Journal of Hydroinformatics

Journal of Hydroinformatics Vol 25 No 2, 511 doi: 10.2166/hydro.2023.184

A Hadoop cloud-based surrogate modelling framework for approximating complex hydrological models

Jinfeng Ma, Hua Zheng, Ruonan Li *, Kaifeng Rao, Yanzheng Yang and Weifeng Li

Research Centre for Eco-Environmental Sciences Chinese Academy of Sciences, Beijing 100085, China
*Corresponding author. E-mail: rnli@rcees.ac.cn

RL, 0000-0003-2955-0236

ABSTRACT

Hydrological simulation has long been a challenge because of the computationally intensive and expensive nature of complex hydrological models. In this paper, a surrogate modelling (SM) framework is presented based on the Hadoop cloud for approximating complex hydrological models. The substantial model runs required by the design of the experiment (DOE) of SM were solved using the Hadoop cloud. Polynomial chaos expansion (PCE) was fitted and verified using the high-fidelity model DOE and was then used as a case study to investigate the approximation capability in a Soil and Water Assessment Tool (SWAT) surrogate model with regard to the accuracy, fidelity, and efficiency. In experiments, the Hadoop cloud reduced the computation time by approximately 86% when used in a global sensitivity analysis. PCE achieved results equivalent to those of the standard Monte Carlo approach, with a flow variance coefficient of determination of 0.92. Moreover, PCE proved to be as reliable as the Monte Carlo approach but significantly more efficient. The proposed framework greatly decreases the computational costs through cloud computing and surrogate modelling, making it ideal for complex hydrological model simulation and optimization.

Key words: Chaospy, Hadoop cloud, polynomial chaos expansion, surrogate modelling, SWAT

HIGHLIGHTS

- Our surrogate modelling framework reduces the computational cost of simulations.
- The design of the experiment was parallelized on a Hadoop cloud.
- PCE was fitted and verified using a high-fidelity model.
- The approximation ability of PCE in the SWAT surrogate model was investigated.

1. INTRODUCTION

Environmental models are increasingly used to evaluate the effectiveness of alternative designs, operational management, and policy options (Black *et al.* 2014; Maier *et al.* 2016). This is largely attributable to the fact that they can emulate the dynamics of real-world systems in both conventional management scenarios and alternative virtual realities (Maier *et al.* 2016; Maier *et al.* 2019). Hydrological models, for instance, have been widely used to aid in the comprehension of natural processes and the investigation of the effects of anthropogenic activity on watershed systems (Fernandez-Palomino *et al.* 2021). However, modern simulation models are often computationally demanding, as they strive to rigorously capture comprehensive knowledge about real-world systems. As a result, the number of choices that can be assessed manually is often restricted, making it difficult to decide which alternatives to explore during the decision-making process (Mugunthan *et al.* 2005; Keating *et al.* 2010; Maier *et al.* 2019). To address this challenge, optimization algorithm (OA) techniques are commonly used to locate the best solution to minimize the number of models evaluated. In addition, high-performance computing (HPC) techniques such as cloud computing are also extensively used to 'divide' the modelling procedures into several small subtasks and 'conquer' – run these subtasks concurrently, to reduce the overall computation time. As a result, simulation optimization frameworks that couple model simulations with OAs and HPC techniques have become increasingly prevalent.

Through HPC, a whole computational task can be separated into several small separate subtasks that can be processed in parallel. HPC commonly uses two primary parallel computing architectures – standalone machines and clusters or gridded networks – that are characterized in terms of the degree of parallelism allowed by the hardware (Zhang *et al.* 2016). The former assigns tasks to diverse processors on a standalone machine, whereas the latter assigns tasks within a cluster or across a gridded network. The major drawback of the former parallelization strategy is it is not as scalable, while the latter strategy has a risk of network bandwidths becoming bottlenecks (White 2012). Moreover, the time and money required to obtain such resources may impede the use of contemporary parallel computing paradigms because they often derive from conventional supercomputing resources. Recently, cloud computing has shown promising capability for parallelizing model simulations and managing massive post-simulation results (Zhang *et al.* 2016). In particular, Hadoop – an open-source software framework extensively used for building cloud computing environments – has recently been the focus of extensive testing to address computational complexity (Hu *et al.* 2015). Ideally, even though HPC can perform large-scale simulations, given the complexity of the model, we still utilize OA to further reduce the number of model runs needed.

With OAs, the goal is to find the best solution given a limited amount of time and computing budget. Many OAs have been proposed in the literature (Tayfur 2017). Typical algorithms include the earlier shuffled complex evolution (SCE-UA) algorithm developed at the University of Arizona, the dynamically dimensioned search (DDS) algorithm designed for calibration problems with many parameters, and the Markov chain Monte Carlo (MCMC) sampler-based differential evolution adaptive metropolis (DREAM) method. In acknowledgement of the inherent multi-objective nature of complex models, various multi-objective algorithms (MOEAs) have also been developed. Notable algorithms include the non-dominated sorting genetic algorithm II (NSGA-II) (Ercan & Goodall 2016), Borg multi-objective evolutionary algorithm (Borg-MOEA) (Hadka & Reed 2013, 2015), and Pareto archived dynamically dimensioned search (PA-DDS) (Asadzadeh & Tolson 2013). Although most OAs strive to drastically reduce computing costs, in some circumstances, they may inversely increase them. In reality, rather than the optimization procedure, it is the model computation that dominates the execution time required by the majority of simulation practices. Recognizing that neither HPC nor OAs may appropriately address the computational burden of simulation optimization, this raises the question of whether it is possible to significantly lower the computational cost of the original complex model while maintaining comparable accuracy to approximate the original models. Fortunately, surrogate modelling (SM) is one approach that can achieve this.

SM, which operates at a higher degree of abstraction than the original simulation, is concerned with creating a 'surrogate' of the raw simulation model that is less expensive to execute (Razavi *et al.* 2012a). Surrogate models have been designed to be used intelligently to substitute simulation models. Response surface modelling and lower-fidelity modelling are the two main categories that fall under SM. The former is experimentally approximated via data-driven function approximation. The latter, in contrast to the original simulation models that are often regarded as high-fidelity models, uses physically based simulation models that are less detailed. The reduced simulation models known as low-fidelity models maintain most of the processes that were represented in the original simulation model. In practice, response surface modelling is more broadly applicable than lower-fidelity modelling since it requires no modifications to the original model, has fewer parameters and is straightforward to use (Razavi *et al.* 2012b). In response to surface modelling, there are three main categories of research: (1) the creation of experimental designs (design of the experiment (DOE)) for efficient approximation, (2) the creation and use of function approximation techniques as surrogates, and (3) the creation of frameworks using surrogates. To date, the second and third groups have received most of the attention in environmental research. However, the first category of DOE-related research has rarely been reported, which is mostly because of its notoriously high computational burden, limiting the potential of this SM tool in real-life applications.

There are numerous SM techniques that have been widely used to surrogate the original computationally intensive models, such as polynomials (Blatman & Sudret 2008; Fen *et al.* 2009), ANN (Papadrakakis *et al.* 1998; Behzadian *et al.* 2009), RBF (Hussain *et al.* 2002; Regis & Shoemaker 2007), Kriging (Sacks *et al.* 1989; Sakata *et al.* 2003), and MARS (Friedman 1991; Jin *et al.* 2001; Chen *et al.* 2018). Recently, a novel data-driven SM methodology called polynomial chaos expansion (PCE) has been proposed. PCE was first proposed by Wiener (Wiener 1938), and the method is now widely used in a variety of fields, ranging from transportation (Stavropoulou & Müller 2015) to chemistry (Paffrath & Wever 2007; Villegas *et al.* 2012), aerodynamics (WU *et al.* 2018), groundwater flow and contaminant transport (Laloy *et al.* 2013; Deman *et al.* 2016), and surface water modelling (Fan *et al.* 2015; Wang *et al.* 2015). Unlike other SM techniques, PCE provides an effective solution to characterize nonlinear effects in stochastic analysis by capturing the model output's dependency on uncertain input parameters using a collection of high-dimensional orthogonal polynomials (Wang *et al.* 2015). According to previous

research, PCE has the potential to be a useful approach when applied to simple conceptual hydrological models, such as HYdrological MODel (HYMOD), to save time and computational resources (Fan *et al.* 2015). A more recent study investigated the capabilities of PCE in generating probabilistic flow predictions and quantifying the uncertainty of Soil and Water Assessment Tool (SWAT) parameters to further establish its applicability and reliability for approximating more complex hydrological models (Ghaith & Li 2020). These findings suggest that PCE is a potentially effective technique that reduces the time and computational effort and that its advantages are more significant when applied to more complex models. As mentioned above, substantial research has been conducted on the creation and use of function approximation techniques as surrogates, as well as on the creation of application frameworks employing surrogates. However, to our knowledge, no studies have fully addressed the DOE implementation process, which is a fundamental component of SM. Due to a lack of DOE research, the adoption of SM approaches capable of supporting computationally intensive investigations such as sensitivity analysis (SA), calibration, and optimization has been hampered.

The primary goal of this research is to provide a framework to solve the DOE of SM implementation using cloud computing techniques. Another objective is to illustrate how SM may be exploited to further counterbalance the very large computational burden of sophisticated hydrological model simulations. We conducted this investigation utilizing the PCE approximation of the SWAT model as an example since both the SWAT model and PCE are widely utilized. The main contributions of our work are to illustrate how cloud computing methodologies can facilitate the surrogate modeling process, and how the trained surrogate modeling reduces the computational cost of complex hydrological model simulation while keeping their high accuracy, with implications for future projects that aim to reduce computation time in the hydrological simulation.

The paper is organized as follows. In Section 2, the Hadoop cloud-based SM framework is introduced for approximating the SWAT model. The major components of the framework, including the physical SWAT model setup, DOE, Hadoop cloud, and construction of PCE, are described. In Section 3, the presented framework is applied to the SWAT model approximation in a case study, and the accuracy, fidelity, and efficiency of the method are evaluated. In Sections 4 and 5, the approximation findings are presented and discussed. Finally, some conclusions are presented in Section 6.

2. HADOOP CLOUD-BASED SM FRAMEWORK FOR APPROXIMATING COMPLEX MODELS

The primary purpose of using a Hadoop cloud framework to surrogate complex models is to execute DOE – by far the most computationally intensive part of SM – into the Hadoop cloud. In this sense, the DOE's extensive evaluations of the original complex model are wrapped as a MapReduce job and then concurrently executed on the Hadoop cluster, allowing flexible integration of the model into cloud computing and guaranteeing that all the model evaluations are successfully processed. Another purpose is to elaborate on how to build a practical SM application by taking advantage of the appropriate toolkits. The entire framework may be logically separated into the following two sections based on the programming languages used. The Python web framework is used to host our main SM approach – a PCE library and an additional core SA library. The Java web framework is used to handle massive model calculations with the help of a Hadoop cloud. It is within the Hadoop cloud that the DOE is decomposed into substantial SWAT model evaluations that can run simultaneously to reduce the overall time needed. Although previous research has confirmed the feasibility of wrapping complex models into the MapReduce framework (Hu *et al.* 2015; Zhang *et al.* 2016; Ma *et al.* 2022a), to our knowledge, it is rarely used in solving the DOE problem of SM.

The proposed framework is primarily developed in Python in a Linux environment by combining the PCE library (Chaospy) and SA library (SALib). Python is used to construct Chaospy, a numerical toolkit for performing uncertainty quantification using non-intrusive PCE and sophisticated Monte Carlo (MC) techniques (Feinberg & Langtangen 2015). This toolkit is designed to help scientists develop customized statistical approaches by fusing many fundamental and sophisticated building blocks. Various techniques, including Sobol, FAST, the delta moment-independent measure, and the derivative-based global sensitivity measure (DGSM), are all implemented in SALib (Herman & Usher 2017).

The methodology of the framework consists of five steps: (i) a hydrological model setup, to establish the original high-fidelity model; (ii) DOE, to use appropriate space-filling strategies to empirically capture the behaviour of the underlying hydrological system over a limited range of variables; (iii) setup of the Hadoop cloud, to convert DOE to a MapReduce job, and fulfil it on multiple computing nodes; (iv) construction of PCE, to conduct SM using DOE; and (v) evaluation of PCE, to assess the approximation capability of PCE in surrogating the original high-fidelity model. The primary components

of the framework are described in detail in the following subsections. The primary tasks are described and depicted in Figure 1.

(i) Hydrological model setup

Initially, a high-fidelity model must be created as the target to be surrogated. Here, the setup of the SWAT model is illustrated as an example. The workflow starts with the compilation and collection of the input data, including meteorology, topography, land use, and soil (Franco *et al.* 2020). Then, the SWAT model was constructed through a professional graphical user interface (GUI) toolkit, either using ArcSWAT or Quantum Geographic Information System (QGIS), because both have a user-friendly interface. Subsequently, a SWAT project that comprises many configuration files can be generated with the help of the above modelling tool. These files must be executed directly by the SWAT model (Fortran-based SWAT executable program compiled for the Linux operating system). The subsequent DOEs are generated by updating these configuration files before running the SWAT executable program.

(ii) DOE

A typical SM framework begins with a DOE, which entails efficiently sampling the feasible parameter space and assessing their associated objective function values using the original model (Razavi *et al.* 2012b). Therefore, a well-distributed DOE program to generate the surrogate is critical to the success of SM practice. Because a priori knowledge of the parameter distribution is generally unavailable, either a uniform or gamma distribution is often suggested instead. Such assumptions of distribution are mainly derived from default and literature parameter values.

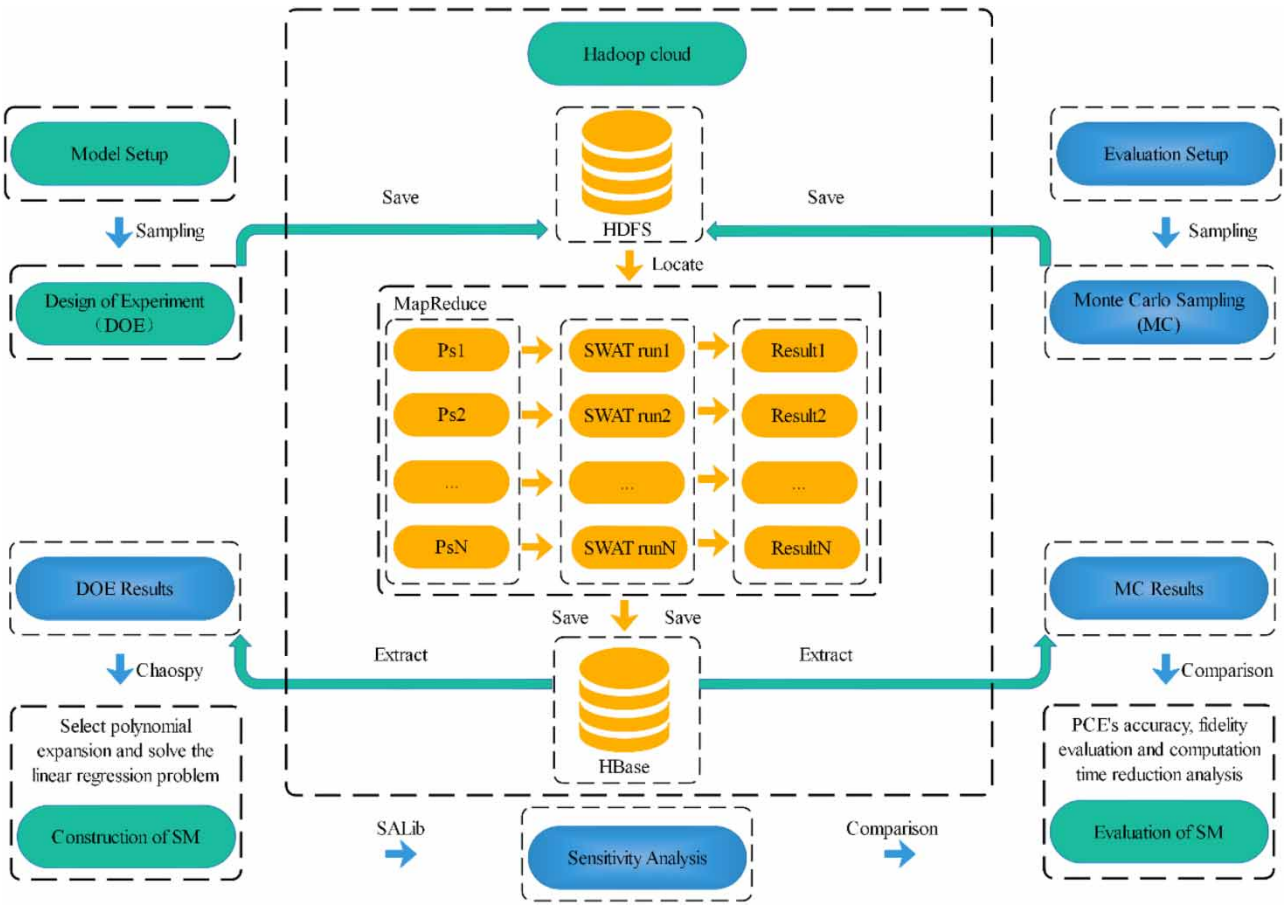

Figure 1 | Schematic diagram of the Hadoop cloud-based framework designed to surrogate the complex model. Within the Hadoop cloud, the original high-fidelity model is wrapped as a MapReduce task.

The DOE is made up of the following tasks. First, the model parameters are initially chosen based on prior experience, and their ranges are carefully determined based on available data. Second, certain sampling processes, such as the Latin hypercube sampling method (LHS) (Mckay *et al.* 2000), are utilized to sample the parameters properly. Namely, each sample is taken from a multi-dimensional space to represent as many potential combinations of the model parameters as possible. LHS is a stratification technique that forces random samples to be distributed more widely than regular random samples. It is similar to other low discrepancy sequences such as halton, hammersley korobov or sobol, but it keeps random samples at its core. The entire sampling process is conducted within the Python web framework, and the sampling results are sent to the Java web framework. Finally, the Java web framework creates a MapReduce task from the parsed samples and submits it to the Hadoop cloud for subsequent model evaluations.

(iii) Hadoop cloud

When the Hadoop cloud receives the MapReduce job, it begins to concurrently run the SWAT models. More specifically, the SWAT project configuration files located at the local computing node are executed by the SWAT executable program. Once the model execution finishes, the parsed simulation results are first stored in the HBase database for persistent storage purposes and then sent to the Python web framework. The Python web framework is responsible for receiving the parsed results sequentially until the final DOE is generated. Typically, the DOE is expressed as a tabular form of input parameter-output variable pairs. For more information, a detailed description of the coupling of the complex model and the Hadoop cloud can be found in prior publications (Zhang *et al.* 2016; Ma *et al.* 2022a).

(iv) Construction of the PCE

Fitting surrogates for the response surfaces of computationally intensive models over a set of previously evaluated samples will provide an approximate representation of the response surfaces. Numerous approximation strategies have been developed in the past and used as surrogates (Razavi *et al.* 2012b). In this study, the PCE was chosen because it is widely used. The popularity of PCE stems from one major feature: it has the simplest type of coefficients within a polynomial, and these coefficients can be flexibly extracted (Ghaith & Li 2020) or even estimated using the least squares regression approach. Second-order polynomial functions are the polynomials that are most frequently utilized as response surface surrogates. These functions contain $(D+1)(D+2)/2$ parameters, where D is the number of explanatory parameters. Once the DOE is generated, these data will be regarded as training data and used as input for the following polynomials. First, an expansion of orthogonal polynomials needs to be selected. Although not mandatory, first-order and second-order orthogonal polynomials are generally preferred over other types of polynomials due to their simplicity and stability (Feinberg & Langtangen 2015). In contrast, higher-order polynomials (third-order or more), which are frequently used in curve fitting, are sometimes impractical when D is large; as a result, they are rarely used (Razavi *et al.* 2012b). Second, coefficients of polynomials are estimated using either quadrature integration or the linear regression method.

(v) Evaluation of the PCE

Performing model analysis on an approximation, as a surrogate for the real model, is the final goal of SM. Following the construction of the PCE, the original model is completely replaced by the PCE to perform the analysis of interest. In this step, the approximation capability of the PCE in surrogating the SWAT model is assessed against extensive MC sampling and evaluation of the original SWAT model with regards to the accuracy, fidelity, and efficiency. Here, the difference (in the mean and variance) between the PCE and MC-simulated values of the variable of interest serves as a measure of accuracy. Fidelity is measured by the difference between the PCE-SA and the SA of the actual model. Efficiency is measured by the difference between the PCE execution time and the MC execution time.

2.1. Polynomial chaos expansion

The collection of random variables that characterize the input stochasticity may be used to define the output as a nonlinear function (Huang *et al.* 2007). By constructing a mapping between an unknown random variable and other (known) random variables, the PCE can be used to determine the probability distribution of the unknown random variable. Therefore, the PCE approach is frequently used to quantify how uncertainty propagates in a dynamic system with random inputs. Wiener (1938) first proposed the PCE approach by decomposing the model stochastic process into Hermite polynomials and Gaussian random variables. However, for non-Gaussian random input variables (e.g., variables from uniform or gamma distributions), the convergence of Hermite polynomial expansion may not be optimal. Consequently, Xiu *et al.* (2002) proposed generalized

PCE (GPCE) to address the above convergence problem for non-Gaussian distributions. The polynomials can be chosen from the hypergeometric polynomials of the Askey scheme based on the type of random input (Shi *et al.* 2009). The general PCE can be written in the form below:

$$y = a_0 + \sum_{i=1}^{n} a_i \Gamma_1(\xi_i) + \sum_{i=1}^{n} \sum_{j=1}^{i} a_{i,j} \Gamma_2(\xi_i, \xi_j) + \cdots\cdots \quad (1)$$

where y is the output, and $\Gamma_p(\xi_1, \xi_2, \cdots \xi_p)$ is the polynomial chaos of order p. For standard normal variables, the Hermite polynomial is used and expressed as

$$\Gamma_p(\xi_{i1}, \xi_{i2}, \ldots, \xi_{iM}) = (-1)^p e^{1/2\xi^T \xi} \frac{\partial^M}{\partial \xi_{i1} \partial \xi_{i2} \ldots \partial \xi_{iM}} e^{-1/2\xi^T \xi} \quad (2)$$

where $(\xi_{i1}, \xi_{i2}, \ldots, \xi_{iM})$ are the standard normal random variables.

In comparison to two other methods, such as Gram-Schmidt and Gill-Cholesky, the discretized Stieltjes procedure is the most widely used method for creating orthogonal polynomials. It is also thought to be the only method for building orthogonal polynomials that is truly numerically stable (Feinberg & Langtangen 2015). The discretized Stieltjes procedure is based on one-dimensional recursion coefficients, which can be easily estimated using numerical integration. Unfortunately, regarding the multivariate case, this approach can be used only if the variables are stochastically independent. Thus, the GPCE approach is used instead when the variables are stochastically dependent.

2.2. Sensitivity analysis library

To identify the parameters that have the greatest effect on the model performance, it is necessary to screen out sensitive parameters and statistically examine the effects of each parameter on the model performance. Many prior studies (van Griensven & Meixner 2006; Borgonovo *et al.* 2012) have employed SA for this purpose. As a result, if any parameters that are not influential can be identified and maintained at reasonable values within their ranges, the computational cost can be reduced without sacrificing model performance. In this research, SALib was used to perform SA for the SWAT model, which aids in estimating the effect of the model inputs or exogenous influences on the desired outputs in simulations (Herman & Usher 2017). SALib includes several global SA approaches that are simple to integrate into regular modelling processes, facilitating the creation of samples from model inputs. Moreover, it includes utilities for analyzing and visualizing model outputs.

3. CASE STUDY

3.1. SWAT model setup

To simplify the research procedure, the Meichuan River Basin, which has been used in previous research (Ma *et al.* 2022b), was chosen as the study area due to the availability of data. The same eight parameters (CANMX, CN2, CH_N2, CH_K2, ALPHA_BNK, SOL_AWC, SOL_K, and SOL_BD) were chosen as target variables, and they have the same upper and lower bounds as they did in previous research (Ma *et al.* 2022b).

3.2. Hadoop cloud setup

Information about the software and hardware of the Hadoop cloud can be found in Table 1 of previous research (Ma *et al.* 2022a). The basic sequential framework (also known as an offline framework) of SM was employed in this study, which is the most basic SM-enabled analytic framework and requires that all DOEs be available in the first step. Therefore, all DOEs were grouped into one single MapReduce job consisting of multiple SWAT model evaluation tasks. Such a design is consistent with the design philosophy of MapReduce and has several advantages in offline computing on the Hadoop cloud.

3.3. Construction of the PCE for the surrogate SWAT model

As previously stated, our offline framework begins (Step 1) with the DOE, which draws a specific number of samples over the feasible parameter space and evaluates their related objective function values using the original SWAT model. These parameter-target value pairs constitute the training dataset. In Step 2, an SM model is globally fitted to the training dataset. In Step 3, the SM model completely replaces the original model in performing the analysis of interest. For example, a specific

Table 1 | SA experimental setup for evaluating the fidelity of the PCE in approximating the SWAT model

SA method	SA measurements	Sampling technique	Sample size
Morris	Mean elementary effect	Morris	9,216
Sobol	Sobol's first and total indices	Satellite	18,432
Delta	Delta first-order index	Latin hypercube	1,024
FAST	FAST first-order index (FAST)	Fast_sampler	8,192
RBD-FAST	RBD-FAST first-order index	Latin hypercube	1,024
DGSM	DGSM Indice	Finite_diff	9,216

search algorithm, such as Bayesian optimization, is typically combined with the SM model for calibration purposes. The result obtained from the SM model is commonly deemed to be the result of the same analysis as the original model; for example, the optimal point found in the SM model is typically first examined by the original function and then judged to be the optimal solution to the original function (Razavi *et al.* 2012b).

Of the three steps, Step 1 is the most computationally expensive since it consumes almost the entire computational budget allocated to solving the problem (Razavi *et al.* 2012b). First, each of the eight parameters of the SWAT model must be assigned a probability density, and they are assumed to be stochastically independent. The choice of the uniform distribution is motivated by the fact that very little information was available from the literature or previous studies. Therefore, it is generally assumed that each of the eight parameters is uniformly distributed, and then a joint distribution derived from the eight uniform distributions is generated using the Chaospy library. Second, the sample points with eight vectors in a probability space can be drawn by using either a quasi-MC scheme, such as a Sobol sequence or Hammersley sequence, or other methods, such as LHS. In this study, we specified a second-order Gaussian quadrature scheme that is customized to the eight-vector joint distribution. With this scheme, 6,561 sample points were required in total. The final step of the DOE is to run the computational model at the sample points. By packaging the 6,561 sample points into a MapReduce job and submitting it to the Hadoop cloud, the evaluation results can be efficiently calculated and retrieved.

In Step 2, a PCE is created and globally fitted to the training dataset. First, the orthogonal second-order polynomials corresponding to the eight-vector joint distribution are generated using the Chaospy library. The approximation polynomial solver, which joins the PCE, quadrature nodes and weights, and model samples together, must estimate the polynomial coefficients before it can be used for later prediction purposes. There are numerous approaches for estimating the polynomial coefficients, which are often classified as either invasive or non-intrusive. Because invasive methods need to use information from the underlying model while calculating the coefficients, they typically need modifications of the source code of the original model. Therefore, the non-intrusive method was instead used in this study. In the field of non-intrusive approaches, there are essentially two viable options available: projection of pseudospectral data (Gottlieb & Orszag 1977) and point collocation (Hosder *et al.* 2007). The former uses a numerical integration approach to determine Fourier coefficients, whereas the latter solves a linear system generated by a statistical regression formulation. Unlike both MC integration and pseudospectral projection, the point collocation method does not assume that the samples follow any particular form, although they are traditionally selected to be random, quasi-random, nodes from quadrature integration, or a combination of the three. For this case study, we selected the samples to follow the Sobol samples from MC integration and optimal quadrature nodes from pseudospectral projection. Moreover, unlike those in the pseudospectral projection, the polynomials in point collocations are not required to be orthogonal. Nonetheless, orthogonal polynomials have been shown to work well with respect to stability (Feinberg & Langtangen 2015).

In Step 3, the approximate solver of the model can efficiently evaluate the model at any point in probability space. The built-in tools of Chaospy can further be used to derive statistical information from the model prediction. For example, the expected value and variance can easily be calculated. Our study focused on the evaluation of the approximation capability of SM with regards to the accuracy, fidelity, and efficiency. The SA technique was applied to measure the fidelity of the PCE in approximating the SWAT model. As shown in Table 1, six SA approaches were simultaneously used to check whether the reduced and statistic-based PCE could inherit the parameter sensitivity information from the target mechanistic model (SWAT).

4. RESULTS

4.1. Accuracy of the PCE model

The results of the PCE were compared to those of the MC simulation – a classic uncertainty quantification technique – to examine the reliability of the PCE for the quantification of the parameter uncertainty of SWAT. As stated in Section 3.3, 6,561 sets of parameter values were used to train the PCE model. Moreover, 10,000 additional sets of parameter values were drawn at random from the parameter distributions and used in the SWAT simulations as an MC analysis. The mean value and variance of the flow were calculated. During both the training and validation periods, the time series of the mean monthly flow for the MC and PCE were almost identical, as illustrated in Figures 2 and 3.

The variance tended to be higher in the MC results than in the PCE results during both the training and validation periods because the trained PCE model was based on the use of training sample points with a limited size (6,561 with training purpose), whose parameter range may have been narrower than that of the MC samples (10,000 with validation purpose), as illustrated in Figures 4 and 5. Notwithstanding this difference, the best fit coefficient of determination was 0.92 during the validation period. To some extent, this might imply that the PCE is a promising substitute to the MC for assessing the parameter uncertainty of SWAT. In summary, the PCE was able to produce uncertainty analysis findings that were comparable to those of the MC by creating a SWAT surrogate.

Another 1,000 MC experiments were undertaken to further explore the capability of the PCE to approximate the SWAT model. The Nash–Sutcliffe efficiency (NSE) metric was used to determine the global trend of the PCE and MC matching. Percent bias (PBIAS) was further used to measure the average tendency of the approximation of monthly flow. Normally, an NSE greater than 0.65 is considered acceptable (Ritter & Muñoz-Carpena 2013), and a PBIAS value of 0 is considered optimal, with low values indicating accurate matching. The results showed that the maximum, minimum, and mean values of the NSE were 0.93, 0.39, and 0.78, respectively. Of the 1,000 results, 99% had NSE values greater than 0.65, indicating good matching between the PCE and MC, as illustrated in Figure 6. This observation is also supported by the good performance statistics of PBIAS, with 99 and 68% of the 1,000 results having PBIAS less than 20 and 15%, respectively.

4.2. Fidelity of the PCE model

To obtain insight into the mechanism of the simulation model, it is possible to explore how the output evolves as a result of all the inputs and their potential interactions using the SA approaches. Therefore, the SA methods were employed experimentally to examine the fidelity of the PCE. Six prior SA experiments using the original SWAT method were compared with those using the PCE model. As illustrated in Figure 7, five of the approaches (all except the DGSM) showed similar weak recognizability of the following four parameters: CH_N2, SOL_AWC, SOL_K, and SOL_BD. For these five approaches, CN2 was the most sensitive parameter, followed by ALPHA_BNK. CANMX was slightly more sensitive than CH_N2. Overall, four approaches (all except DGSM and Morris) showed almost the same recognizability of the following four relatively sensitive parameters: CN2, ALPHA_BNK, CANMX, and CH_N2. Interestingly, CH_N2 was deemed to be non-sensitive in the original Morris SA findings. However, it was found to be slightly more sensitive than CANMX in the SA results of the PCE. This implies that the PCE is merely a rough approximation of the original model, and therefore, there is a risk that important

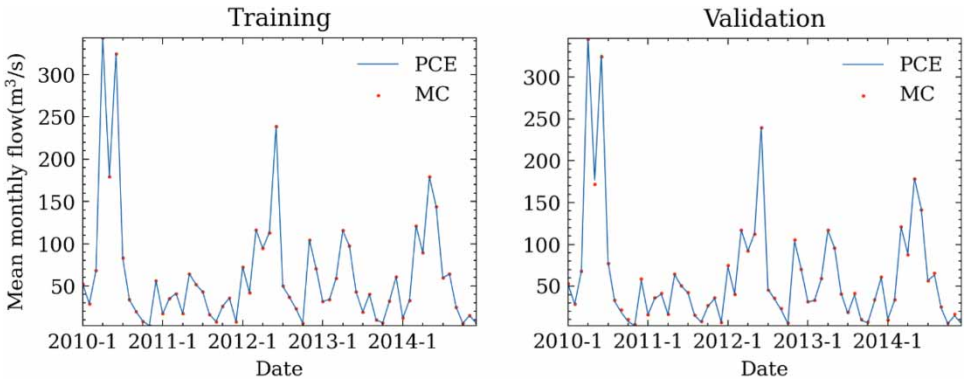

Figure 2 | Comparison of the mean monthly flow time series generated by the PCE and MC for the training period (left) and validation period (right).

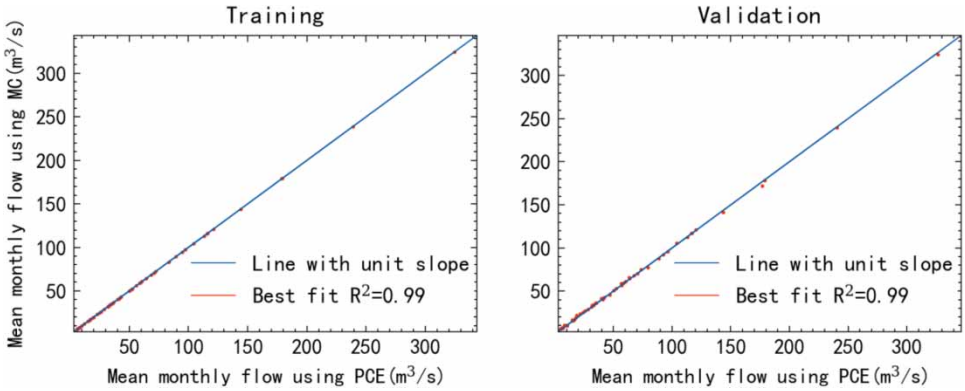

Figure 3 | Scatter plots of the mean monthly flow generated by the PCE and MC for the training period (left) and validation period (right).

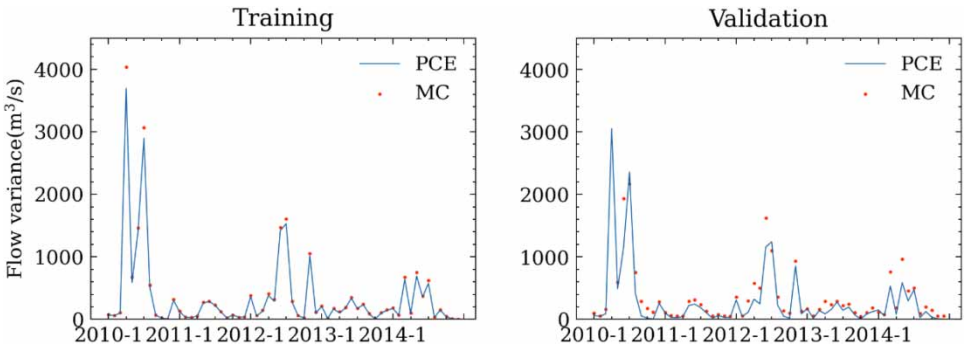

Figure 4 | Time series of the PCE and MC monthly flow variance for the training period (left) and validation period (right).

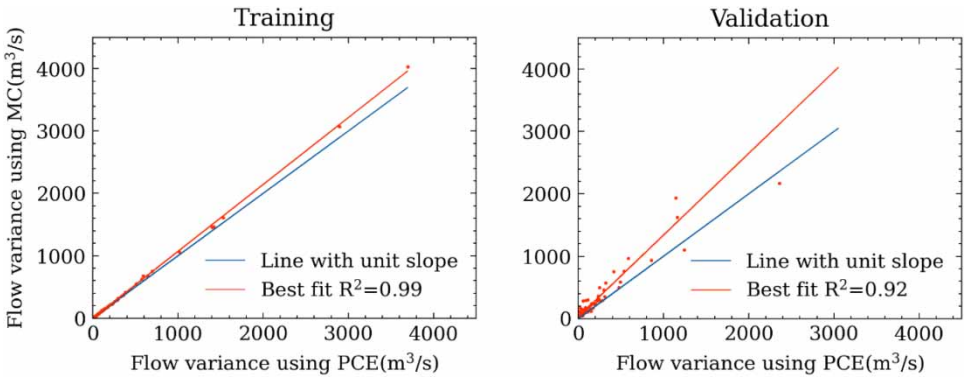

Figure 5 | Scatter plots of the PCE and MC monthly flow variance for the training period (left) and validation period (right).

information from the original model could be lost. Nevertheless, the PCE could still help to capture the sensitive parameters of the SWAT model.

To further evaluate the fidelity of the PCE model, the PCE was coupled with a global BOBYQA (Powell 2009) algorithm to determine whether the PCE model is capable of guiding the optimization algorithm for parameter identifiability. The experiment was repeated 1,000 times with varied initial parameter values for each experiment. It is well known that a parameter becomes more recognizable as the parameter range narrows or when its expected values are concentrated in a certain area of the feasible range (Fernandez-Palomino *et al.* 2021). Figure 8 shows that CN2, ALPHA_BNK, and SOL_AWC are

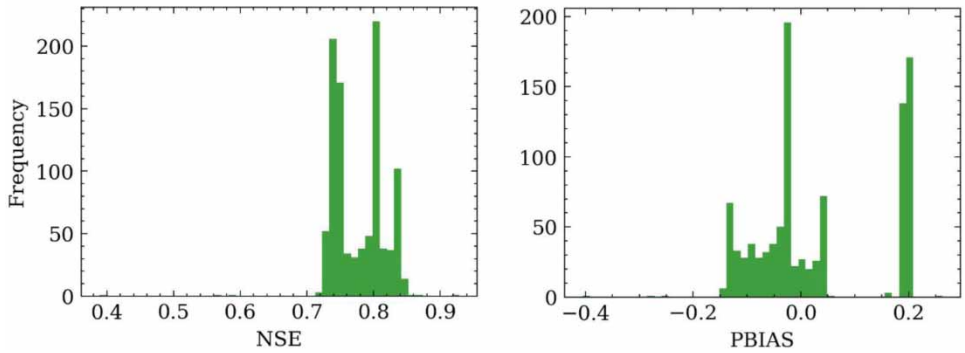

Figure 6 | Nash–Sutcliffe efficiency (NSE) and percent bias (PBIAS) of the simulated streamflow of the PCE using 1,000 MC experiments.

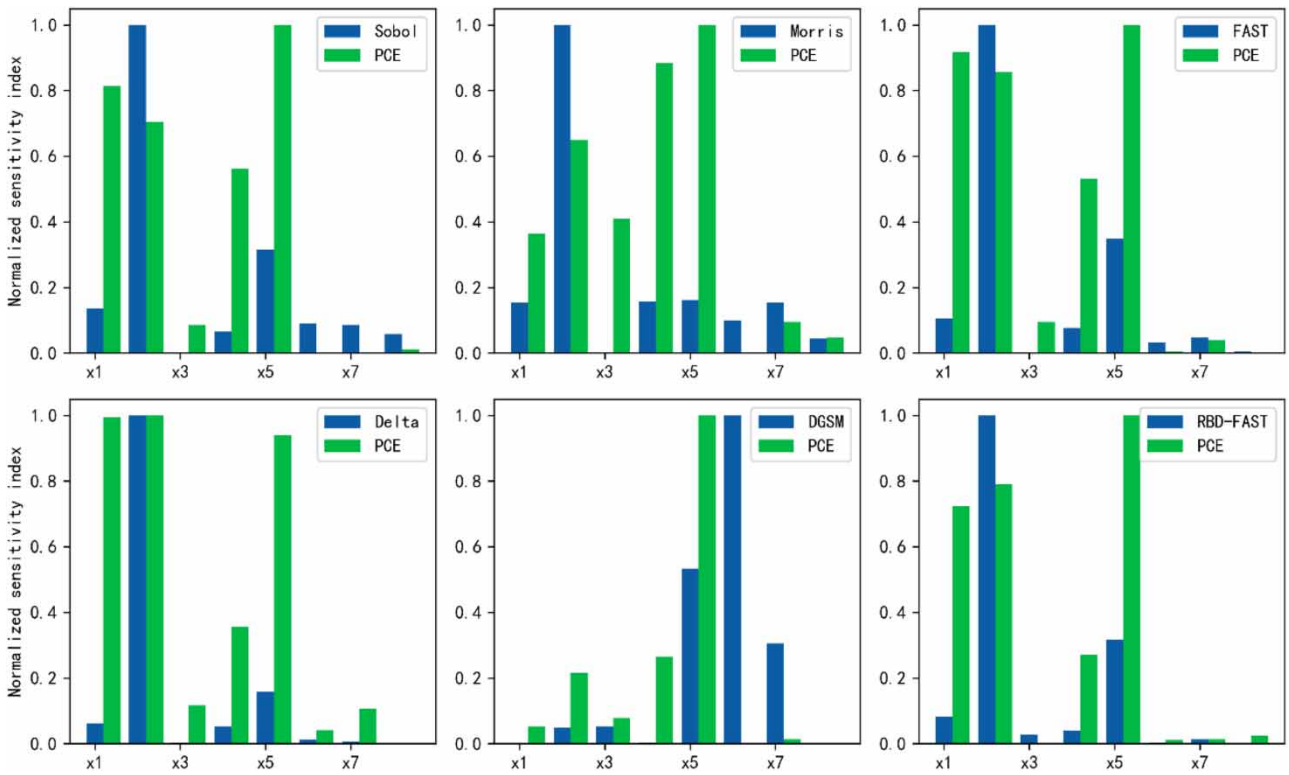

Figure 7 | Parameter sensitivity analysis comparing six original SWAT models with their corresponding PCE surrogate models. *x*1–*x*8 denote the parameters CANMX, CN2, CH_N2, CH_K2, ALPHA_BNK, SOL_AWC, SOL_K, and SOL_BD, respectively.

recognizable and that the other five parameters are hardly recognizable or not recognizable because different values of these parameters may produce comparable effects when combined with the other parameters. This finding also shows that the PCE model is vulnerable to the consequences of equifinality, with alternative calibration procedures using different sets of parameters producing equivalent simulations (Beven 2006), which is common in complex hydrological nonlinear models. This result confirms the capability of the PCE model to approximate complex models such as SWAT, i.e., the PCE inherits the equifinality property of complex models.

4.3. Reduction of the computation time

A previous study demonstrated that 100 tasks (SWAT executions) took an average of 198 s (Ma *et al.* 2022a). Conversely, a single task took an average of 112 s when using a single core. Thus, when 100 cores were used, an average speedup of 57

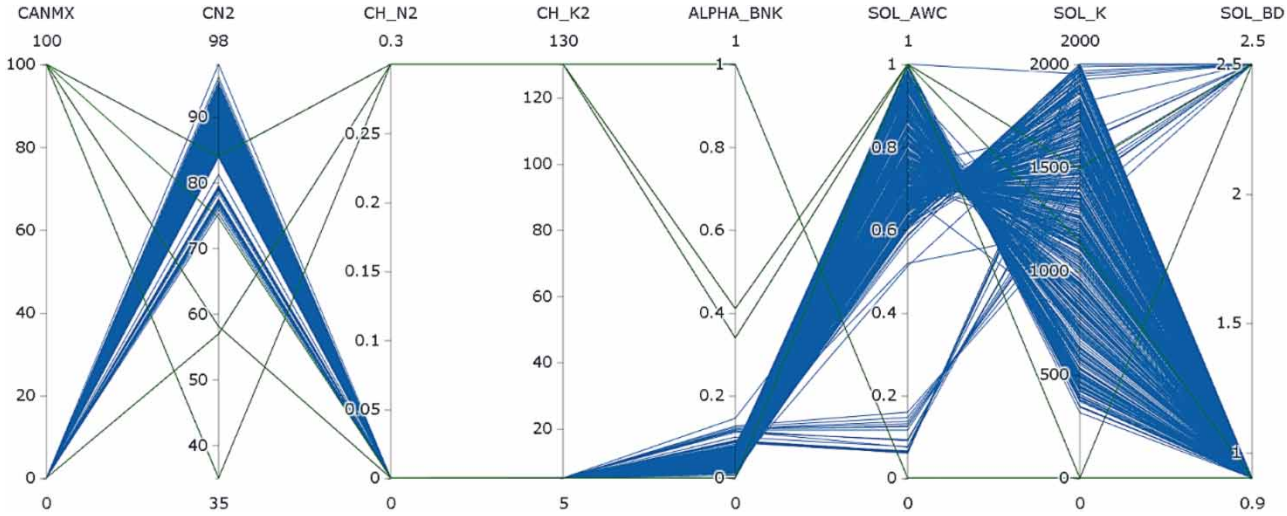

Figure 8 | Parallel coordinates plot of 1,000 optimal solutions using the global BOBYQA optimization algorithm. For each solution, the initial parameter set is randomly sampled according to their ranges. The Nash–Sutcliffe efficiency (NSE) of simulated streamflow is used as the objective function. NSE values over 0.95 are represented by the dark green colour, while values between 0.65 and 0.96 are shown by the blue colour. Please refer to the online version of this paper to see this figure in colour: http://dx.doi.org/10.2166/hydro.2023.184.

was obtained, which is faster than the traditional standalone model evaluation. The global SA in this case study required 47,104 model evaluations to ensure convergence of the sensitivity indices of SA and took 1,584 min on the Hadoop cloud with 100 cores using the original model. In contrast, the PCE model required 6,561 evaluations of the original model and took just 22 min. Furthermore, because of the fast processing performance of the PCE, 47,104 surrogate SA assessments took approximately 7 s on a standalone machine with an Intel 16-core i7-10700 processor and 32 GB RAM. Overall, the proposed global SA process reduced the number of model evaluations by 40,543 and the computation time by approximately 86%.

5. DISCUSSION

A hydrological simulation framework usually consists of three key parts: HPC, OAs, and a complex simulation model. Despite significant progress in the first two parts, the development of such a framework is still constrained by the inherent computational complexity of the simulation models. For this reason, the emphasis seems to have shifted from the first two parts to the simulation models themselves. Consequently, SM is being widely investigated as a viable solution because it is believed to preserve the high accuracy of the original model while significantly decreasing the computational costs. Within SM, the DOE is deemed to be the most computationally expensive part because it requires nearly the whole computing budget to solve the problem (Razavi *et al.* 2012b). To date, no studies have fully addressed the DOE implementation process. This deficiency has hampered the adoption of SM methods that may aid in simulation optimization. We have extended the previously presented framework to address the DOE process of SM implementation using cloud computing techniques (Ma *et al.* 2022a). The framework allows substantial high-fidelity model evaluations of DOEs to be implemented elegantly on the Hadoop cloud. The case study uses the PCE model to surrogate the SWAT model, demonstrating how SM-based PCE and cloud computing methodologies could be exploited to successfully counterbalance the very large computational burden of complex hydrological model simulations. Moreover, the Hadoop cloud-based framework can provide an ideal environment for assessing the applicability of SM, which is critical before SM is used. The approximation capability of PCEs in surrogate SWAT models was examined within the framework. The case study demonstrated how SM libraries, such as Chaospy, and SA libraries, such as SALIB, can be conveniently coupled with cloud computing to perform approximation analysis regarding the accuracy, fidelity, and efficiency.

Previous SM-related research efforts have focused on either Step 1 (described in Section 3.3): constructing and using function approximation methods as substitutes or Step 3: constructing frameworks that utilize surrogates (Razavi *et al.* 2012b). To date, there has been little research focusing on Step 1 of SM (DOE), mostly because it is commonly thought to be biased

towards computational aspects rather than SM itself. In practice, the DOE consumes nearly the whole computing budget allocated to perform SM. This was confirmed by our case study, in which 6,561 original model evaluations required by the DOE took 22 min, whereas 47,104 surrogate SA assessments took only approximately 7 s. These observations were also reported in other studies (Westermann & Evins 2019; Chen *et al.* 2020; Zhang 2020). Notable examples include the development of a surrogate method of groundwater modelling using a gated recurrent unit, in which 2,000 evaluations of the original models required to fit the surrogate took nearly 8.2 h, whereas 70,000 evaluations of the surrogate took nearly 3 min (Chen *et al.* 2020). In this regard, our framework is the first of its kind to introduce cloud computing to address the computational burden of DOEs. Various DOE strategies may be readily explored within the framework to facilitate the construction of the appropriate SM model, whose objective may range from finding the optimum to mapping out the whole surface of the original model. Furthermore, reports of previous studies provided little information on the modelling and analysis toolkits or libraries that they used. Conversely, our framework explicitly introduces SM-related libraries, such as Chaospy and SALib, and thereby provides environmental modellers who are considering SM with a more complete description of the components of the SM process, as well as recommendations for the subjective choices needed when using surrogate models for hydrological simulations.

A significant number of model simulations are typically required by simulation optimization frameworks that combine model simulations with OAs. Despite the extensive effort that has been devoted to improving HPC and OAs, one such approach could be challenging when the single-model simulation is computationally expensive. A promising framework would therefore alleviate the very large computational burden as much as possible. Using the Chaospy library, our framework can successfully surrogate the SWAT model with a second-order PCE and then examine the approximation capability of the PCE with SALib. This framework is particularly suitable for addressing the DOE problem because it inherits the parallel processing feature of cloud computing. Separating the DOE from the other two steps of SM is beneficial for programming because it allows the DOE to be extended to more complex models. Because it adopts the Python-based web framework to host the main SM library Chaospy and core SA library SALIB, the framework can be flexibly extended to other Python-based SM-related libraries and statistical analysis libraries, such as DAKOTA (Eldred *et al.* 2010), Open TURNS (Andrianov *et al.* 2007), and the Surrogate Modelling Toolbox (Hwang *et al.* 2019). Therefore, it is a generic framework for tackling a wide range of SM tasks and is not specifically limited to PCE-based SM and global SA.

Despite its advantages, the proposed framework has two potential drawbacks. One drawback is that it is currently a basic sequential framework (also known as an offline framework) and is not an adaptive–recursive framework (also known as an online framework). The basic sequential framework – the simplest type of SM analysis framework – consists of three main components: DOE, global fitting on the DOE to construct a surrogate, and replacement for the original model in conducting the relevant analyses (Razavi *et al.* 2012b). This indicates that the size of the DOE in Step 1 of the offline framework is often fixed and larger than that of the initial DOE in more advanced online SM frameworks. Because the DOE in Step 1 of the offline framework consumes nearly all computing resources allocated to address the problem, the other frameworks may be thought of as being online in this sense since they update the surrogate regularly as new data become available. Currently, only one DOE is performed in our framework to fit the PCE, and PCE updates are not yet supported. This might raise the concern that the PCE method is sometimes not good at identifying sensitive parameters or locating the optimum because the surrogate model is smoother than the true response surface, which is unavoidable due to the information limitation of a 'once-for-all' DOE – a limited number of samples. Notwithstanding this drawback, our framework can support the online framework through an appropriate orchestration of continuous DOEs, as required by an online framework, and earlier studies have provided us with insight in this regard (Ma *et al.* 2022b). Therefore, establishing an adaptive sampling technique capable of detecting and refining the informative region will be a critical task in the future. Another possible drawback of this framework comes from the assumption that the input variables are stochastically independent; however, this is not often the case. Reformulating the issue as a set of independent variables using GPCE is one approach for quantitatively addressing stochastically dependent variables (Xiu *et al.* 2002). Because the computational architecture of Chaospy includes the Rosenblatt transformation, which allows for mapping any domain to the unit hypercube $[0, 1]^D$, GPCE is available for all densities. Future research that incorporates GPCE within the framework is thus recommended. Furthermore, the framework provides a ideal computational environment for investigating whether alternative time series prediction-based machine learning models (Mosavi *et al.* 2018; Kaya *et al.* 2019; Fu *et al.* 2020; Ehteram *et al.* 2022; Wang *et al.* 2022; Xie *et al.* 2022) could be used as quick and accurate surrogate models.

6. CONCLUSIONS

In this paper, a Hadoop cloud-based SM framework is presented for approximating complex hydrological models, with the goals of improving the computational efficiency of DOEs that commonly require substantial model evaluations and of evaluating the applicability of SM before it is used. The framework is the first of its kind to solve the problem of the very large computational burden of DOEs on the Hadoop cloud. Because it employs a Python web framework to host both the SM libraries and other statistical analysis libraries, the framework is also capable of providing an ideal environment for assessing the applicability of SM with regards to the accuracy, fidelity, and efficiency. The case study using PCEs to approximate the SWAT model indicates that the framework is capable of solving the DOE problem as well as evaluating the applicability of SM. The framework is particularly well suited to complex hydrological model optimization, but it may also be extended to solve various optimization problems. It is expected to provide hydrological modellers who are contemplating SM with a more detailed description of the components of the process than what was previously available, as well as a new perspective on how to make the necessary subjective decisions when using surrogate models.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (41925005) and the National Key R&D Program of China (2019YFD0901105).

DATA AVAILABILITY STATEMENT

All relevant data are available from an https://github.com/JinfengM/PCE-Hadoop.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Andrianov, G., Burriel, S., Cambier, S., Dutfoy, A. & Pendola, M. 2007 Open TURNS, an open source initiative to Treat Uncertainties, Risks'N Statistics in a structured industrial approach. In: Proceedings of the European Safety and Reliability Conference 2007, ESREL 2007, Stavanger, Norway.

Asadzadeh, M. & Tolson, B. 2013 Pareto archived dynamically dimensioned search with hypervolume-based selection for multi-objective optimization. *Engineering Optimization* **45**, 1489–1509.

Behzadian, K., Kapelan, Z., Savic, D. & Ardeshir, A. 2009 Stochastic sampling design using a multi-objective genetic algorithm and adaptive neural networks. *Environmental Modelling & Software* **24**, 530–541.

Beven, K. 2006 A manifesto for the equifinality thesis. *Journal of Hydrology* **320**, 18–36.

Black, D. C., Wallbrink, P. J. & Jordan, P. W. 2014 Towards best practice implementation and application of models for analysis of water resources management scenarios. *Environmental Modelling & Software* **52**, 136–148.

Blatman, G. & Sudret, B. 2008 Sparse polynomial chaos expansions and adaptive stochastic finite elements using a regression approach. *Comptes Rendus Mécanique* **336**, 518–523.

Borgonovo, E., Castaings, W. & Tarantola, S. 2012 Model emulation and moment-independent sensitivity analysis: an application to environmental modeling. *Environmental Modelling & Software* **34**, 105–115.

Chen, M., Izady, A., Abdalla, O. A. & Amerjeed, M. 2018 A surrogate-based sensitivity quantification and Bayesian inversion of a regional groundwater flow model. *Journal of Hydrology* **557**, 826–837.

Chen, Y., Liu, G., Huang, X., Chen, K., Hou, J. & Zhou, J. 2020 Development of a surrogate method of groundwater modeling using gated recurrent unit to improve the efficiency of parameter auto-calibration and global sensitivity analysis. *Journal of Hydrology* **598**, 125726.

Deman, G., Konakli, K., Sudret, B., Kerrou, J., Perrochet, P. & Benabderrahmane, H. 2016 Using sparse polynomial chaos expansions for the global sensitivity analysis of groundwater lifetime expectancy in a multi-layered hydrogeological model. *Reliability Engineering & System Safety* **147**, 156–169.

Ehteram, M., Kalantari, Z., Ferreira, C. S., Chau, K.-w. & Emami, S.-M.-K. 2022 Prediction of future groundwater levels under representative concentration pathway scenarios using an inclusive multiple model coupled with artificial neural networks. *Journal of Water and Climate Change* **13**, 3620–3643.

Eldred, M. S., Dalbey, K. R., Bohnhoff, W. J., Adams, B. M., Swiler, L. P., Hough, P. D., Gay, D. M., Eddy, J. P. & Haskell, K. H. 2010 *DAKOTA: A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis. Version 5.0, User's Manual*.

Ercan, M. B. & Goodall, J. L. 2016 Design and implementation of a general software library for using NSGA-II with SWAT for multi-objective model calibration. *Environmental Modelling & Software* **84**, 112–120.

Fan, Y. R., Huang, W., Huang, G. H., Huang, K. & Zhou, X. 2015 A PCM-based stochastic hydrological model for uncertainty quantification in watershed systems. *Stochastic Environmental Research and Risk Assessment* **29**, 915–927.

Feinberg, J. & Langtangen, H. P. 2015 Chaospy: an open source tool for designing methods of uncertainty quantification. *Journal of Computational Science* **11**, 46–57.

Fen, C.-S., Chan, C. & Cheng, H.-C. 2009 Assessing a response surface-based optimization approach for soil vapor extraction system design. *Journal of Water Resources Planning and Management* **135**, 198–207.

Fernandez-Palomino, C. A., Hattermann, F. F., Krysanova, V., Vega-Jácome, F. & Bronstert, A. 2021 Towards a more consistent eco-hydrological modelling through multi-objective calibration: a case study in the Andean Vilcanota River basin, Peru. *Hydrological Sciences Journal* **66**, 59–74.

Franco, A. C. L., Oliveira, D. Y. d. & Bonumá, N. B. 2020 Comparison of single-site, multi-site and multi-variable SWAT calibration strategies. *Hydrological Sciences Journal* **65**, 2376–2389.

Friedman, J. H. 1991 Multivariate adaptive regression splines. *Annals of Statistics* **19**, 1–67.

Fu, M., Fan, T., Ding, Z. 'a., Salih, S. Q., Al-Ansari, N. & Yaseen, Z. M. 2020 Deep learning data-intelligence model based on adjusted forecasting window scale: application in daily streamflow simulation. *IEEE Access* **8**, 32632–32651.

Ghaith, M. & Li, Z. 2020 Propagation of parameter uncertainty in SWAT: a probabilistic forecasting method based on polynomial chaos expansion and machine learning. *Journal of Hydrology* **586**, 124854.

Gottlieb, D. & Orszag, S. A. 1977 *Numerical analysis of spectral methods: Theory and applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Hadka, D. & Reed, P. 2013 Borg: an auto-adaptive many-objective evolutionary computing framework. *Evolutionary Computation* **21**, 231–259.

Hadka, D. & Reed, P. 2015 Large-scale parallelization of the Borg multiobjective evolutionary algorithm to enhance the management of complex environmental systems. *Environmental Modelling & Software* **69**, 353–369.

Herman, J. & Usher, W. 2017 SALib: an open-source python library for sensitivity analysis. *JOSS* **2**, 97.

Hosder, S., Walters, R. & Balch, M. 2007 Efficient sampling for non-intrusive polynomial chaos applications with multiple uncertain input variables. In: *48th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*. American Institute of Aeronautics and Astronautics, Reston, Virigina.

Hu, Y., Cai, X. & DuPont, B. 2015 Design of a web-based application of the coupled multi-agent system model and environmental model for watershed management analysis using Hadoop. *Environmental Modelling & Software* **70**, 149–162.

Huang, S., Mahadevan, S. & Rebba, R. 2007 Collocation-based stochastic finite element analysis for random field problems. *Probabilistic Engineering Mechanics* **22**, 194–205.

Hussain, M. F., Barton, R. R. & Joshi, S. B. 2002 Metamodeling: radial basis functions, versus polynomials. *European Journal of Operational Research* **138**, 142–154.

Hwang, J. T., Bartoli, N., Lafage, R., Morlier, J. & Martins, J. R. R. A. 2019 A Python surrogate modeling framework with derivatives. *Advances in Engineering Software* **135**, 1–13.

Jin, R., Chen, W. & Simpson, T. W. 2001 Comparative studies of metamodelling techniques under multiple modelling criteria. *Structural and Multidisciplinary Optimization* **23**, 1–13.

Kaya, C. M., Tayfur, G. & Gungor, O. 2019 Predicting flood plain inundation for natural channels having no upstream gauged stations. *Journal of Water and Climate Change* **10**, 360–372.

Keating, E. H., Doherty, J., Vrugt, J. A. & Kang, Q. 2010 Optimization and uncertainty assessment of strongly nonlinear groundwater models with high parameter dimensionality. *Water Resources Research* **46**, 10517.

Laloy, E., Rogiers, B., Vrugt, J. A., Mallants, D. & Jacques, D. 2013 Efficient posterior exploration of a high-dimensional groundwater model from two-stage Markov chain Monte Carlo simulation and polynomial chaos expansion. *Water Resources Research* **49**, 2664–2682.

Ma, J., Rao, K., Li, R., Yang, Y., Li, W. & Zheng, H. 2022a Improved Hadoop-based cloud for complex model simulation optimization: calibration of SWAT as an example. *Environmental Modelling & Software* **149**, 105330.

Ma, J., Zhang, J., Li, R., Zheng, H. & Li, W. 2022b Using Bayesian optimization to automate the calibration of complex hydrological models: framework and application. *Environmental Modelling & Software* **147**, 105235.

Maier, H. R., Guillaume, J. H. A., van Delden, H., Riddell, G. A., Haasnoot, M. & Kwakkel, J. H. 2016 An uncertain future, deep uncertainty, scenarios, robustness and adaptation: how do they fit together? *Environmental Modelling & Software* **81**, 154–164.

Maier, H. R., Razavi, S., Kapelan, Z., Matott, L. S., Kasprzyk, J. & Tolson, B. A. 2019 Introductory overview: optimization using evolutionary algorithms and other metaheuristics. *Environmental Modelling & Software* **114**, 195–213.

Mckay, M. D., Beckman, R. J. & Conover, W. J. 2000 A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **42**, 55–61.

Mosavi, A., Ozturk, P. & Chau, K.-w. 2018 Flood prediction using machine learning models: literature review. *Water* **10**, 1536.

Mugunthan, P., Shoemaker, C. A. & Regis, R. G. 2005 Comparison of function approximation, heuristic, and derivative-based methods for automatic calibration of computationally expensive groundwater bioremediation models. *Water Resources Research* **41**, 11427.

Paffrath, M. & Wever, U. 2007 Adapted polynomial chaos expansion for failure detection. *Journal of Computational Physics* **226**, 263–281.

Papadrakakis, M., Lagaros, N. D. & Tsompanakis, Y. 1998 Structural optimization using evolution strategies and neural networks. *Computer Methods in Applied Mechanics and Engineering* **156**, 309–333.

Powell, M. J. 2009 The BOBYQA algorithm for bound constrained optimization without derivatives. Cambridge NA Report NA2009/06, University of Cambridge, Cambridge, UK.

Razavi, S., Tolson, B. A. & Burn, D. H. 2012a Numerical assessment of metamodelling strategies in computationally intensive optimization. *Environmental Modelling & Software* **34**, 67–86.

Razavi, S., Tolson, B. A. & Burn, D. H. 2012b Review of surrogate modeling in water resources. *Water Resources Research* **48**, 07401.

Regis, R. G. & Shoemaker, C. A. 2007 A stochastic radial basis function method for the global optimization of expensive functions. *INFORMS Journal on Computing* **19**, 497–509.

Ritter, A. & Muñoz-Carpena, R. 2013 Performance evaluation of hydrological models: statistical significance for reducing subjectivity in goodness-of-fit assessments. *Journal of Hydrology* **480**, 33–45.

Sacks, J., Welch, W. J., Mitchell, T. J. & Wynn, H. P. 1989 Design and analysis of computer experiments. *Statistical Science* **4**, 409–423.

Sakata, S., Ashida, F. & Zako, M. 2003 Structural optimization using Kriging approximation. *Computer Methods in Applied Mechanics and Engineering* **192**, 923–939.

Shi, L., Yang, J., Zhang, D. & Li, H. 2009 Probabilistic collocation method for unconfined flow in heterogeneous media. *Journal of Hydrology* **365**, 4–10.

Stavropoulou, F. & Müller, J. 2015 Parametrization of random vectors in polynomial chaos expansions via optimal transportation. *SIAM Journal on Scientific Computing* **37**, A2535–A2557.

Tayfur, G. 2017 Modern optimization methods in water resources planning, engineering and management. *Water Resources Management* **31**, 3205–3233.

van Griensven, A. & Meixner, T. 2006 Methods to quantify and identify the sources of uncertainty for river basin water quality models. *Water Science and Technology: A Journal of the International Association on Water Pollution Research* **53**, 51–59.

Villegas, M., Augustin, F., Gilg, A., Hmaidi, A. & Wever, U. 2012 Application of the polynomial chaos expansion to the simulation of chemical reactors with uncertainties. *Mathematics and Computers in Simulation (MATCOM)* **82**, 805–817.

Wang, G. C., Zhang, Q., Band, S. S., Dehghani, M., Chau, K. w., Tho, Q. T., Zhu, S., Samadianfard, S. & Mosavi, A. 2022 Monthly and seasonal hydrological drought forecasting using multiple extreme learning machine models. *Engineering Applications of Computational Fluid Mechanics* **16**, 1364–1381.

Wang, S., Huang, G. H., Huang, W., Fan, Y. R. & Li, Z. 2015 A fractional factorial probabilistic collocation method for uncertainty propagation of hydrologic model parameters in a reduced dimensional space. *Journal of Hydrology* **529**, 1129–1146.

Westermann, P. & Evins, R. 2019 Surrogate modelling for sustainable building design – a review. *Energy and Buildings* **198**, 170–186.

White, T. 2012 *Hadoop: The Definitive Guide*. O'Reilly Media, Inc., Farnham, NY.

Wiener, N. 1938 The homogeneous chaos. *American Journal of Mathematics* **60**, 897.

Wu, X., Zhang, W., Song, S. & Ye, Z. 2018 Sparse grid-based polynomial chaos expansion for aerodynamics of an airfoil with uncertainties. *Chinese Journal of Aeronautics* **31**, 997–1011.

Xie, H., Randall, M. & Chau, K.-w. 2022 Green roof hydrological modelling with GRU and LSTM networks. *Water Resources Management* **36**, 1107–1122.

Xiu, D., Lucor, D., Su, C.-H. & Karniadakis, G. E. 2002 Stochastic modeling of flow-Structure interactions using generalized polynomial chaos. *Journal of Fluids Engineering* **124**, 51–59.

Zhang, W. 2020 *MARS Applications in Geotechnical Engineering Systems*. Springer Singapore, Singapore.

Zhang, D., Chen, X., Yao, H. & James, A. 2016 Moving SWAT model calibration and uncertainty analysis to an enterprise Hadoop-based cloud. *Environmental Modelling & Software* **84**, 140–148.

First received 25 October 2022; accepted in revised form 25 February 2023. Available online 10 March 2023