:10.1016/j.jhydrol.2004.03.040.

Smith, R. C. (2013), *Uncertainty quantification: theory, implementation, and applications*, Siam.

Sobol', I. M. (1993), Sensitivity Estimates for Nonlinear Mathematical Models, *Mathematical Modeling and Computational Experiment*, *1*(4), 407-414.

Sobol', I. M. (2001), Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, *Mathematics and Computers in Simulation*, *55*(1-3), 271-280, doi:10.1016/s0378-4754(00)00270-6.

Sochala, P., and O. P. Le Maître (2013), Polynomial Chaos expansion for subsurface flows with uncertain soil parameters, *Advances in Water Resources*, *62*, 139-154, doi:10.1016/j.advwatres.2013.10.003.

Stedinger, J. R., R. M. Vogel, S. U. Lee, and R. Batchelder (2008), Appraisal of the generalized likelihood uncertainty estimation (GLUE) method, *Water Resources Research*, *44*(12), doi:10.1029/2008wr006822.

Sudret, B. (2007), Uncertainty propagation and sensitivity analysis in mechanical models Contributions to structural reliability and stochastic spectral methods, Habilitation thesis, Universite Blaise Pascal, Clermont-Ferrand, France.

Sudret, B. (2008), Global sensitivity analysis using polynomial chaos expansions, *Reliability Engineering & System Safety*, *93*(7), 964-979, doi:10.1016/j.ress.2007.04.002.

Sukegawa, N., and Y. Kitagawa (1992), Flood runoff model for small urban watershed with detention basins, *Doboku Gakkai Ronbunshu*, *1992*(443), 1-8.

Tang, Y., P. Reed, K. van Werkhoven, and T. Wagener (2007a), Advancing the identification and evaluation of distributed rainfall-runoff models using global sensitivity analysis, *Water Resources Research*, *43*(6), doi:10.1029/2006wr005813.

Tang, Y., P. Reed, T. Wagener, and K. van Werkhoven (2007b), Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation, *Hydrology and Earth System Sciences*, *11*(2), 793-817, doi:10.5194/hess-11-793-2007.

Tanoue, M., Y. Hirabayashi, and H. Ikeuchi (2016), Global-scale river flood vulnerability in the last 50 years, *Sci Rep*, *6*, 36021, doi:10.1038/srep36021.

Tarantola, A. (2005), *Inverse problem theory and methods for model parameter estimation*, SIAM.

Thiboult, A., F. Anctil, and M. A. Boucher (2016), Accounting for three sources of uncertainty in ensemble hydrological forecasting, *Hydrol. Earth Syst. Sci.*, *20*(5), 1809-1825, doi:10.5194/hess-20-1809-2016.

Thiemann, M., M. Trosset, H. Gupta, and S. Sorooshian (2001), Bayesian recursive parameter estimation for hydrologic models, *Water Resources Research*, *37*(10), 2521-2535, doi:10.1029/2000wr900405.

Thomas E. Adams, I. (2016), *FLOOD FORECASTING: A Global Perspective*.

Thompson, J. R., H. R. Sørenson, H. Gavin, and A. Refsgaard (2004), Application of the coupled MIKE SHE/MIKE 11 modelling system to a lowland wet grassland in southeast England, *Journal of Hydrology*, *293*(1-4), 151-179, doi:10.1016/j.jhydrol.2004.01.017.

Todini, E. (1999), Using phase-state modelling for inferring forecasting uncertainty in nonlinear stochastic decision schemes, *Journal of Hydroinformatics*, *1*(2), 75-82, doi:10.2166/hydro.1999.0007.

Todini, E. (2004), Role and treatment of uncertainty in real-time flood forecasting, *Hydrological Processes*, *18*(14), 2743-2746, doi:10.1002/hyp.5687.

Tokar, A. S., and P. A. Johnson (1999), Rainfall-Runoff Modeling Using Artificial Neural Networks, *Journal of Hydrologic Engineering*, *4*(3), 232-239, doi:10.1061/(asce)1084-0699(1999)4:3(232).

Torre, E., S. Marelli, P. Embrechts, and B. Sudret (2019), Data-driven polynomial chaos expansion for machine learning regression, *Journal of Computational Physics*, *388*, 601-623, doi:10.1016/j.jcp.2019.03.039.

Tossavainen, O.-P., J. Percelay, M. Stacey, J. P. Kaipio, and A. Bayen (2011), State estimation and modeling error approach for 2-D shallow water equations and Lagrangian measurements, *Water Resources Research*, *47*(10), doi:10.1029/2010wr009401.

Tran, T. D., V. N. Tran, and J. Kim (2021), Improving the Accuracy of Dam Inflow Predictions Using a Long Short-Term Memory Network Coupled with Wavelet Transform and Predictor Selection, *Mathematics*, *9*(5), 551, doi:10.3390/math9050551.

Tran, V. N., M. C. Dwelle, K. Sargsyan, V. Y. Ivanov, and J. Kim (2020), A novel modeling framework for computationally efficient and accurate real-time ensemble flood forecasting with uncertainty quantification, *Water Resources Research*, doi:https://doi.org/10.1029/2019WR025727.

Tran, V. N., and J. Kim (2019), Quantification of predictive uncertainty with a metamodel: Toward more efficient hydrologic simulations, *Stochastic Environmental Research and Risk Assessment*, doi:10.1007/s00477-019-01703-0.

Tran, V. N., and J. Kim (2021a), A Robust Surrogate Data Assimilation Approach to Real-Time Forecasting using Polynomial Chaos Expansion, *Journal of Hydrology*, 126367, doi:10.1016/j.jhydrol.2021.126367.

Tran, V. N., and J. Kim (2021b), Toward an Efficient Uncertainty Quantification of Streamflow Predictions Using Sparse Polynomial Chaos Expansion, *Water*, *2*(203), doi:https://doi.org/10.3390/w13020203.

Tran, V. N., and J. Kim (2022), Robust and Efficient Uncertainty Quantification for Extreme Events that Deviate Significantly from the Training Dataset Using Polynomial Chaos-Kriging, *Journal of Hydrology*, 127716, doi:https://doi.org/10.1016/j.jhydrol.2022.127716.

Uhlenbrook, S., and A. Sieber (2005), On the value of experimental data to reduce the prediction uncertainty of a process-oriented catchment model, *Environmental Modelling & Software*, *20*(1), 19-32, doi:10.1016/j.envsoft.2003.12.006.

UNDP (1999), Viet Nam: Flood Damage Summary 06 Nov 1999, edited, United Nations Development Programme.

van Werkhoven, K., T. Wagener, P. Reed, and Y. Tang (2009), Sensitivity-guided reduction of parametric dimensionality for multi-objective calibration of watershed models, *Advances in Water Resources*, *32*(8), 1154-1169, doi:10.1016/j.advwatres.2009.03.002.

Vapnik, V. (2013), *The nature of statistical learning theory*, Springer science & business media.

Vigsnes, M., O. Kolbjørnsen, V. L. Hauge, P. Dahle, and P. Abrahamsen (2017), Fast and Accurate Approximation to Kriging Using Common Data Neighborhoods, *Mathematical Geosciences*, *49*(5), 619-634, doi:10.1007/s11004-016-9665-7.

Vrugt, J. A. (2016), Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation, *Environmental Modelling & Software*, *75*, 273-316, doi:10.1016/j.envsoft.2015.08.013.

Vrugt, J. A., C. G. H. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten (2005), Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, *Water Resources Research*, *41*(1), doi:10.1029/2004wr003059.

Vrugt, J. A., H. V. Gupta, L. A. Bastidas, W. Bouten, and S. Sorooshian (2003a), Effective and efficient algorithm for multiobjective optimization of hydrologic models, *Water Resources Research*, *39*(8), doi:10.1029/2002wr001746.

Vrugt, J. A., H. V. Gupta, W. Bouten, and S. Sorooshian (2003b), A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters, *Water Resources Research*, *39*(8), doi:10.1029/2002wr001642.

Vrugt, J. A., H. V. Gupta, B. Nualláin, and W. Bouten (2006a), Real-Time Data Assimilation for Operational Ensemble Streamflow Forecasting, *Journal of Hydrometeorology*, *7*(3), 548-565, doi:10.1175/jhm504.1.

Vrugt, J. A., B. Ó Nualláin, B. A. Robinson, W. Bouten, S. C. Dekker, and P. M. A. Sloot (2006b), Application of parallel computing to stochastic parameter estimation in environmental models, *Computers & Geosciences*, *32*(8), 1139-1155, doi:10.1016/j.cageo.2005.10.015.

Vrugt, J. A., and B. A. Robinson (2007), Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, *Water Resources Research*, *43*(1), doi:10.1029/2005wr004838.

Vrugt, J. A., P. H. Stauffer, T. Wöhling, B. A. Robinson, and V. V. Vesselinov (2008a), Inverse Modeling of Subsurface Flow and Transport Properties: A Review with New Developments, *Vadose Zone Journal*, *7*(2), 843, doi:10.2136/vzj2007.0078.

Vrugt, J. A., C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson (2008b), Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, *Water Resources Research*, *44*(12), doi:doi:10.1029/2007WR006720.

Vrugt, J. A., C. J. F. ter Braak, H. V. Gupta, and B. A. Robinson (2008c), Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling?, *Stochastic Environmental Research and Risk Assessment*, *23*(7), 1011-1026, doi:10.1007/s00477-008-0274-y.

Wang, D., Y. Chen, and X. Cai (2009), State and parameter estimation of hydrologic models using the constrained ensemble Kalman filter, *Water Resources Research*, *45*(11), doi:10.1029/2008wr007401.

Wang, G. G., and S. Shan (2007), Review of Metamodeling Techniques in Support of Engineering Design Optimization, *Journal of Mechanical Design*, *129*(4), 370, doi:10.1115/1.2429697.

Wang, H., W. Gong, Q. Duan, and Z. Di (2020), Evaluation of parameter interaction effect of hydrological models using the sparse polynomial chaos (SPC) method, *Environmental Modelling & Software*, *125*, 104612, doi:10.1016/j.envsoft.2019.104612.

Wang, S., B. C. Ancell, G. H. Huang, and B. W. Baetz (2018), Improving Robustness of Hydrologic Ensemble Predictions Through Probabilistic Pre- and Post-Processing in Sequential Data Assimilation, *Water Resources Research*, *54*(3), 2129-2151, doi:10.1002/2018wr022546.

Wang, S., G. H. Huang, B. W. Baetz, and B. C. Ancell (2017), Towards robust quantification and reduction of uncertainty in hydrologic predictions: Integration of particle Markov chain Monte Carlo and factorial polynomial chaos expansion, *Journal of Hydrology*, *548*, 484-497, doi:10.1016/j.jhydrol.2017.03.027.

Wang, S., G. H. Huang, B. W. Baetz, and W. Huang (2015), A polynomial chaos ensemble hydrologic prediction system for efficient parameter inference and robust uncertainty assessment, *Journal of Hydrology*, *530*, 716-733, doi:10.1016/j.jhydrol.2015.10.021.

Wang, X. (2021), Uncertainty quantification and global sensitivity analysis for transient wave propagation in pressurized pipes, *Water Resources Research*, doi:10.1029/2020wr028975.

Ward, P. J., B. Jongman, J. C. J. H. Aerts, P. D. Bates, W. J. W. Botzen, A. Diaz Loaiza, S. Hallegatte, J. M. Kind, J. Kwadijk, P. Scussolini, and H. C. Winsemius (2017), A global framework for future costs and benefits of river-flood protection in urban areas, *Nature Climate Change*, *7*(9), 642-646, doi:10.1038/nclimate3350.

Ward, P. J., B. Jongman, F. S. Weiland, A. Bouwman, R. van Beek, M. F. P. Bierkens, W. Ligtvoet, and H. C. Winsemius (2013), Assessing flood risk at the global scale: model setup, results, and sensitivity, *Environmental Research Letters*, *8*(4), 044019, doi:10.1088/1748-9326/8/4/044019.

Weerts, A. H., and G. Y. H. El Serafy (2006), Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models, *Water Resources Research*, *42*(9), doi:10.1029/2005wr004093.

Wei, C., and M. M. Dewoolkar (2006), Formulation of capillary hysteresis with internal state variables, *Water Resources Research*, *42*(7), doi:10.1029/2005wr004594.

Whitaker, J. S. (2012), Developments in ensemble data assimilation, paper presented at Proceedings of the Seminar on Data assimilation for atmosphere and ocean, ECMWF, 6-9 September 2011.

Whitaker, J. S., and T. M. Hamill (2002), Ensemble Data Assimilation without Perturbed Observations, *Monthly Weather Review*, *130*(7), 1913-1924, doi:10.1175/1520-0493(2002)130<1913:edawpo>2.0.co;2.

Wiener, N. (1938), The Homogeneous Chaos, *American Journal of Mathematics*, *60*(4), 897, doi:10.2307/2371268.

Wilson, A. G., E. Gilboa, J. P. Cunningham, and A. Nehorai (2014), Fast Kernel Learning for Multidimensional Pattern Extrapolation, paper presented at NIPS.

Wing, O. E. J., C. C. Sampson, P. D. Bates, N. Quinn, A. M. Smith, and J. C. Neal (2019), A flood inundation forecast of Hurricane Harvey using a continental-scale 2D hydrodynamic model, *Journal of Hydrology X*, *4*, 100039, doi:10.1016/j.hydroa.2019.100039.

Winsemius, H. C., Jeroen C. J. H. Aerts, Ludovicus P. H. van Beek, Marc F. P. Bierkens, A. Bouwman, B. Jongman, Jaap C. J. Kwadijk, W. Ligtvoet, Paul L. Lucas, Detlef P. van Vuuren, and Philip J. Ward (2015), Global drivers of future river flood risk, *Nature Climate Change*, *6*(4), 381-385, doi:10.1038/nclimate2893.

Wittmann, R., H.-J. Bungartz, and P. Neumann (2017), High performance shallow water kernels for parallel overland flow simulations based on FullSWOF2D, *Computers & Mathematics with Applications*, *74*(1), 110-125, doi:10.1016/j.camwa.2017.01.005.

WMO (2018), Impact-based Forecasting and Warning: Weather Ready Nations*Rep.*

Wu, B., Y. Zheng, Y. Tian, X. Wu, Y. Yao, F. Han, J. Liu, and C. Zheng (2014), Systematic assessment of the uncertainty in integrated surface water-groundwater modeling based on the probabilistic collocation method, *Water Resources Research*, *50*(7), 5848-5865, doi:10.1002/2014wr015366.

Xie, X., and D. Zhang (2010), Data assimilation for distributed hydrological catchment modeling via ensemble Kalman filter, *Advances in Water Resources*, *33*(6), 678-690, doi:10.1016/j.advwatres.2010.03.012.

Xie, X., and D. Zhang (2013), A partitioned update scheme for state-parameter estimation of distributed hydrologic models based on the ensemble Kalman filter, *Water Resources Research*, *49*(11), 7350-7365, doi:10.1002/2012wr012853.

Xing, Z., R. Qu, Y. Zhao, Q. Fu, Y. Ji, and W. Lu (2019), Identifying the release history of a groundwater contaminant source based on an ensemble surrogate model, *Journal of Hydrology*, *572*, 501-516, doi:10.1016/j.jhydrol.2019.03.020.

Xiong, L., and K. M. O’Connor (2008), An empirical method to improve the prediction limits of the GLUE methodology in rainfall–runoff modeling, *Journal of Hydrology*, *349*(1-2), 115-124, doi:10.1016/j.jhydrol.2007.10.029.

Xiu, D., and G. E. Karniadakis (2002), The Wiener--Askey Polynomial Chaos for Stochastic Differential Equations, *SIAM Journal on Scientific Computing*, *24*(2), 619-644, doi:10.1137/s1064827501387826.

Xu, D. (2020), Addressing Uncertainty in Understanding Hydroclimate, Hydrology and Hydraulics Across Scales, University of Michigan.

Yapo, P. O., H. V. Gupta, and S. Sorooshian (1996), Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data, *Journal of Hydrology*, *181*(1-4), 23-48, doi:10.1016/0022-1694(95)02918-4.

Young, P. C. (2002), Advances in real-time flood forecasting, *Philos Trans A Math Phys Eng Sci*, *360*(1796), 1433-1450, doi:10.1098/rsta.2002.1008.

Zahmatkesh, Z., M. Karamouz, and S. Nazif (2015), Uncertainty based modeling of rainfall-runoff: Combined differential evolution adaptive Metropolis (DREAM) and K-means clustering, *Advances in Water Resources*, *83*, 405-420, doi:10.1016/j.advwatres.2015.06.012.

Zak, S. K., and K. J. Beven (1999), Equifinality, sensitivity and predictive uncertainty in the estimation of critical loads, *Science of The Total Environment*, *236*(1-3), 191-214, doi:10.1016/s0048-9697(99)00282-x.

Zhang, C., J. Chu, and G. Fu (2013), Sobolʹʹs sensitivity analysis for a distributed hydrological model of Yichun River Basin, China, *Journal of Hydrology*, *480*, 58-68, doi:10.1016/j.jhydrol.2012.12.005.

Zhang, H., H.-J. Hendricks Franssen, X. Han, J. A. Vrugt, and H. Vereecken (2017), State and parameter estimation of two land surface models using the ensemble Kalman filter and the particle filter, *Hydrology and Earth System Sciences*, *21*(9), 4927-4958, doi:10.5194/hess-21-4927-2017.

Zhang, J., G. Lin, W. Li, L. Wu, and L. Zeng (2018a), An Iterative Local Updating Ensemble Smoother for Estimation and Uncertainty Assessment of Hydrologic Model Parameters With Multimodal Distributions, *Water Resources Research*, *54*(3), 1716-1733, doi:10.1002/2017wr020906.

Zhang, J., Q. Zheng, D. Chen, L. Wu, and L. Zeng (2020), Surrogate-Based Bayesian Inverse Modeling of the Hydrological System: An Adaptive Approach Considering Surrogate Approximation Error, *Water Resources Research*, *56*(1), doi:10.1029/2019wr025721.

Zhang, X., P. Liu, L. Cheng, Z. Liu, and Y. Zhao (2018b), A back-fitting algorithm to improve real-time flood forecasting, *Journal of Hydrology*, *562*, 140-150, doi:10.1016/j.jhydrol.2018.04.051.

Zhao, D., and D. Xue (2010), A multi-surrogate approximation method for metamodeling, *Engineering with Computers*, *27*(2), 139-153, doi:10.1007/s00366-009-0173-y.

Abstract in Korean

극심한 홍수는 기후 온난화로 인해 과거보다 더 자주 발생하며, 그것들은 더 심오한 사회-경제적 영향을 미친다. 홍수 예측은 홍수 위험 관리와 완화의 중요한 구성 요소 중 하나이지만 기상학적 투입, 초기 상태, 모델 구조 및 모델 매개변수에 의해 야기되는 여러 가지 불확실성의 영향을 받는다. 수많은 연구 노력이 홍수 예측 작업의 불확실성을 조사했다. 그러나 현재, 우리는 불확실한 정량적 방법으로 홍수 예측에서 계산 부담, 부정확성 및 신뢰할 수 없는 예측 가능성에 대한 장기간 지속되는 도전을 처리할 수 있는 포괄적인 연구가 완전히 부족하다. 본 논문은 불확실성 정량화로 현재의 홍수 예측을 효율적이고 정확하게 계산하기 위한 새로운 모델링 프레임워크 구축에 대한 포괄적인 지식을 얻는 것을 목표로 한다.

본 논문에서는 홍수 예측에서 수문학적 모델의 정확하고 강력하며 효율적인 불확실성 정량화를 위한 일련의 혁신적인 방법론이 개발되었다. 이러한 방법에는 다음이 포함됩니다. (i) 홍수 예측에서 수문 모델의 매개변수 불확실성을 빠르고 강력한 정량화 및 이해를 위한 PCE(polynomial chaos expansion)와 결합된 GLUE(generalized likelihood uncertainty estimation) 프레임워크를 기반으로 하는 통합 모델링 프레임워크; (ii) 처음으로 대리 모델링, 매개변수 추론 및 데이터 동화의 세 가지 모델링 기술을 결합한 계산상 효율적이고 정확한 실시간 앙상블 홍수 예측과 불확실성 정량화를 위한 새로운 모델링 프레임워크 (iii) Ensemble Kalman filter (EnKF)의 내부 프로세스를 대체하기 위해 PCE를 사용한 실시간 홍수 예측을 위한 새롭고 강력하며 효율적인 대리 데이터 동화 접근

방식; 및 (iv) 학습 데이터 공간에서 크게 벗어나는 극단적인 이벤트에 대해서도 신뢰할 수 있는 결합 결과를 제공할 수 있는 PCK(다항식 교란 크리깅)라는 새로운 대리 모델.

본 논문의 주요 업적들은 다음과 같이 요약된다. (i) PCE 대리 모델은 불확실성 정량화 작업의 계산 요구를 상쇄하기 위해 GLUE 프레임워크에 통합된다. 이는 불확실한 투입에 대한 모델 출력의 민감도에 대해 모델 행동의 동인에 대해 추론할 수 있는 해석 가능한 확률론적 프레임워크의 이점을 제공한다. (ii) 홍수 예측의 새로운 프레임워크는 처음으로 세 가지 모델링 기술의 이점을 포함한다. (1) PCE 대리물은 계산 시간을 크게 줄일 수 있다. (2) 매개 변수 추론(GLUE)은 모델의 더 빠른 수렴, 불확실성 감소 및 시뮬레이션 결과의 우수한 정확도를 허용한다. (3) EnKF는 예측 중에 발생하는 오류를 동화한다. 이 프레임워크는 수문학적 매개 변수의 불확실성을 설명하고 이해하는 데 총체적이고 강력한 접근 방식을 제공하며 실시간 홍수 예측에서 시뮬레이션의 계산 부담을 크게 줄인다. 이 모델링 프레임워크는 실시간 및 앙상블 홍수 예측을 위해 복잡하고 충실도가 높은 수문 및 유압 모델이 점점 더 채택되어야 한다고 주장하는 모델링 패러다임의 변화에 기여한다. (iii) EnKF의 내부 프로세스를 PCE로 대체하여 새로운 대리 필터를 개발하기 위해 대리 접근 방식의 힘을 더욱 활용합니다. 대리 필터를 구성하는 방법에 대한 포괄적인 조사에 따르면 새로운 부분(원래 필터의 일부 대체) 및 불변(전체 기간 동안 유효) 접근 방식이 정확도와 효율성 측면에서 선호되며, 이는 차원 및 차원 수를 직접 줄이는 데 도움이 된다. 그리고 실시간 예측 간의 격차를 해소한다. 이 제안된 대리 필터는 차원 문제에서 계산 집약적인 데이터 동기를 수행하기 위한 유망한 대체도구가 될 것이

다. 그리고 (iv) 다항식 교란 크리깅(PCK: polynomial chaos-kriging)이라는 새로운 대리 모델은 잘 알려진 두 가지 대리 모델인 PCE와 kriging의 장점을 결합하여 개발되었다. 이 이 조합은 훈련된 데이터 공간에서 크게 벗어난 극한 사건에 대한 스트림 흐름 예측을 가능하게 했고, 예측 불확실성을 강력하고 효율적으로 정량화할 수 있게 했다. 이 발견은 궁극적으로 잠재적으로 더 포괄적인 대리 모델을 향한 새로운 설계를 고무할 것이다.