Journal of Hydrology 564 (2018) 697–711

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

A systematic assessment and reduction of parametric uncertainties for a distributed hydrological model

Yanjun Gan[a,*], Xin-Zhong Liang[b,c], Qingyun Duan[d,*], Aizhong Ye[d], Zhenhua Di[d], Yang Hong[e], Jianduo Li[a]

[a] State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing, China
[b] Climate, Environment and Sustainability Center, Nanjing University of Information Science and Technology, Nanjing, Jiangsu, China
[c] Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD, USA
[d] State Key Laboratory of Earth Surface Processes and Resource Ecology, Faculty of Geographical Science, Beijing Normal University, Beijing, China
[e] School of Civil Engineering and Environmental Sciences, University of Oklahoma, Norman, OK, USA

ARTICLE INFO

This manuscript was handled by A. Bardossy, Editor-in-Chief, with the assistance of Roger Moussa, Associate Editor

Keywords:
Hydrological modeling
Uncertainty quantification
Sensitivity analysis
Surrogate model
Parameter optimization

ABSTRACT

Quantifying and reducing uncertainties in physics-based hydrological model parameters will improve model reliability for hydrological forecasting. We present an uncertainty quantification framework that combines the strengths of stepwise sensitivity analysis and adaptive surrogate-based multi-objective optimization to facilitate practical assessment and reduction of model parametric uncertainties. Framework performance was tested using the distributed hydrological model Coupled Routing and Excess Storage (CREST) for daily streamflow simulation over ten watersheds. By identifying sensitive parameters stepwisely, we reduced the number of parameters requiring calibration from twelve to seven, thus limiting the dimensionality of calibration problem. By updating surrogate models adaptively, we found the optimal sets of sensitive parameters with the surrogate-based multi-objective optimization. The calibrated CREST was able to satisfactorily simulate observed streamflow for all watersheds, improving one minus Nash-Sutcliffe efficiency (1 − NSE) by 65–90% and percentage absolute relative bias (|RB|) by 60–95% compared to the default. The validation result demonstrated that the calibrated CREST was also able to reproduce observed streamflow outside the calibration period, improving 1 − NSE by 40–85% and |RB| by 35–90% compared to the default. Overall, this uncertainty quantification framework is effective for assessment and reduction of model parametric uncertainties, the results of which improve model simulations and enhance understanding of model behaviors.

1. Introduction

Hydrological models have become indispensable tools for both operational applications such as flood forecasting and water resource management decision support, and scientific research to understand hydrological processes (Kavetski and Clark, 2010). The accuracy of model predictions is greatly influenced by uncertainties such as data errors and model deficiencies (Renard et al., 2010). Prediction uncertainties are mainly due to model parameters, if model structures are supposed to be correct and input data and observed data are assumed to be error-free (Gupta et al., 2005; Gourley and Vieux, 2006; Matonse and Kroll, 2013). Appropriate calibration could reduce these parametric uncertainties to enable the models to better represent the real systems and thus provide faithful predictions (Vrugt et al., 2003; Duan et al., 2017; Sikorska and Renard, 2017).

Model parameters are calibrated by tuning parameter values to minimize differences between simulated and observed variables (Duan et al., 1992). Practical experience suggests that since model calibration problem is inherently multi-objective, a set of non-commensurate criteria measuring different aspects of the system should be considered simultaneously (Gupta et al., 1998; Khu and Madsen, 2005; Moussa and Chahinian, 2009). The multi-objective calibration of hydrological models is often time-consuming and requires many model runs to determine optimal sets (Efstratiadis and Koutsoyiannis, 2010). The computational burden is further increased by the growing complexity of hydrological models, which include more sub-physical processes and represent spatial heterogeneities of various hydrological processes (Fatichi et al., 2016). In addition, complex models include more parameters, but still are calibrated on limited observational data, which has led to an "over-parameterization" problem (Jakeman and Hornberger,

* Corresponding authors.
E-mail addresses: yjgan@cma.gov.cn (Y. Gan), qyduan@bnu.edu.cn (Q. Duan).

https://doi.org/10.1016/j.jhydrol.2018.07.055
Received 7 May 2018; Received in revised form 12 July 2018; Accepted 20 July 2018
Available online 24 July 2018

1993). Furthermore, the highly nonlinear and complex interactions of the hydrological processes result in many acceptable parameter sets that produce equal model performances, known as "equifinality" (Beven, 2006; Stedinger et al., 2008), which also increases the difficulty of finding the optimal sets. Therefore, the efficient calibration of computationally-intensive distributed hydrological models under a multi-objective framework is a difficult exercise with great uncertainty (Zhang et al., 2009a; Shi et al., 2014).

One way to alleviate the computational burden and uncertainty associated with model calibration is to reduce parameter dimensionality before calibration. Sensitivity analyses can identify parameters that dominate model behavior (Gan et al., 2014), which allows models to be simplified without affecting the reliability of simulation by setting insensitive parameters at fixed values within their feasible ranges (Saltelli et al., 2008). Sensitivity analysis methods can be qualitative or quantitative. The qualitative methods assess parameter sensitivity by visualization tools such as scatter plots showing the parameter-response relationships, or by relative simple measures such as the screening methods. Whereas the quantitative methods quantify each parameter's contribution to the response variance using variance-based measures (Pianosi et al., 2016). A comprehensive evaluation of various qualitative and quantitative sensitivity analysis methods can be found in Gan et al. (2014). Their comparisons show that the latter are more accurate and robust but less efficient than the former.

Another way to ease the computational burden is to use surrogate models to approximate the more expensive hydrological models for quantitative sensitivity analysis and model calibration. The surrogate model is constructed based on parameter-response relationships obtained from simulations with the hydrological model (Marie and Simioni, 2014). The way of using surrogate modeling methods can be sequential or adaptive-recursive. The sequential framework uses the globally fitted surrogate model for a fully substitution of the original model. Whereas the adaptive-recursive framework adaptively update the globally/locally fitted surrogate model and recursively use the newly build surrogate model for substitution of the original model (Razavi et al., 2012). The former is more appropriate for quantitative sensitivity analysis while the latter is more efficient for parameter optimization. A variety of surrogate modeling methods are reviewed and compared in Villa-Vialaneix et al. (2012) and Razavi et al. (2012).

Most previous studies have used either sensitivity analyses or surrogate models to facilitate model calibration. For example, Muleta and Nicklow (2005), Foglia et al. (2009) and van Werkhoven et al. (2009) conducted sensitivity analyses to reduce parameter dimensionality before model calibration. Wang et al. (2014), Gong et al. (2015), and Chen et al. (2017) adopted surrogate-based optimization methods to improve computational efficiency for model calibration. Some studies have used both sensitivity analyses and surrogate models in the way of surrogate-based sensitivity analyses (Shahsavani and Grimvall, 2011; Borgonovo et al., 2012; Daneshkhah and Bedford, 2013).

We combined the strengths of the state-of-the-art methods for multi-objective calibration in a systematic stepwise way: (1) qualitative sensitivity analysis for parameter screening of a high-dimensional model; (2) sequential surrogate-based quantitative sensitivity analysis for identifying the most sensitive parameters of the dimension-reduced model; and (3) adaptive surrogate-based optimization for calibrating the most sensitive parameters. This stepwise way would save considerable computational budget for assessment and reduction of parametric uncertainties. We conducted hydrological simulation experiments to construct surrogate models that can be used both to facilitate quantitative sensitivity analysis for parameter reduction and to provide reliable information for promising regions of the global optimum. This reuse of the hydrological simulation experiments would greatly reduce computational cost, since the majority of the computational budget for surrogate modeling is the original simulation model run.

The main objective of this study is to assess and reduce parametric uncertainties for the distributed hydrological model Coupled Routing and Excess Storage (CREST) (Wang et al., 2011) with a systematic uncertainty quantification framework. Using this framework, we demonstrate a method to improve daily streamflow simulation effectively in the presence of parametric uncertainties, for the CREST over ten representative watersheds with varying hydroclimatic, soil, and vegetation conditions. During this process, we also try to understand model behaviors for heterogeneous spatial domain across these watersheds.

2. Model, experimental data and setup

2.1. CREST model

CREST was jointly developed by the University of Oklahoma (http://hydro.ou.edu) and National Aeronautics and Space Administration (NASA) SERVIR (named for the Spanish verb "to serve") Project Team (http://www.servir.net), to simulate the spatiotemporal variation of land surface and subsurface water fluxes and storages (Wang et al., 2011). Its main features include: (1) coupling between distributed rainfall-runoff generation and cell-to-cell routing processes via three feedback mechanisms; and (2) scalability by a representation of subgrid soil moisture variability and routing processes. Further model details can be found in Wang et al. (2011). CREST has been successfully implemented in a variety of multiscale hydrological studies (Khan et al., 2011; Wu et al., 2012; Xue et al., 2013; Tang et al., 2016) and several flood warning systems including the Flooded Locations And Simulated Hydrographs Project (FLASH; http://blog.nssl.noaa.gov/flash/) and the Global Flood Monitoring System (GFMS; http://eos.ou.edu).

Although CREST has been widely used, its parametric uncertainties have not yet been analyzed and its model behaviors across watersheds have not been fully understood. Previous works on the calibration of CREST are often performed for all parameters, which takes a high computational cost for this distributed hydrological model (Khan et al., 2011; Wu et al., 2012; Xue et al., 2013). Consequently, we would like to identify the most important parameters of CREST and simplify it for future use. Moreover, almost all calibration works of CREST are performed with single-objective for the watersheds outside of China, experiences from which are insufficient and nontransferable for its use in China domain. Therefore, we would also like to make a systematic multi-objective calibration of the most important parameters for representative watersheds in China. The latest version CREST v2.1 was used in this study to evaluate the impact of model parameters on streamflow simulation across watersheds over the China domain at a 0.125-degree latitude-longitude grid.

2.2. Experimental data

2.2.1. Study area and evaluation data sets

We selected one representative watershed for each of the ten large river basins in China, which span different hydroclimatic, soil, and vegetation conditions. Fig. 1 shows the geographic location, elevation, and streamflow gauging station of each representative watershed. At these stations, observed daily streamflow discharge data from 1 January 2008 to 31 December 2012 were acquired from the hydrological yearbooks published by the Ministry of Water Resources of China. Table 1 summarizes the station and watershed characteristics, as well as watershed-averaged mean daily precipitation (P), potential evapotranspiration (PET), and station mean daily streamflow discharge (SD) for the observation period.

2.2.2. Hydrography data

The digital elevation model (DEM) was obtained from the 30 arc-second HydroSHEDS (Hydrological data and map based on SHuttle Elevation Derivatives at multiple Scales) product (available at http://hydrosheds.cr.usgs.gov/) (Lehner et al., 2008). The raw data were aggregated by neighborhood averaging onto a 0.125-degree grid, and

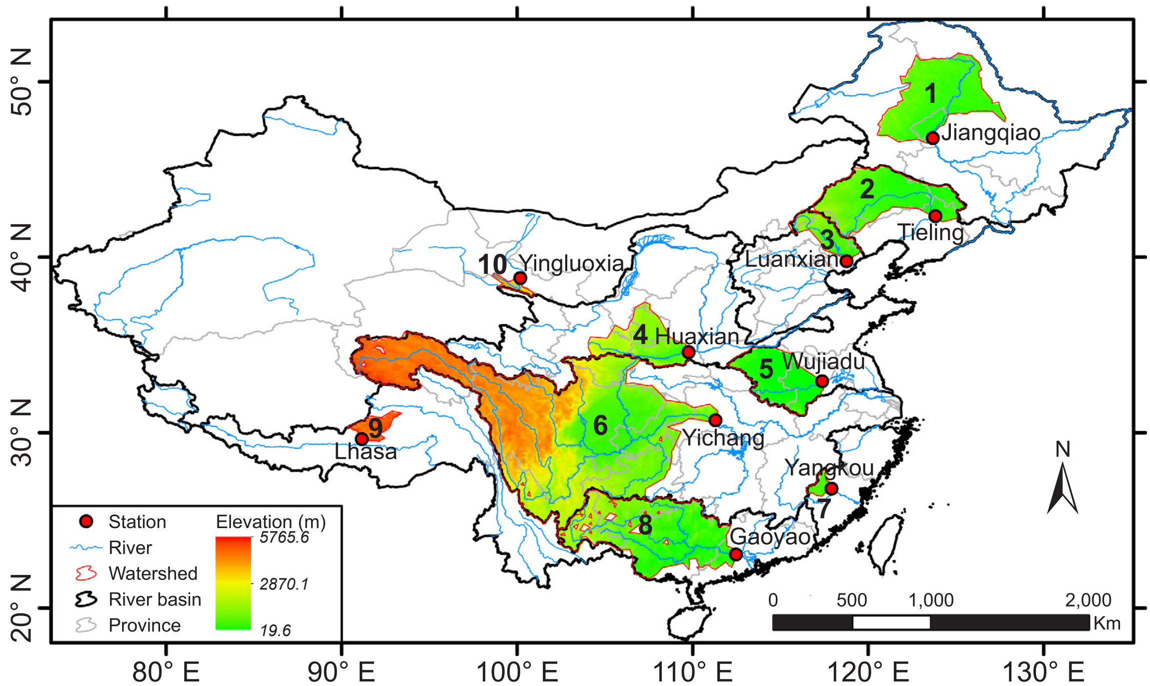

Fig. 1. Model domain showing geographic location, elevation, and streamflow gauging station of the representative watershed over ten large river basins of China.

then corrected by algorithms such as stream burning and sink filling. Based on the corrected DEM data, a drainage direction map followed by a flow accumulation map were generated using the eight-direction (D8) flow model (Jenson and Domingue, 1988). A river network map was extracted by considering the grids with flow accumulation values (in number of grids) greater than or equal to 50 using the raster calculator tool.

2.2.3. *Forcing data*

A gauge-satellite merged precipitation product with a daily 0.25 deg × 0.25 deg resolution was used to drive CREST (available at http://data.cma.cn/). It was generated using hourly rain gauge data at more than 30,000 automatic weather stations in China, in conjunction with the Climate Precipitation Center Morphing (CMORPH) precipitation product (Shen et al., 2014). The daily 1 deg × 1 deg global PET product was acquired from the Famine Early Warning Systems Network (FEWS NET; available at http://earlywarning.usgs.gov/fews). It was estimated by the Penman-Monteith equation using climate data extracted from Global Data Assimilation System (GDAS). Both precipitation and PET data covering the period from 1 January 2007 to 31 December 2012 were remapped onto the 0.125 deg × 0.125 deg study domain by bilinear interpolation.

2.3. *Experimental setup*

The control version CREST with default parameter values was run for a spin-up period of six years (2007–2012) at a daily time step to achieve an equilibrium state. All subsequent experiments were initialized from this state at the end of the spin-up. The experiments for qualitative and quantitative sensitivity analysis as well as parameter optimization were run from 1 January 2007 to 31 December 2010, while those for validation were run from 1 January 2010 to 31 December 2012. To minimize the influence of initial conditions, the first year's outputs of all experiments were excluded from subsequent analyses.

3. Uncertainty quantification framework

An uncertainty quantification framework that combines the strengths of stepwise sensitivity analysis and adaptive surrogate-based optimization was designed to assess and reduce parametric uncertainties of CREST at low computational cost. Fig. 2 breaks down the four key steps of the uncertainty quantification framework. More detailed information regarding the qualitative sensitivity analysis and surrogate-based quantitative sensitivity analysis can be found in Gan et al. (2015). A detailed description of the adaptive surrogate-based optimization is presented in Wang et al. (2014).

3.1. *Problem specification*

The first step of the uncertainty quantification framework is to specify the uncertainty quantification problem, including the choice of model parameters and objective functions. Table 2 lists the twelve CREST parameters, whose ranges were determined according to physical meanings and previous literature (Xue et al., 2013, Xue et al. 2015). All parameters were assumed to follow a uniform distribution. Following McCuen et al. (2006), two objective functions – relative bias (RB) and Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970) – were used to evaluate the goodness-of-fit for streamflow simulation. The former metric is the model bias normalized with respect to the observed mean, while the latter is the root mean square error normalized with respect to the observed variance. The equations of both statistics are given in appendix A.

3.2. *Qualitative sensitivity analysis*

Qualitative sensitivity analysis methods reflect the relative importance of model parameters, and aim to screen out insensitive parameters using a small number of model evaluations. We adopted the Latin Hypercube-based One-At-a-Time (LH-OAT) method (van Griensven et al., 2006) for qualitative sensitivity analysis. It combines the robustness of Latin hypercube sampling, which ensures that the parameter space has been fully explored, with the precision of one-at-a-time sampling, assuring that changes in the response in each model run

Table 1
Characteristics of the selected watersheds and corresponding streamflow gauging stations.

River Basin	Watershed No.	Area (km^2)	Station Name	Lon (°E)	Lat (°N)	P (mm/d)	PET (mm/d)	SD (m^3/s)	Soil texture	Vegetation type	Hydrologic characteristic (N: Normal; W: Wet; D: Dry)
Songhua River	W1	151,275	Jiangqiao	123.68	46.78	0.71	2.08	446.9	44.2% loam; 30.9% silt loam; 13.0% clay loam	30.6% deciduous needleleaf forest; 29.7% meadow; 17.0% farmland; 16.9% deciduous broadleaf forest	2008 (D), 2009 (N), 2010 (D), 2011 (W), 2012 (N)
Liao River	W2	153,460	Tieling	123.83	42.33	0.62	2.83	84.2	35.7% loam; 25.6% loamy sand; 13.2% sandy loam; 8.7% silt loam	42.9% farmland; 17.2% meadow; 9.2% desert grassland; 8.6% deciduous broadleaf forest	2008 (N), 2009 (D), 2010 (W), 2011 (D), 2012 (N)
Hai River	W3	42,788	Luanxian	118.78	39.76	0.71	2.96	20.0	66.5% loam; 13.9% silt loam; 9.8% sandy loam; 7.0% loamy sand	36.4% deciduous broadleaf forest; 20.9% meadow; 20.3% farmland; 13.3% evergreen needleleaf forest	2008 (D), 2009 (D), 2010 (D), 2011 (D), 2012 (N)
Yellow River	W4	102,987	Huaxian	109.77	34.58	0.73	2.76	136.5	86.4% silt loam; 7.4% loam; 5.0% loamy sand	57.1% farmland; 21.4% slope grassland; 9.7% deciduous broadleaf forest; 5.2% bush	2008 (D), 2009 (D), 2010 (N), 2011 (W), 2012 (N)
Huai River	W5	120,240	Wujiadu	117.38	32.93	1.29	3.16	874.7	30.3% silty clay loam; 27.6% silt loam; 15.0% clay loam; 13.1% loam	84.0 farmland; 4.7% slope grassland; 4.3% bush	2008 (N), 2009 (D), 2010 (W), 2011 (D), 2012 (D)
Yangtze River	W6	984,794	Yichang	111.30	30.70	1.25	2.41	12601.4	35.5% loam; 29.3% silt loam; 11.4% silty clay loam; 10.6% clay loam; 10.4% sandy loam	29.4% alpine and sub-alpine meadow; 22.2% evergreen needleleaf forest; 20.1% farmland; 9.4% evergreen broadleaf forest	2008 (N), 2009 (N), 2010 (N), 2011 (N), 2012 (N)
Southeast River	W7	12,501	Yangkou	117.92	26.80	2.66	2.79	408.9	57.7% clay loam; 25.4% loam; 5.6% silt loam	53.5% evergreen broadleaf forest; 39.4% evergreen needleleaf forest	2008 (D), 2009 (D), 2010 (W), 2011 (D), 2012 (W)
Pearl River	W8	347,505	Gaoyao	112.47	23.05	1.95	3.11	6648.8	34.5% clay loam; 27.4% silty clay loam; 17.9% loam; 6.5% silt loam	28.7% bush; 26.5% evergreen needleleaf forest; 13.3% farmland; 12.8% evergreen broadleaf forest	2008 (W), 2009 (D), 2010 (N), 2011 (D), 2012 (N)
Southwest River	W9	26,216	Lhasa	91.15	29.63	0.79	1.94	287.9	51.6% silt loam; 32.1% sandy loam; 15.7% loam	82.4% alpine and sub-alpine meadow; 13.2% alpine and sub-alpine plain grass	2008 (W), 2009 (D), 2010 (D), 2011 (N), 2012 (N)
Inland River	W10	9189	Yingluoxia	100.18	38.81	0.63	1.67	67.1	58.5% silt loam; 21.5% sandy loam; 13.8% loam	63.1% meadow; 26.2% alpine and sub-alpine meadow; 7.7% bare rocks	2008 (W), 2009 (W), 2010 (N), 2011 (W), 2012 (W)

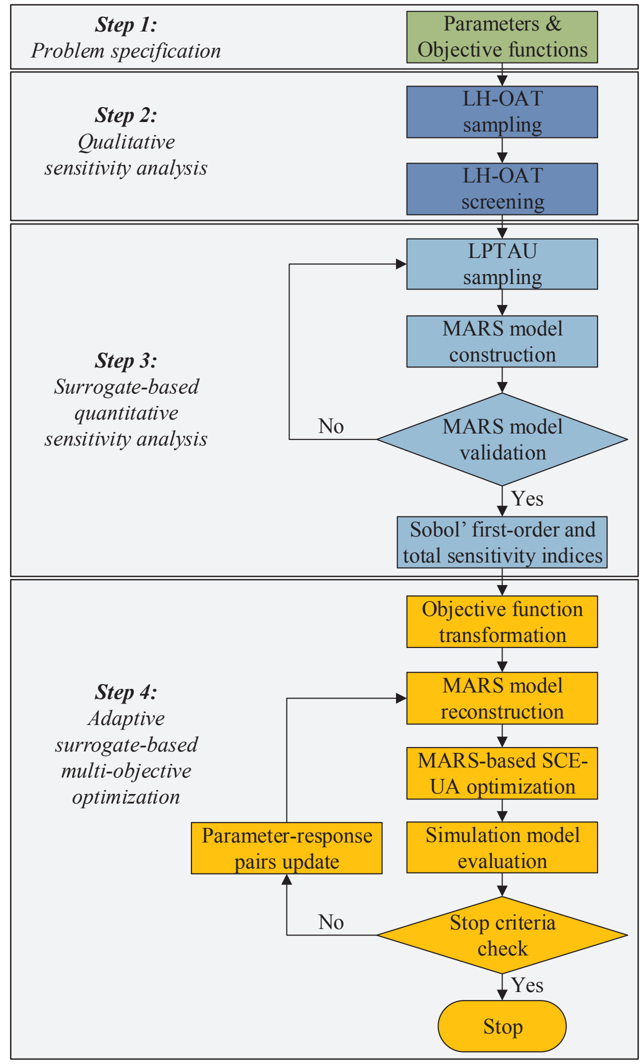

Fig. 2. Flowchart of the uncertainty quantification framework.

Table 2
Description of CREST model parameters and their feasible ranges.

No.	Parameter	Description	Range	Default
P1	RainFact	Multiplier on the precipitation field	0.5–1.2	1.00
P2	Ksat	Soil saturated hydraulic conductivity (mm/d)	1–1000	2.84
P3	WM	Maximum water-holding capacity (mm)	1–500	129.95
P4	B	Exponent of the variable infiltration curve	0.05–1.5	0.48
P5	IM	Impervious area ratio	0–0.2	0.07
P6	KE	Multiplier on the potential evapotranspiration	0.1–1.5	0.85
P7	coeM	Overland flow velocity coefficient	1–150	58.89
P8	expM	Overland flow velocity exponent	0.1–2	0.25
P9	coeR	Multiplier used to convert overland flow speed to channel flow speed	0.2–3	0.73
P10	coeS	Multiplier used to convert overland flow speed to interflow flow speed	0.001–1	0.63
P11	KS	Overland reservoir discharge parameter	0–1	0.41
P12	KI	Interflow reservoir discharge parameter	0–1	0.22

can be unambiguously attributed to the parameter that was changed. Overall and interaction effects of each parameter can be represented respectively by the mean and standard deviation of its elementary effects. The former is the overall influence of the parameter on the response, while the latter is the ensemble of the parameter's higher order effects (Campolongo et al., 2007).

3.3. Surrogate-based quantitative sensitivity analysis

Quantitative sensitivity analysis methods estimate the percentage of response variance each parameter is responsible for, due to the parameter's first-order effect and interaction effects with other parameters. We used the Sobol' (1993,2001) method, which performs ANOVA (Analysis of Variance)-like decomposition of model response variance and measures the specific order sensitivity index of a parameter by attributing all relevant variances. The total sensitivity index can be estimated by adding up all the sensitivity indices containing the parameter. Sobol's method is model independent and can be used for non-linear and non-monotonic models. Its main drawback is that calculating higher-order terms is prohibitive, especially when parameter dimensionality is large. We adopted the total sensitivity index proposed by Homma and Saltelli (1996) as a measure of the total effect of a given parameter at reasonable cost, by subtracting all sensitivity indices not related to this parameter from 1.

Even though the calculation of total sensitivity indices is practical, a large number of samples are still needed to get reliable results. We employed Multivariate Adaptive Regression Splines (MARS) (Friedman, 1991) to construct a surrogate model for computing Sobol' first-order (Sobol'-1) and total (Sobol'-t) sensitivity indices. Several other methods such as Gaussian process, radial basis functions, and support vector machines can also be considered. The selection of surrogate modeling methods depends on the purpose of the research, the applicability of the method, and the complexity of the simulation model. We chose MARS because we need a globally accurate surrogate model that can be used not only for sensitivity analysis but also for providing reliable information for the promising regions of the global optimum, and MARS is particularly competitive in accuracy and efficiency (Jin et al., 2001; Chen et al., 2013; Gan et al., 2014).

Quasi-random LP_τ (LPTAU) sequence sampling (Sobol', 1967) was adopted to generate deterministic and uniformly distributed samples, because it allows the addition of more points to the initial samples with the same uniformity characteristics (Tong, 2005). The idea of LPTAU is to generate uniformly distributed quasi-random sequence within the n-dimensional hypercube $I^n = [0,\ 1]^n$. Unlike random sequence, the position of sample points of the quasi-random sequence is deterministic. A finite sequence of points in I^n is called a P_τ net if it contains $N = 2^k$ points and every dyadic box of volume $2^{\tau-k}$ contains exactly 2^τ points of the sequence, where k and τ are integers and $k > \tau$ (Niederreiter, 1978). Other sampling methods such as Monte Carlo, Latin hypercube, and orthogonal array-based Latin hypercube can also be considered. A comparison of different sampling methods and surrogate modeling methods for surrogate-based sensitivity analysis can be found in Gan et al. (2014).

We employed the k-fold cross-validation strategy (Meckesheimer et al., 2002) to assess the fidelity of MARS. Namely, N parameter-response pairs $(X^j,\ Y^j)$ for $j = 1,\ 2,\ \cdots, N$ were split into k subsets of approximately equal size, and the MARS was fitted k times, each time leaving out one of the subsets from training and using the omitted subset to compute the coefficient of determination (R^2, see Appendix B). A tenfold cross-validation scheme as suggested by Zhang et al. (2009b) was employed to obtain the average R^2-values assessing the accuracy of MARS. Additional model runs were made to increase the accuracy of MARS when R^2 is less than a predefined value (e.g., 0.85), and the refitted MARS was assessed again using the enlarged parameter-response pairs. This trail-and-error process was repeated until we trained a reliable surrogate model.

3.4. Adaptive surrogate-based multi-objective optimization

The multi-objective optimization seeks a compromise solution from a set of Pareto solutions that represent the tradeoffs among different objectives. We employed the adaptive surrogate-based optimization framework presented in Wang et al. (2014) for multi-objective optimization of the most sensitive parameters, and used the LPTAU sampling, MARS surrogate modeling, and SCE-UA optimization methods for corresponding processes. The entire adaptive surrogate-based multi-objective optimization procedure is as follows.

(1) Objective function transformation

To make the two objective functions of streamflow discharge comparable in variability range and direction, we took the transformations $F_1(\mathbf{X}) = |RB|$ and $F_2(\mathbf{X}) = 1{-}NSE$, and expressed the multi-objective optimization problem as:

$$\min Y = \min F(\mathbf{X}) = \min(F_1(\mathbf{X}),\ F_2(\mathbf{X})) = \min(|RB|,\ 1{-}NSE) \tag{1}$$

The two objective functions were then transformed into a Euclidian distance function following Madsen (2000):

$$F_{agg}(\mathbf{X}) = \sqrt{\sum_{j=1}^{2} [F_j(\mathbf{X}) + \mathrm{A_j}]^2} \tag{2}$$

where $F_j(\mathbf{X})$ is the jth objective function, and A_j is the transformation constant to make $F_j(\mathbf{X}) + A_j$ have the same distance to the origin. A_j can be determined through LPTAU experiments inherited from surrogate-based quantitative sensitivity analysis by:

$$A_j = \max\{F_{1,\min},\ F_{2,\min}\}{-}F_{j,\min},\ j = 1,\ 2 \tag{3}$$

If $F_{1,\min} \geqslant F_{2,\min}$, $A_1 = 0$ and $A_2 = F_{1,\min}{-}F_{2,\min}$; or else, $A_1 = F_{2,\min}{-}F_{1,\min}$ and $A_2 = 0$.

(2) Surrogate model reconstruction

To improve the optimization search efficiency, MARS was reconstructed to approximate the relationship between the sensitive parameters and the aggregated objective function F_{agg} using the parameter-response pairs of the simulation model (i.e., CREST). We took the LPTAU experiments inherited from surrogate-based quantitative sensitivity analysis as the initial samples to construct a globally accurate surrogate model, which can provide reliable information for promising regions of the global optimum in the parameter space. It is unlikely, however, that the surrogate would be sufficiently accurate in the region of the optimum. A succession of infill samples at the predicted optimum were therefore added to enhance the accuracy of the surrogate in the promising region to find the optimum efficiently.

(3) Surrogate-based optimization

We employed the global optimization method Shuffled Complex Evolution (SCE-UA) (Duan et al., 1992) to search the minimum F_{agg} value and corresponding optimal parameter set of the surrogate model. For effectiveness and efficiency, the SCE-UA method must properly specify a few algorithmic parameters. In particular, the number of complexes (p) is primarily determined by the dimensionality of the calibration problem (i.e., the number of model parameters to be optimized), and a value equal to two or larger is sufficient for a six-dimension problem (Duan et al., 1994). We set p to four, since the dimensionality of our calibration problem would be smaller than twelve after the parameter reduction. Other algorithmic parameters and the stop criteria were assigned to the default values as suggested by Duan et al. (1994).

(4) Simulation model evaluation and stop criteria check

The simulation model (CREST) was run using the optimal parameter set of the surrogate model to find the corresponding F_{agg} value. The optimal parameter set and corresponding F_{agg} value were used to update the parameter-response pairs, and steps 2–4 were continued until the F_{agg} value of the simulation model converged to a constant.

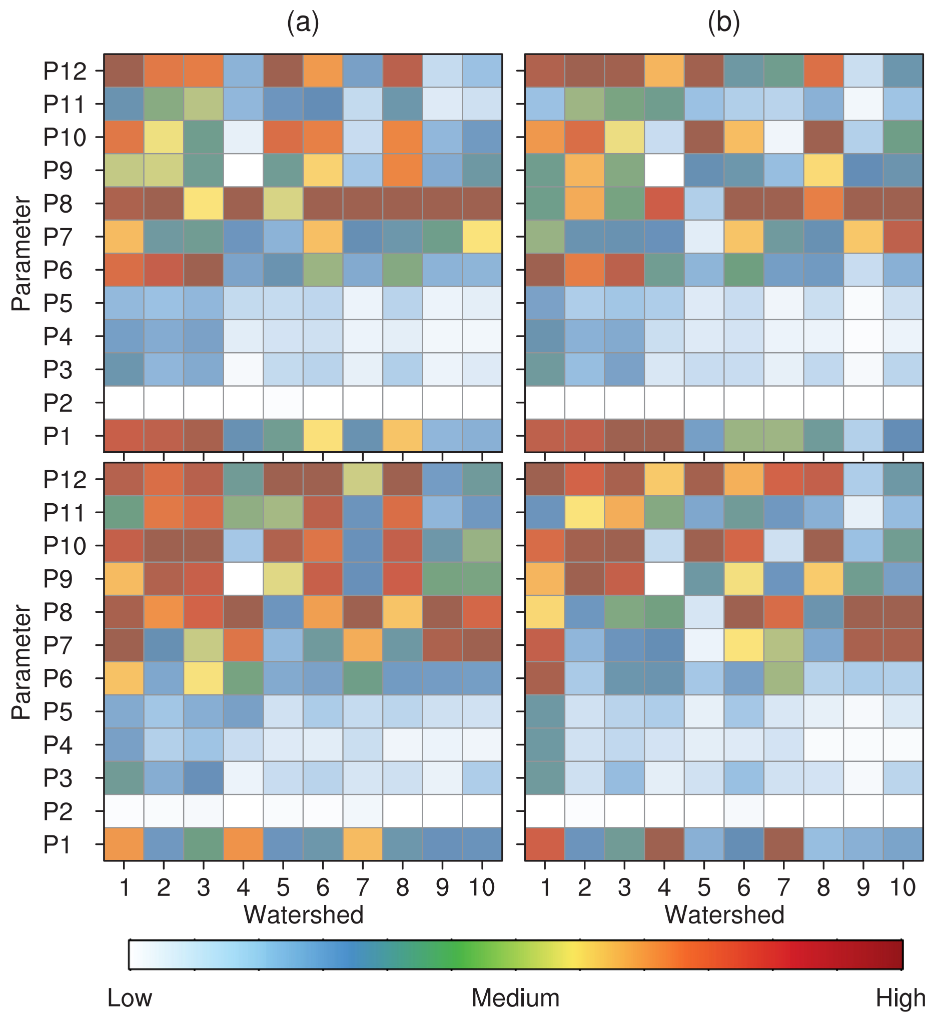

Fig. 3. Relative parameter sensitivities to overall effects (top row) and interaction effects (bottom row) of ten watersheds for (a) RB and (b) NSE of streamflow discharge.

4. Results and discussion

4.1. LH-OAT-based parameter screening

The range for each of the 12 CREST parameters was evenly divided into 20 levels. Thus, 260 ($=20 \times (12 + 1)$) LH-OAT samples were generated from the 12-dimensional parameter space. Fig. 3 shows the relative parameter sensitivities to overall effects and interaction effects of the ten watersheds for two objective functions (RB and NSE). Parameter rankings varied among watersheds as well as among objective functions even for the same sensitivity index. However, certain dominant patterns were evident. Overall, P2, P3, P4, and P5 can be regarded as insensitive parameters, while the other eight parameters can be classified as sensitive. All the insensitive parameters are physically based and related to specific subprocesses of runoff generation and routing, while the sensitive parameters are conceptually based and related to a collection of aggregated hydrological processes (Madsen, 2003; Wagener et al., 2003). The former can be determined from the physical characteristics of watersheds, while the latter cannot (Madsen, 2003).

The two objective functions find similar parameter sensitivities across watersheds, with a few exceptions. The most obvious difference is that in watershed W4, RB identifies P1 with a low overall effect (top row of Fig. 3a), while NSE identifies it with a high overall effect (top row of Fig. 3b). Another distinct difference is that in watershed W8, RB identifies P11 with a high interaction effect (bottom row of Fig. 3a), while NSE identifies it with a low interaction effect (bottom row of Fig. 3b). This variance of parameter sensitivities between objective functions was also found in several previous studies (Tang et al., 2007; van Werkhoven et al., 2008; Gan et al., 2015), since different objective functions measure different aspects of model behaviors.

The number of sensitive parameters per watershed varies from two (W9 and W10) to eight (W1) for RB, and from two (W9 and W10) to seven (W1) for NSE. These sensitive parameters may be explained by characteristics of the specific watersheds. For example, watersheds W9 and W10 have very similar hydroclimatic, soil, and vegetation conditions, and both are sensitive only to P7 and P8, which influence the overland flow velocity. In these sparsely vegetated small semi-arid watersheds, infiltration-excess overland flow contributes most to the total runoff. Variation of overland flow velocity leads to alterations in the timing of runoff delivery from slopes to streams (Holden et al., 2008). Meanwhile, humid watershed W7 is also sensitive to P7 and/or P8, because saturation-excess overland flow constitutes a major part of the total runoff. Watersheds W1, W2, and W3 have similar hydroclimatic conditions, and are sensitive to P1, P6, P8, P10, and P12. The first three parameters reflect the influence of vegetation type, resulting in different throughfall, evapotranspiration, and overland flow velocity, respectively. The latter two parameters reflect the influence of soil texture, and control the flow speed and discharge of interflow, respectively. For humid watersheds W1, W2, W3, W5, W6, and W8, P10 and P12 show strong overall effects, suggesting the importance of interflow, and P9 yields a significant sensitivity, indicating the

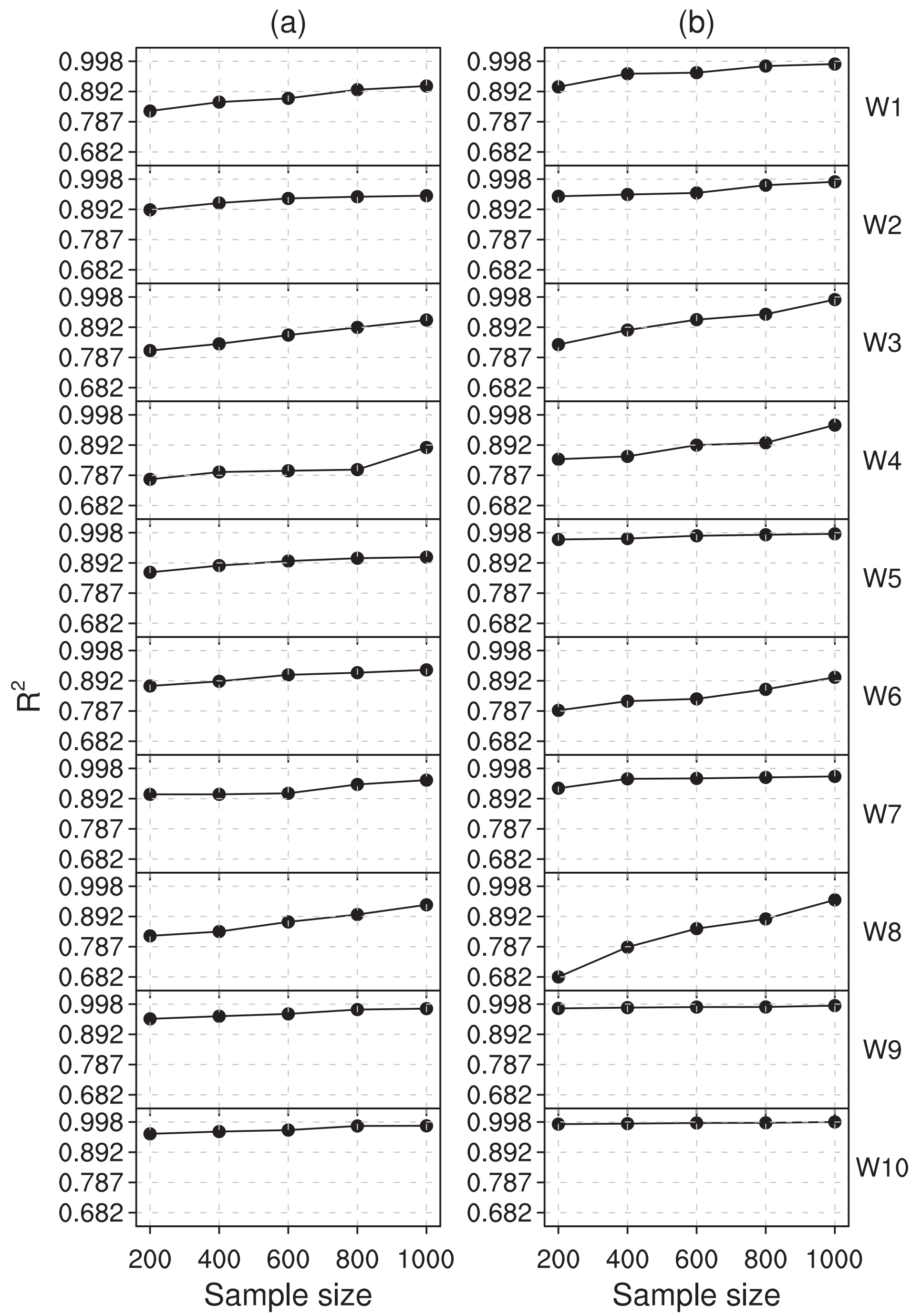

Fig. 4. The average R^2-values obtained by MARS with tenfold cross-validation scheme and different training sample sizes across ten watersheds for (a) RB and (b) NSE of streamflow discharge.

importance of channel flow. In some cases a parameter with a low overall effect (e.g., P7 in watershed W4 for RB) may have a significant interaction effect, and thus should also be regarded as a sensitive parameter (Gan et al., 2014).

4.2. Construction and validation of MARS model

Four parameters (P2, P3, P4, and P5) identified as insensitive by LH-OAT screening were set to their default values as shown in Table 2, and the other eight parameters that govern the hydrological response of CREST were selected for further analysis. We designed five sets of LPTAU samples, equally spaced between 200 and 1000 with an increment of 200, to investigate the effect of training sample size on the performance of MARS in approximating CREST. Fig. 4 shows that MARS's performance is substantially affected by the training sample size. It can well approximate CREST at a low sample size for most of the watersheds. For example, The R^2 values are greater than 0.80 for all watersheds except W4, W6, and W8 when the sample size equals to 200. The surrogate model trained with more samples tended to perform better, but had higher computational cost. Although the ranges of R^2-values varied across watersheds and objective functions, the overall performance of MARS with 1000 samples was sufficient in all cases, indicating that the surrogate model mimics the performance of CREST fairly well. The R^2-values with 1000 training samples range from 0.884 to 0.997, which are close to those reported in previous studies obtained by support vector machine and artificial neural network (Johnson and

Rogers, 2000; Zhang et al., 2009b).

We then used the MARS trained with 1000 samples as a pseudo simulator of CREST to ensure that it is globally accurate not only for sensitivity analysis but also for providing reliable information for the promising regions of the global optimum. For the purpose of efficient sensitivity analysis, MARS trained with a small sample size (e.g., 200) can even be used for most of the watersheds. Moreover, other efficient methods such as the Gaussian process-based probability sensitivity analysis (Oakley and O'Hagan, 2004; Daneshkhah and Bedford, 2013) could also be considered. However, the surrogate model trained with much less samples for surrogate-based sensitivity analysis may easily miss the global optimum when it is reused for surrogate-based optimization (Razavi et al., 2012). The efficiency of the surrogate model could be better improved by comparing different surrogate modeling methods and sampling methods against the research purpose (Gan et al., 2014; Wang et al., 2014).

4.3. MARS-based Sobol' sensitivity indices

The eight parameters (P1, P6, P7, P8, P9, P10, P11, and P12) that play important roles in explaining the uncertainty of streamflow simulation were tested for their sensitivities. Using the pseudo simulator, we designed 100,000 LPTAU samples to estimate Sobol' sensitivity indices for each objective function in each watershed. Fig. 5 presents Sobol' first-order and total sensitivity indices of ten watersheds for RB and NSE of streamflow discharge. The residual between Sobol' total and first-order sensitivity indices represents the interaction of a specific parameter with other parameters at all orders. Parameter sensitivities vary across objective functions and watersheds, and are mainly dominated by interactions. The significant parameter interactions indicate that CREST may be overparameterized, or its model structures are incorrect (Bastidas et al., 2006; Saltelli et al., 2008; Rosero et al., 2010). In addition, all watersheds exhibit higher parameter interaction for NSE than for RB, suggesting that the functional relationship is more complex for the former than for the latter.

It should be noted that the surrogate-based sensitivity indices, especially the interactions we obtained may differ from the true indices that would have been obtained had we been able to carry out the sensitivity analysis on the simulation model (O'Hagan, 2006). The Gaussian process-based probability sensitivity analysis method provides us a way of presenting variance-based sensitivity indices with uncertainty ranges at low computational cost (Oakley and O'Hagan, 2004; Daneshkhah and Bedford, 2013). This method is especially attractive when the simulation model is computationally expensive and its response surface is smooth. The uncertainty ranges of the sensitivity indices decrease as more samples are used to approximate corresponding sensitivity indices (Daneshkhah and Bedford, 2013). Although no uncertainty ranges are given here along with the MARS-based Sobol' sensitivity indices, the sensitivity categories based on the total sensitivity indices should be reliable since the accuracy of the surrogate model has validated to be acceptable (Shahsavani and Grimvall, 2011).

In general, all parameters are sensitive in at least one watershed, except that parameter P11 (overland reservoir discharge parameter) is insensitive in all watersheds for all objective functions. In most watersheds, the most sensitive parameters are P6 (multiplier on potential evapotranspiration), P8 (overland flow velocity exponent), and P12 (interflow reservoir discharge parameter) for RB, and P1 (multiplier on the precipitation field), P6, and P12 for NSE. P1 and P6 are the major driving parameters in runoff generation, and thus are expected to be highly sensitive. The high sensitivity of P12 indicates the important role of interflow to the total runoff in CREST. Meanwhile, the low P11 sensitivity and the high P8 sensitivity suggest that the velocity of overland flow influences the streamflow situation more than its amount.

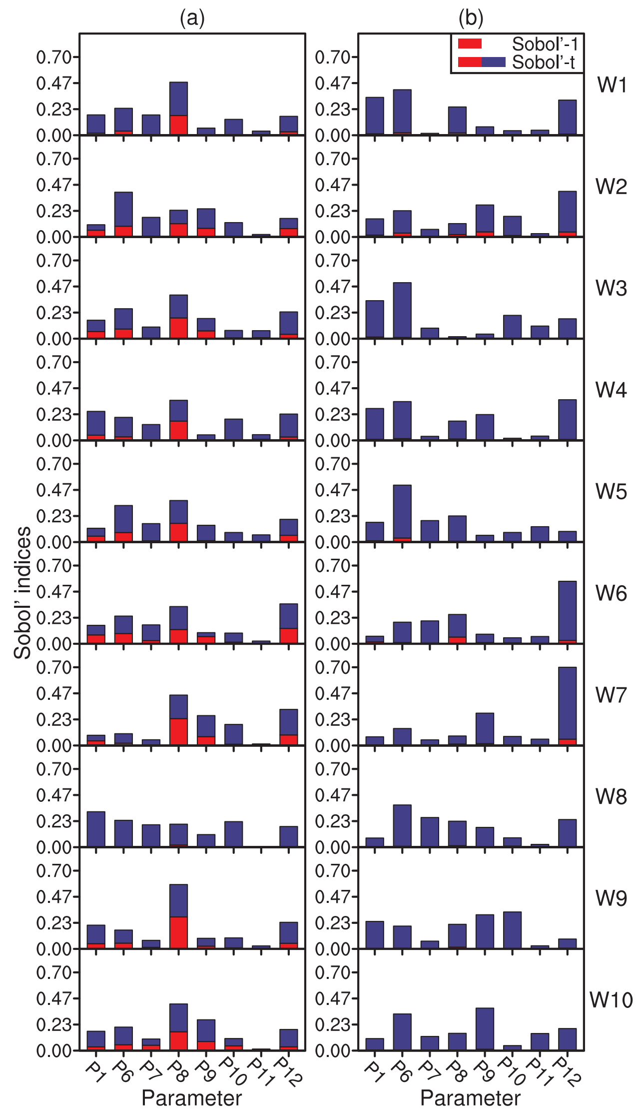

Fig. 5. Sobol' sensitivity indices of ten watersheds for (a) RB and (b) NSE of streamflow discharge.

4.4. Adaptive MARS-based multi-objective optimization

4.4.1. Feasibility of calibration

We systemically checked the feasibility of calibration to determine whether the actual observations could be captured by the ensemble ranges of the 1000 LPTAU experiments (Hou et al., 2012; Huang et al., 2013), which were inherited from the surrogate-based quantitative sensitivity analysis and would be used to construct surrogate models for adaptive surrogate-based optimization. Fig. 6 depicts boxplots of mean daily streamflow discharge for the ensemble experiments in ten watersheds. The boxplots present the streamflow simulation uncertainties by five-number summaries (i.e., the minimum, first quartile (25%), median (50%), third quartile (75%), and the maximum) of the ensemble experiments. Generally, the control experiment underestimates discharges, particularly for the summer months, whereas the ensemble ranges contain the observations. The result suggests that the parameter space is adequately represented by the samples of model parameters with reasonable physical bounds. It also indicates the necessity and possibility of calibrating parameters using observations.

4.4.2. Model calibration

Seven sensitive parameters (P1, P6, P7, P8, P9, P10, and P12) were

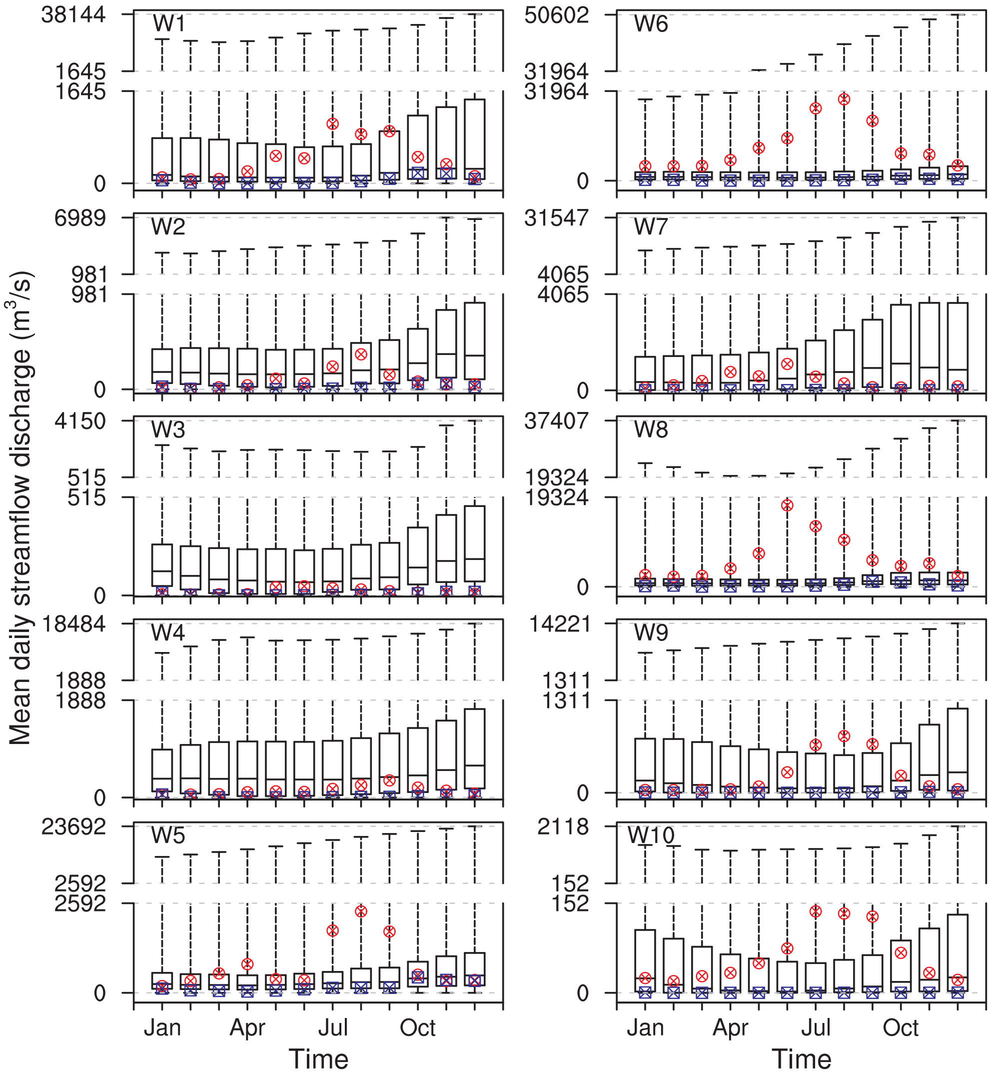

Fig. 6. Boxplots for mean daily streamflow discharge of the 1000 LPTAU ensemble experiments across ten watersheds. The red circles denote observations and the blue boxes denote the outputs of the control experiment using default parameter set.

selected as the calibration parameters, while other insensitive parameters were set to their default values. The 1000 LPTAU experiments, used in constructing MARS for quantitative sensitivity analysis, were taken to obtain the transformation constants A_1 and A_2 for $|RB|$ and $1-NSE$, respectively. The parameter-response pairs of these experiments were used to reconstruct MARS to approximate CREST for searching for the optimal solution. The newly found optimal parameter-response pair was appended to the parameter-response pool to adaptively update MARS. This adaptive surrogate-based optimization procedure was continued until the best compromise solution was found. Sample sizes needed to find such solutions for watersheds W1 to W10 were 240, 310, 340, 280, 265, 289, 210, 236, 198, and 180, respectively.

Fig. 7 presents the searching route in the objective function space across different watersheds, including the initial, intermediate, and final optimal solutions. A significant tradeoff between $|RB|$ and $1-NSE$ is observed in all watersheds except watershed W10, indicating that a parameter set that gives small percentage volume bias results in a poor hydrograph shape, and vice versa. This result is in line with the findings of Madsen (2003) and Bekele and Nicklow (2007), who have conducted automatic calibration of hydrological models involving multiple objectives. The significant tradeoff between different criteria may be due to unresolved errors in the model structure (Fenicia et al., 2007; Pokhrel et al., 2008) and/or measurement errors in the model data. The relatively nonobvious tradeoff between $|RB|$ and $1-NSE$ for watershed W10 means both percentage volume bias and hydrograph shape can be improved simultaneously. The absence of significant tradeoff indicates that the model structure is generally well conceptualized and the most relevant hydrological processes are taken into consideration for this watershed (Schoups et al., 2005). The disassociation of the final optimal solutions from clusters of the intermediate solutions for W1, W7, and W9 indicates that the response surfaces for these watersheds are pretty rough and the adaptive MARS-based SCE-UA successfully found the global optimal solutions but spent many experiments on the local optimum regions. The final optimal solutions, which compromise the Pareto outcomes, perform significantly better than those of the initial solutions, reducing $1-NSE$ by 5–50% and $|RB|$ by 20–95%.

Fig. 8 shows variations in the parameter sets, which have been normalized to the feasible bounds in Table 2 so that all range from 0 to 1. The parametric uncertainties have been significantly reduced by the adaptive surrogate-based multi-objective optimization. In particular, in watersheds W6 and W8, most parameters have variation ranges smaller than 30%, so they were very precisely determined. On the other hand, in watersheds W1 and W5, most parameters have larger variability ranges and indistinct trends, so parameter sets with high variability could give equally good simulations. Some parameters show obvious trends when moving along the searching route of the optimal solutions. For example, P1 values are small in watershed W3 but large in

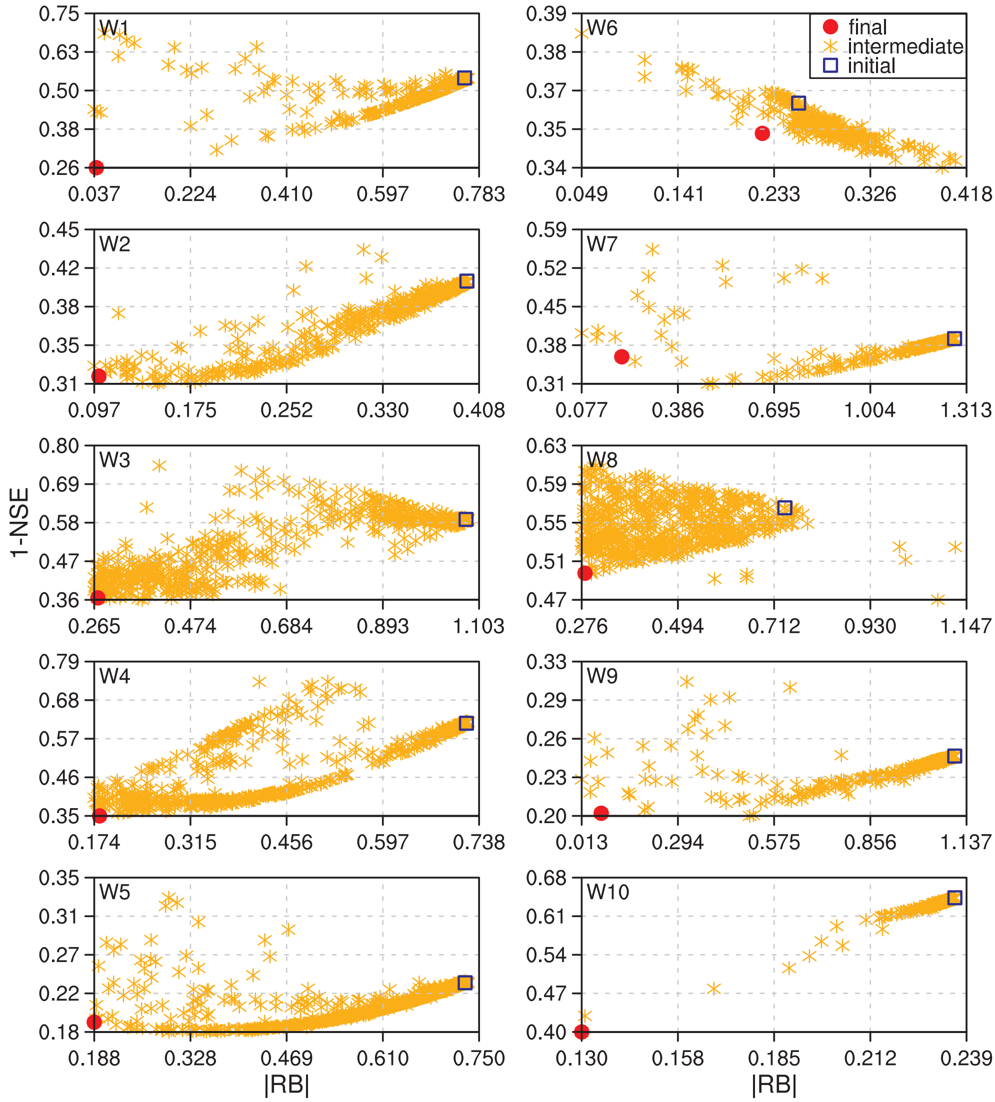

Fig. 7. Searching route in the objective function space showing streamflow |RB| versus $1 - \text{NSE}$ across ten watersheds. Blue squares represent initial optimal solutions, orange stars denote intermediate optimal solutions, and red dots indicate final optimal solutions of the adaptive surrogate-based multi-objective optimization.

watersheds W6 and W8, whereas P6 values are large in watershed W3 but small in watersheds W6 and W8. This is because increases in P1 result in increased precipitation and thus greater streamflow discharge, while increases in P6 result in increased evapotranspiration and thus smaller streamflow discharge. That is, larger P1 and smaller P6 lead to larger streamflow discharge, which is consistent with the observation that the peak flows are much higher in watersheds W6 and W8 than in watershed W3 during the calibration period.

Fig. 9 compares simulated and observed streamflow discharge at different watershed outlets. The uncertainty intervals show that CREST produces a range of simulated hydrographs corresponding to the "equally good" parameter sets along the Pareto front. The final optimal solutions not only closely reproduce the timing and magnitude of the peak flows, but also well represent the shape of the recession curves. The simulation results of the default solutions tend to underestimate streamflow by about 70–100% in all watersheds, while the optimal solutions show significant improvement, reducing $1 - \text{NSE}$ by 65–90% and |RB| by 60–95%. For watersheds W3, W6 and W8, where streamflow data contain many extended periods of low flow values or extreme high peaks, it is relatively more difficult to obtain a reliable calibration. The relatively poor ability for the model to simulate the recession curve (e.g., watersheds W1, W6, and W8) indicates that the groundwater module could be better improved. The tendency for the model to underestimate spring peaks (e.g. watersheds W1, W2, and W3) could be attributed to the lack of a snowmelt module in CREST. It should also be noted that human activities such as reservoir regulation and water diversion can affect the observed streamflow discharge, which inevitably lead to disagreement between simulated and observed streamflow.

4.4.3. Model validation

We validated the calibrated model for streamflow simulation outside the calibration period. Fig. 10 presents the validation results for streamflow simulations over the ten watersheds. The calibrated model simulation showed a large improvement over the default simulation, reducing $1 - \text{NSE}$ by 40–85% and |RB| by 35–90%, although some performance indices remain unsatisfactory. Since the parameters were calibrated to simulate intense flood events, the model performance decreases during the validation period when rainfall rates are lower, and vice versa (Moussa et al., 2007; Zhang et al., 2009a). For example, the calibration period for the watershed W3 is characterized by three dry years which causes low peak flows (less than 150 m^3/s), while the validation period is characterized by high rainfall rates during the year 2012 which causes high peak flows (more than 3000 m^3/s). In addition, the relatively short duration (three years) of the observational data used

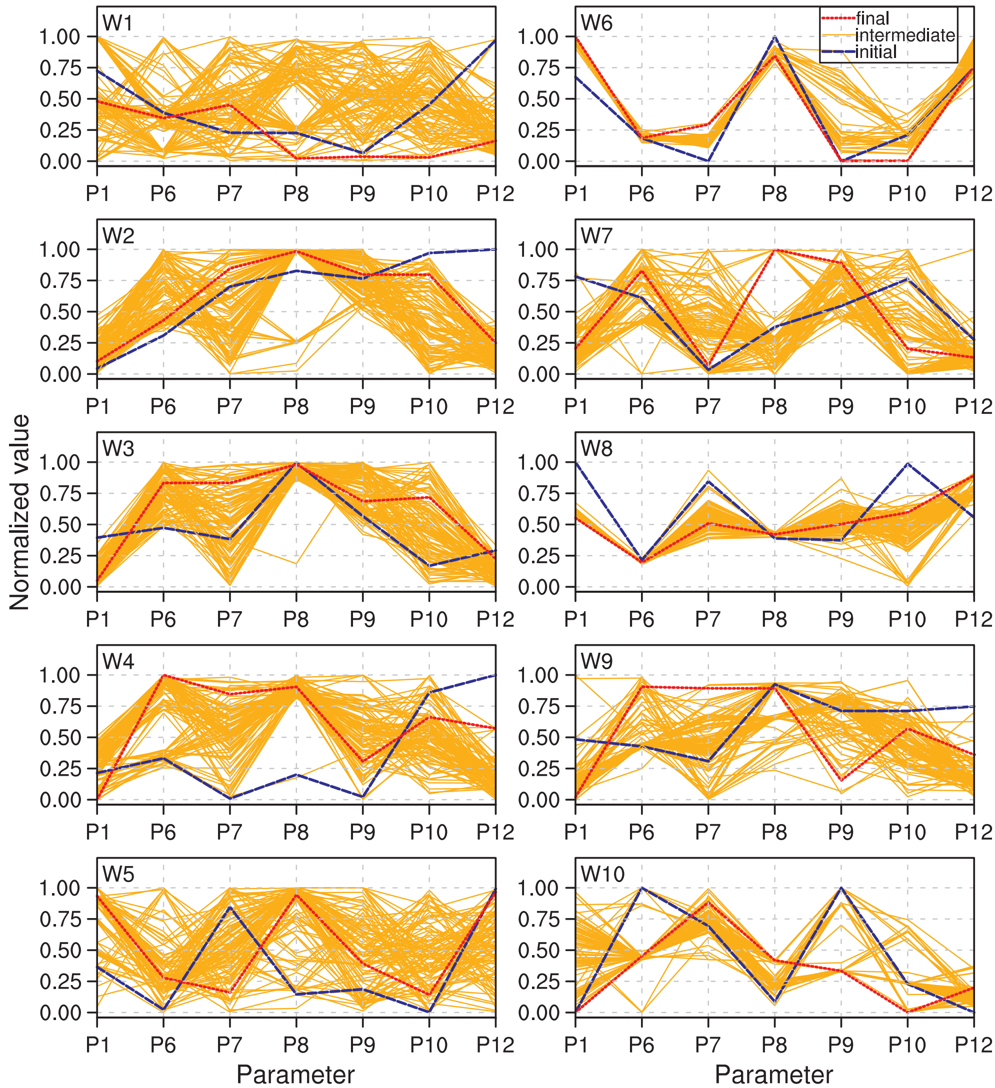

Fig. 8. Normalized parameter sets along the searching route across ten watersheds. Long blue dashed lines represent initial optimal solutions, orange solid lines denote intermediate optimal solutions, and short red dashed lines indicate final optimal solutions of the adaptive surrogate-based multi-objective optimization.

in calibration could also lead to unsatisfactory validation results (Sorooshian and Gupta, 1995; Muleta and Nicklow, 2005). Furthermore, the model limitations in representing the spatial heterogeneity of watershed may cause unstable model performance (Bekele and Nicklow, 2007).

5. Summary and conclusions

We presented an uncertainty quantification framework that combines the strengths of stepwise sensitivity analysis and adaptive surrogate-based multi-objective optimization to facilitate practical assessment and reduction of model parametric uncertainties. The framework was tested over ten watersheds using the distributed hydrological model CREST for daily streamflow simulation during the period 2008–2010. This generated optimal sets of sensitive parameters, which were used to improve CREST with validation against the daily streamflow simulation during the period 2011–2012.

Parameter sensitivities vary across watersheds and objective functions, but clearly demonstrate dominant patterns. Of the twelve parameters tested, we identified four that had little effect on streamflow simulation in any watershed. They were physically-based and determined by the physical mechanisms of runoff generation and routing processes and thus the physical characteristics of the watersheds. The remaining eight parameters were all conceptually-based and related to aggregated hydrological processes, and thus generally cannot be determined from the physical characteristics of the watersheds but need to be calibrated (Madsen, 2003). Of these, we identified that the contribution of the overland reservoir discharge parameter to the response variances is also negligible, indicating that more accurate quantitative evaluation of parameter sensitivities is needed after a qualitative parameter screening. Generally, the stepwise sensitivity analysis efficiently reduced the number of parameters needing calibration from twelve to seven, and thus constrained the dimensionality of calibration problem and enhanced the efficiency of parameter calibration.

The calibration exercise satisfactorily reproduced observed streamflow for all watersheds. The optimal solutions significantly improved streamflow simulation over the default, reducing $1-\text{NSE}$ by 65–90% and $|\text{RB}|$ by 60–95%. The validation exercise also indicated a large improvement of the optimal simulation over the default, reducing $1-\text{NSE}$ by 40–85% and $|\text{RB}|$ by 35–90%, although some performance indices were still not satisfactory as noted in other distributed hydrological models (Muleta and Nicklow, 2005; Zhang et al., 2009a; Sun et al., 2017). Even after calibration, there are potentially large uncertainties because no simulation model is capable of representing all physical processes, and observational data are likely incomplete. Additionally, the optimal parameter sets are realistic only for the same

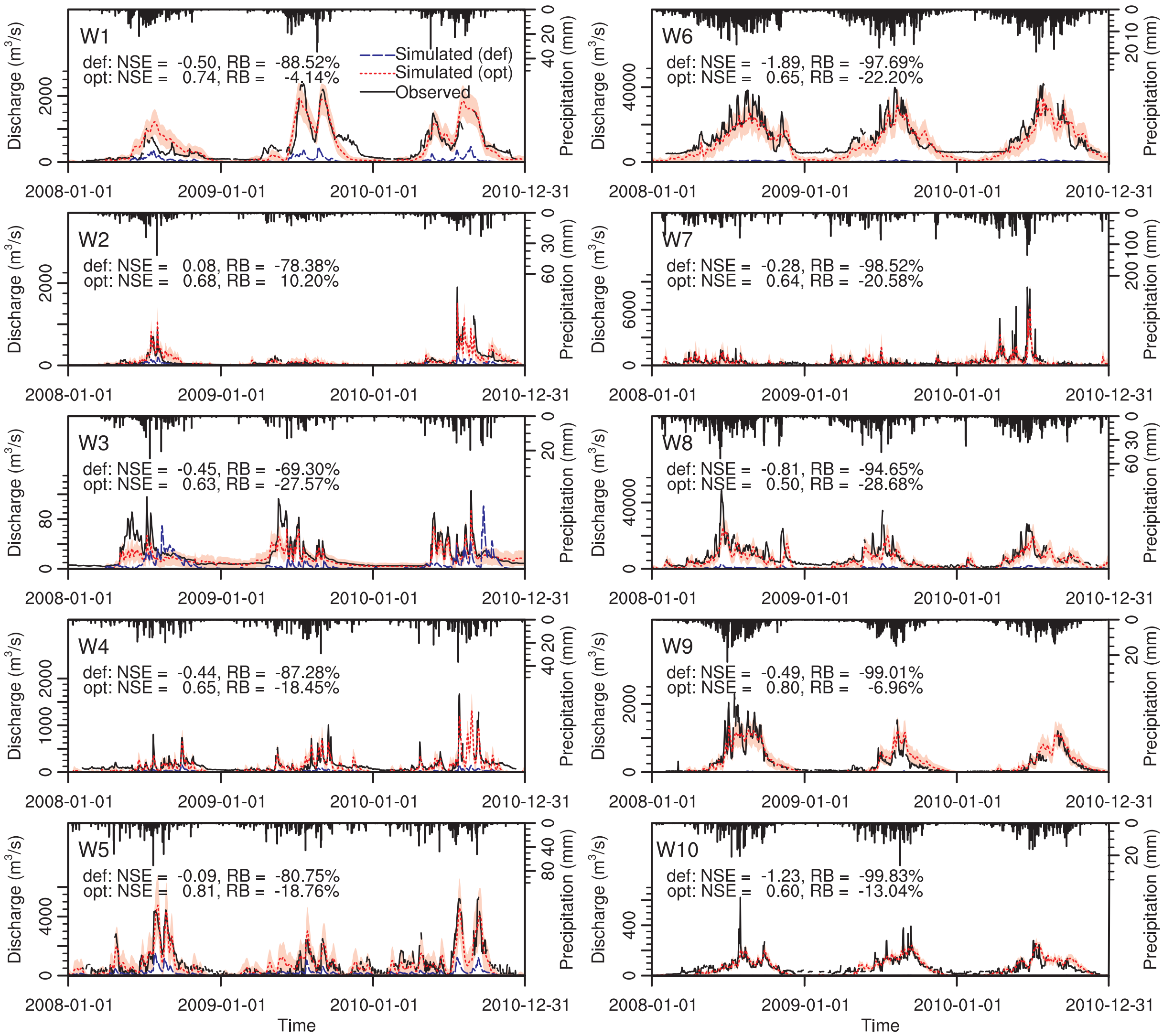

Fig. 9. Comparison of observed and simulated hydrographs for the calibration period across ten watersheds. For each subfigure, the long blue dashed curve corresponds to the hydrograph simulated by default solution (def), the short red dashed curve corresponds to the hydrograph simulated by final optimal solution (opt), the light red band corresponds to the uncertainty interval simulated by Pareto optimal solutions, and the black solid curve corresponds to the observed hydrograph.

watershed under similar meteorological conditions. A comprehensive evaluation is necessary to transfer the optimal parameter sets from one watershed to another with "similar" watershed properties and hydro-climatic conditions (Zhang et al., 2009a; Kumar et al., 2013; Gan et al., 2015), or even to the same watershed with greatly different meteorological conditions.

Overall, our uncertainty quantification framework is effective for multi-watershed multi-objective parametric uncertainty quantification, and provides useful information to understand the model behaviors and improve the model simulations. The framework designs various simulation experiments to reduce parameter dimensionality, and thus is well suited for calibrating high-dimensionality problems requiring systematic quantification of parametric uncertainties. The qualitative or even quantitative sensitivity analysis could be discarded for the calibration of low-dimensionality problems.

We considered only streamflow simulation due to the lack of observational data for other variables. In addition, the duration of the observed streamflow data is relatively short. The capability of this framework, as well as the behavior of CREST, could be more effectively tested with more observational data for multiple variables in multiple watersheds. Further improvements of the framework and its application to some of the more sophisticated models such as the Conjunctive Surface-Subsurface Process (CSSP) (Choi et al., 2013) and Noah multiparameterization (Noah-MP) (Niu et al., 2011) land surface models are still ongoing. The results of these works will be reported in due course.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (41505092), National Key Research and Development Program of China (2017YFC1404000), Open Research Fund of the State Key Laboratory of Severe Weather (2015LASW-A05), and National Natural Science Foundation of China (41605073). Liang was partially supported by the Beijige Open Research Fund of the Nanjing Meteorological Joint Research Center (NJCAR2016ZD03). We thank Jennifer A. Kennedy for her careful revision of the manuscript. We thank the editors and three anonymous reviewers for their insightful comments and suggestions that helped improve the manuscript. The forcing data and hydrography data can be accessed through the links given in the text. The observed streamflow discharge data are available upon request by contacting the first author.

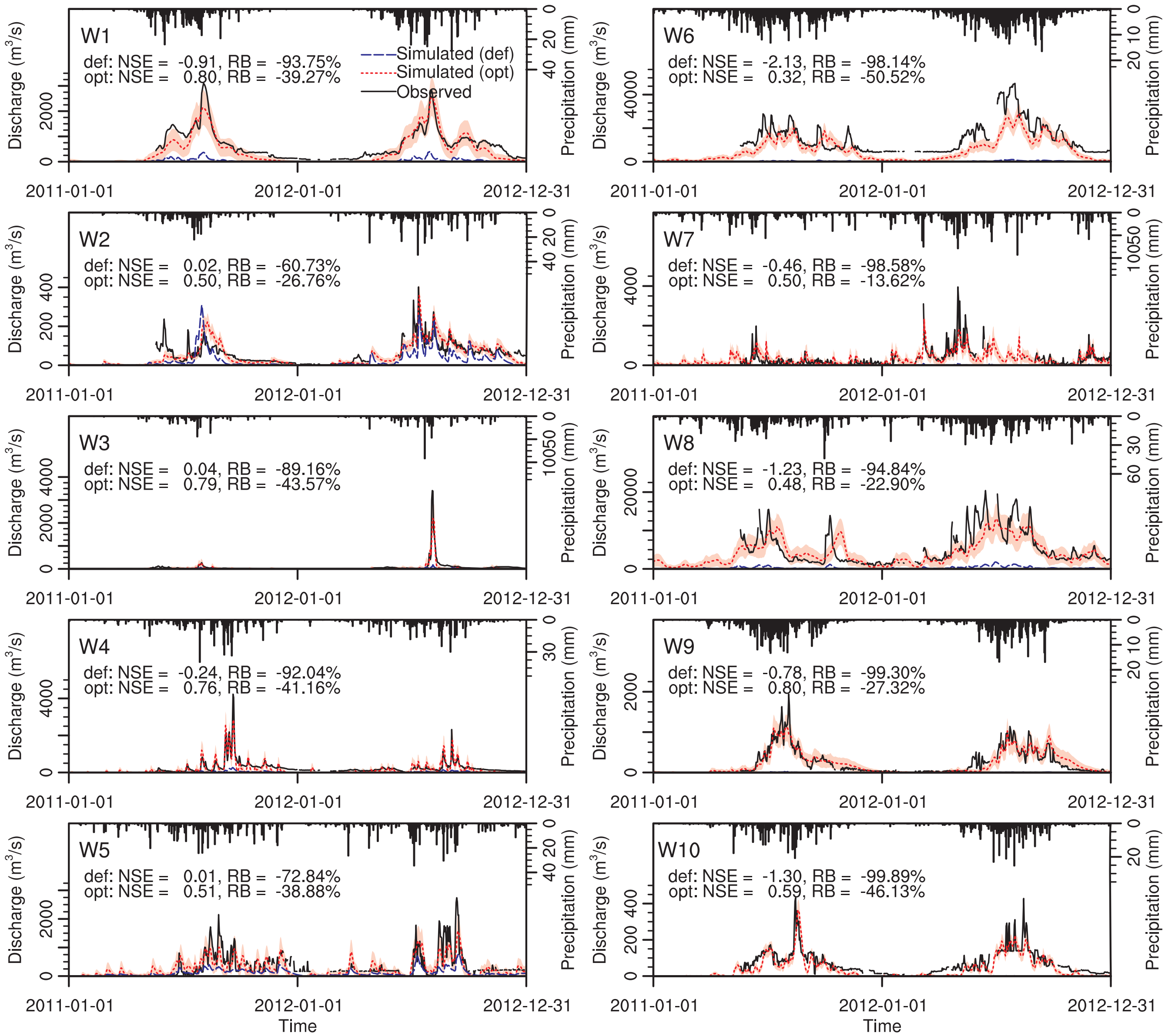

Fig. 10. Comparison of observed and simulated hydrographs for the validation period across ten watersheds. For each subfigure, the long blue dashed curve corresponds to the hydrograph simulated by default solution (def), the short red dashed curve corresponds to the hydrograph simulated by final optimal solution (opt), the light red band corresponds to the uncertainty interval simulated by Pareto optimal solutions, and the black solid curve corresponds to the observed hydrograph.

Appendix A. Objective functions

Relative bias (RB) and Nash-Sutcliffe efficiency (NSE) are defined as follows to measure the overall water balance and the overall shape of the hydrograph, respectively

$$RB = \frac{\frac{1}{N}\sum_{i=1}^{N}(S_i - O_i)}{\overline{O}} \tag{A1}$$

$$NSE = 1 - \frac{\sum_{i=1}^{N}(O_i - S_i)^2}{\sum_{i=1}^{N}(O_i - \overline{O})^2} \tag{A2}$$

where S_i and O_i are simulated and observed values at time i, respectively; $\overline{O}$ is the mean of observations; and N is the total number of observations (or simulations). RB ranges from minus infinity to plus infinity with lower absolute value indicating better agreement, while NSE ranges from minus infinity to 1.0 with higher value indicating better agreement.

Appendix B. *k*-fold cross-validation measure

The *k*-fold cross-validation measure used in this study is the coefficient of determination R^2, which is calculated as

$$R^2 = \frac{1}{k}\sum_{p=1}^{k}\left[\frac{\sum_{j=1}^{N/k}(Y_{p,j}-\overline{Y})(\widehat{Y}_{p,j}-\overline{\widehat{Y}})}{\sqrt{\sum_{j=1}^{N/k}(Y_{p,j}-\overline{Y})^2}\sqrt{\sum_{j=1}^{N/k}(\widehat{Y}_{p,j}-\overline{\widehat{Y}})^2}}\right]^2 \tag{B1}$$

where $\widehat{Y}_{p,j}$ and $Y_{p,j}$ are the predicted and actual response values of the *j*th sample point of the *p*th subset, respectively. $\overline{\widehat{Y}}$ and $\overline{Y}$ are mean values of $\widehat{Y}_{p,j}$ and $Y_{p,j}$, respectively.

References

Bastidas, L.A., Hogue, T.S., Sorooshian, S., Gupta, H.V., Shuttleworth, W.J., 2006. Parameter sensitivity analysis for different complexity land surface models using multicriteria methods. J. Geophys. Res. 111, D20101. https://doi.org/10.1029/2005JD006377.

Bekele, E.G., Nicklow, J.W., 2007. Multi-objective automatic calibration of SWAT using NSGA-II. J. Hydrol. 341 (3–4), 165–176. https://doi.org/10.1016/j.jhydrol.2007.05.014.

Beven, K., 2006. A manifesto for the equifinality thesis. J. Hydrol. 320 (1–2), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007.

Borgonovo, E., Castaings, W., Tarantola, S., 2012. Model emulation and moment-independent sensitivity analysis: an application to environmental modelling. Environ. Modell. Softw. 34, 105–115. https://doi.org/10.1016/j.envsoft.2011.06.006.

Campolongo, F., Cariboni, J., Saltelli, A., 2007. An effective screening design for sensitivity analysis of large models. Environ. Modell. Softw. 22 (10), 1509–1518. https://doi.org/10.1016/j.envsoft.2006.10.004.

Chen, M., Sun, Y., Fu, P., Carrigan, C.R., Lu, Z., Tong, C.H., Buscheck, T.A., 2013. Surrogate-based optimization of hydraulic fracturing in pre-existing fracture networks. Comput. Geosci.-UK 58, 69–79. https://doi.org/10.1016/j.cageo.2013.05.006.

Chen, M., Izady, A., Abdalla, O.A., 2017. An efficient surrogate-based simulation-optimization method for calibrating a regional MODFLOW model. J. Hydrol. 544, 591–603. https://doi.org/10.1016/j.jhydrol.2016.12.011.

Choi, H.I., Liang, X.-Z., Kumar, P., 2013. A conjunctive surface-subsurface flow representation for mesoscale land surface models. J. Hydrometeorol. 14, 1421–1441. https://doi.org/10.1175/JHM-D-12-0168.1.

Daneshkhah, A., Bedford, T., 2013. Probabilistic sensitivity analysis of system availability using Gaussian processes. Reliab. Eng. Syst. Safe. 112, 82–93. https://doi.org/10.1016/j.ress.2012.11.001.

Duan, Q., Sorooshian, S., Gupta, V., 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resour. Res. 28 (4), 1015–1031. https://doi.org/10.1029/91WR02985.

Duan, Q., Sorooshian, S., Gupta, V.K., 1994. Optimal use of the SCE-UA global optimization method for calibrating watershed models. J. Hydrol. 158 (3–4), 265–284. https://doi.org/10.1016/0022-1694(94)90057-4.

Duan, Q., et al., 2017. Automatic model calibration: a new way to improve numerical weather forecasting. B. Am. Meteorol. Soc. 98 (5), 959–970. https://doi.org/10.1175/BAMS-D-15-00104.1.

Efstratiadis, A., Koutsoyiannis, D., 2010. One decade of multi-objective calibration approaches in hydrological modelling: a review. Hydrolog. Sci. J. 55 (1), 58–78. https://doi.org/10.1080/02626660903526292.

Fatichi, S., et al., 2016. An overview of current applications, challenges, and future trends in distributed process-based models in hydrology. J. Hydrol. 537, 45–60. https://doi.org/10.1016/j.jhydrol.2016.03.026.

Fenicia, F., Savenije, H.H., Matgen, P., Pfister, L., 2007. A comparison of alternative multiobjective calibration strategies for hydrological modeling. Water Resour. Res. 43 (3), W03434. https://doi.org/10.1029/2006WR005098.

Foglia, L., Hill, M.C., Mehl, S.W., Burlando, P., 2009. Sensitivity analysis, calibration, and testing of a distributed hydrological model using error-based weighting and one objective function. Water Resour. Res. 45 (6), W06427. https://doi.org/10.1029/2008WR007255.

Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Stat. 19 (1), 1–67. https://doi.org/10.1214/aos/1176347963.

Gan, Y., Duan, Q., Gong, W., Tong, C., Sun, Y., Chu, W., Ye, A., Miao, C., Di, Z., 2014. A comprehensive evaluation of various sensitivity analysis methods: a case study with a hydrological model. Environ. Modell. Softw. 51, 269–285. https://doi.org/10.1016/j.envsoft.2013.09.031.

Gan, Y., Liang, X.-Z., Duan, Q., Choi, H.I., Dai, Y., Wu, H., 2015. Stepwise sensitivity analysis from qualitative to quantitative: application to the terrestrial hydrological modeling of a Conjunctive Surface-Subsurface Process (CSSP) land surface model. J. Adv. Model. Earth Sy. 7 (2), 648–669. https://doi.org/10.1002/2014MS000406.

Gong, W., Duan, Q., Li, J., Wang, C., Di, Z., Ye, A., Miao, C., Dai, Y., 2015. Multiobjective adaptive surrogate modeling-based optimization for parameter estimation of large, complex geophysical models. Water Resour. Res. https://doi.org/10.1002/2015WR018230.

Gourley, J.J., Vieux, B.E., 2006. A method for identifying sources of model uncertainty in rainfall-runoff simulations. J. Hydrol. 327 (1–2), 68–80. https://doi.org/10.1016/j.jhydrol.2005.11.036.

Gupta, H.V., Sorooshian, S., Yapo, P.O., 1998. Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information. Water Resour. Res. 34 (4), 751–763. https://doi.org/10.1029/97WR03495.

Gupta, H. V., K. J. Beven, and T. Wagener (2005), Model calibration and uncertainty estimation, in Encyclopedia of Hydrological Sciences, edited by M. G. Anderson and J. J. McDonnell, John Wiley & Sons, Ltd., doi:10.1002/0470848944.hsa138.

Holden, J., Kirkby, M.J., Lane, S.N., Milledge, D.G., Brookes, C.J., Holden, V., McDonald, A.T., 2008. Overland flow velocity and roughness properties in peatlands. Water Resour. Res. 44 (6), W06415. https://doi.org/10.1029/2007WR006052.

Homma, T., Saltelli, A., 1996. Importance measures in global sensitivity analysis of nonlinear models. Reliab. Eng. Syst. Safe. 52 (1), 1–17. https://doi.org/10.1016/0951-8320(96)00002-6.

Hou, Z., Huang, M., Leung, L.R., Lin, G., Ricciuto, D.M., 2012. Sensitivity of surface flux simulations to hydrologic parameters based on an uncertainty quantification framework applied to the Community Land Model. J. Geophys. Res. 117, D15108. https://doi.org/10.1029/2012JD017521.

Huang, M., Hou, Z., Leung, L.R., Ke, Y., Liu, Y., Fang, Z., Sun, Y., 2013. Uncertainty analysis of runoff simulations and parameter identifiability in the Community Land Model: evidence from MOPEX basins. J. Hydrometeorol. 14, 1754–1772. https://doi.org/10.1175/JHM-D-12-0138.1.

Jakeman, A.J., Hornberger, G.M., 1993. How much complexity is warranted in a rainfall-runoff model? Water Resour. Res. 29 (8), 2637–2649. https://doi.org/10.1029/93WR00877.

Jenson, S.K., Domingue, J.O., 1988. Extracting topographic structure from digital elevation data for geographic information system analysis. Photogramm. Eng. Rem. S. 54 (11), 1593–1600.

Jin, R., Chen, W., Simpson, T.W., 2001. Comparative studies of metamodelling techniques under multiple modelling criteria. Struct. Multidiscip. O. 23 (1), 1–13. https://doi.org/10.1007/s00158-001-0160-4.

Johnson, V.M., Rogers, L.L., 2000. Accuracy of neural network approximators in simulation-optimization. J. Water Res. Plan. Man. 126 (2), 48–56. https://doi.org/10.1061/(ASCE)0733-9496(2000) 126:2(48).

Kavetski, D., Clark, M.P., 2010. Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on model analysis and prediction. Water Resour. Res. 46, W10511. https://doi.org/10.1029/2009WR008896.

Khan, S.I., Adhikari, P., Hong, Y., Vergara, H., Adler, R.F., Policelli, F., Irwin, D., Korme, T., Okello, L., 2011. Hydroclimatology of Lake Victoria region using hydrologic model and satellite remote sensing data. Hydrol. Earth Syst. Sc. 15 (1), 107–117. https://doi.org/10.5194/hess-15-107-2011.

Khu, S.T., Madsen, H., 2005. Multiobjective calibration with Pareto preference ordering: an application to rainfall-runoff model calibration. Water Resour. Res. 41 (3), W03004. https://doi.org/10.1029/2004WR003041.

Kumar, R., Samaniego, L., Attinger, S., 2013. Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. Water Resour. Res. 49, 360–379. https://doi.org/10.1029/2012WR012195.

Lehner, B., Verdin, K., Jarvis, A., 2008. New global hydrography derived from spaceborne elevation data. EOS, Transactions, AGU 89 (10), 93–94. https://doi.org/10.1029/2008EO100001.

Madsen, H., 2000. Automatic calibration of a conceptual rainfall-runoff model using multiple objectives. J. Hydrol. 235 (3), 276–288. https://doi.org/10.1016/S0022-1694(00)00279-1.

Madsen, H., 2003. Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. Adv. Water Resour. 26 (2), 205–216. https://doi.org/10.1016/S0309-1708(02)00092-1.

Marie, G., Simioni, G., 2014. Extending the use of ecological models without sacrificing details: a generic and parsimonious meta-modelling approach. Methods Ecol. Evol. 5 (9), 934–943. https://doi.org/10.1111/2041-210X.12250.

Matonse, A.H., Kroll, C.N., 2013. Applying hillslope-storage models to improve low flow estimates with limited streamflow data at a watershed scale. J. Hydrol. 494, 20–31. https://doi.org/10.1016/j.jhydrol.2013.04.032.

McCuen, R.H., Knight, Z., Cutter, A.G., 2006. Evaluation of the Nash-Sutcliffe efficiency index. J. Hydrol. Eng. 11 (6), 597–602. https://doi.org/10.1061/(ASCE)1084-0699(2006) 11:6(597).

Meckesheimer, M., Booker, A.J., Barton, R.R., Simpson, T.W., 2002. Computationally inexpensive metamodel assessment strategies. AIAA J. 40 (10), 2053–2060. https://doi.org/10.2514/2.1538.

Moussa, R., Chahinian, N., Bocquillon, C., 2007. Distributed hydrological modelling of a Mediterranean mountainous catchment–Model construction and multi-site validation. J. Hydrol. 337 (1), 35–51. https://doi.org/10.1016/j.jhydrol.2007.01.028.

Moussa, R., Chahinian, N., 2009. Comparison of different multi-objective calibration criteria using a conceptual rainfall-runoff model of flood events. Hydrol. Earth Syst. Sc. 13 (4), 519–535. https://doi.org/10.5194/hess-13-519-2009.

Muleta, M.K., Nicklow, J.W., 2005. Sensitivity and uncertainty analysis coupled with automatic calibration for a distributed watershed model. J. Hydrol. 306 (1), 127–145. https://doi.org/10.1016/j.jhydrol.2004.09.005.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. J. Hydrol. 10 (3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.

Niederreiter, H., 1978. Quasi-Monte Carlo methods and pseudo-random numbers. B. Am. Math. Soc. 84 (6), 957–1041.

Niu, G.-Y., Yang, Z.-L., Mitchell, K.E., Chen, F., Ek, M.B., Barlage, M., Kumar, A., Manning, K., Niyogi, D., Rosero, E., 2011. The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. J. Geophys. Res. 116, D12109. https://doi.org/10.1029/2010JD015139.

Oakley, J.E., O'Hagan, A., 2004. Probabilistic sensitivity analysis of complex models: a Bayesian approach. J. R. Stat. Soc. B 66 (3), 751–769. https://doi.org/10.1111/j.1467-9868.2004.05304.x.

O'Hagan, A., 2006. Bayesian analysis of computer code outputs: a tutorial. Reliab. Eng. Syst. Safe. 91 (10), 1290–1300. https://doi.org/10.1016/j.ress.2005.11.025.

Pianosi, F., Beven, K., Freer, J., Hall, J.W., Rougier, J., Stephenson, D.B., Wagener, T., 2016. Sensitivity analysis of environmental models: a systematic review with practical workflow. Environ. Modell. Softw. 79, 214–232. https://doi.org/10.1016/j.envsoft.2016.02.008.

Pokhrel, P., Gupta, H.V., Wagener, T., 2008. A spatial regularization approach to parameter estimation for a distributed watershed model. Water Resour. Res. 44, W12419. https://doi.org/10.1029/2007WR006615.

Razavi, S., Tolson, B.A., Burn, D.H., 2012. Review of surrogate modeling in water resources. Water Resour. Res. 48, W07401. https://doi.org/10.1029/2011WR011527.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S.W., 2010. Understanding predictive uncertainty in hydrologic modeling: the challenge of identifying input and structural errors. Water Resour. Res. 46, W05521. https://doi.org/10.1029/2009WR008328.

Rosero, E., Yang, Z.-L., Wagener, T., Gulden, L.E., Yatheendradas, S., Niu, G.-Y., 2010. Quantifying parameter sensitivity, interaction, and transferability in hydrologically enhanced versions of the Noah land surface model over transition zones during the warm season. J. Geophys. Res. 115, D03106. https://doi.org/10.1029/2009JD012035.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., 2008. In: Global Sensitivity Analysis: The Primer. John Wiley & Sons, Ltd, Chichester, U.K., pp. 292.

Schoups, G., Addams, C.L., Gorelick, S.M., 2005. Multi-objective calibration of a surface water–groundwater flow model in an irrigated agricultural region: yaqui Valley, Sonora, Mexico. Hydrol. Earth Syst. Sc. 9, 549–568. https://doi.org/10.5194/hess-9-549-2005.

Shahsavani, D., Grimvall, A., 2011. Variance-based sensitivity analysis of model outputs using surrogate models. Environ. Modell. Softw. 26 (6), 723–730. https://doi.org/10.1016/j.envsoft.2011.01.002.

Shen, Y., Zhao, P., Pan, Y., Yu, J., 2014. A high spatiotemporal gauge-satellite merged precipitation analysis over China. J. Geophys. Res. 119 (6), 3063–3075. https://doi.org/10.1002/2013JD020686.

Shi, Y., Davis, K.J., Zhang, F., Duffy, C.J., Yu, X., 2014. Parameter estimation of a physically based land surface hydrologic model using the ensemble Kalman filter: a synthetic experiment. Water Resour. Res. 50, 706–724. https://doi.org/10.1002/2013WR014070.

Sikorska, A.E., Renard, B., 2017. Calibrating a hydrological model in stage space to account for rating curve uncertainties: general framework and key challenges. Adv. Water Resour. 105, 51–66. https://doi.org/10.1016/j.advwatres.2017.04.011.

Sobol', I.M., 1967. On the distribution of points in a cube and the approximate evaluation of integrals. USSR Comput. Math. Math. Phys. 7 (4), 86–112.

Sobol', I.M., 1993. Sensitivity analysis for nonlinear mathematical models. Math. Model. Comput. Exp. 1, 407–414.

Sobol', I.M., 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math. Comput. Simulat. 55 (1–3), 271–280. https://doi.org/10.1016/S0378-4754(00)00270-6.

Sorooshian, S., Gupta, V.K., 1995. Model calibration. In: Singh, V.P. (Ed.), Computer Models of Watershed Hydrology. Water Resources Publications, Highlands Ranch, Colorado, pp. 23–68.

Stedinger, J.R., Vogel, R.M., Lee, S.U., Batchelder, R., 2008. Appraisal of the generalized likelihood uncertainty estimation (GLUE) method. Water Resour. Res. 44, W00B06. https://doi.org/10.1029/2008WR006822.

Sun, W., Wang, Y., Wang, G., Cui, X., Yu, J., Zuo, D., Xu, Z., 2017. Physically based distributed hydrological model calibration based on a short period of streamflow data: case studies in four Chinese basins. Hydrol. Earth Syst. Sc. 21 (1), 251–265. https://doi.org/10.5194/hess-21-251-2017.

Tang, G., Zeng, Z., Long, D., Guo, X., Yong, B., Zhang, W., Hong, Y., 2016. Statistical and hydrological comparisons between TRMM and GPM level-3 products over a mid-latitude basin: is day-1 IMERG a good successor for TMPA 3B42V7? J. Hydrometeorol. 17 (1), 121–137. https://doi.org/10.1175/JHM-D-15-0059.1.

Tang, Y., Reed, P., Wagener, T., van Werkhoven, K., 2007. Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation. Hydrol. Earth Syst. Sc. 11 (2), 793–817. https://doi.org/10.5194/hess-11-793-2007.

Tong, C., 2005. In: PSUADE User's Manual. Lawrence Livermore National Laboratory (LLNL), Livermore, CA, pp. 109.

van Griensven, A., Meixner, T., Grunwald, S., Bishop, T., Diluzio, M., Srinivasan, R., 2006. A global sensitivity analysis tool for the parameters of multi-variable catchment models. J. Hydrol. 324 (1–4), 10–23. https://doi.org/10.1016/j.jhydrol.2005.09.008.

van Werkhoven, K., Wagener, T., Reed, P., Tang, Y., 2008. Characterization of watershed model behavior across a hydroclimatic gradient. Water Resour. Res. 44, W01429. https://doi.org/10.1029/2007WR006271.

van Werkhoven, K., Wagener, T., Reed, P., Tang, Y., 2009. Sensitivity-guided reduction of parametric dimensionality for multi-objective calibration of watershed models. Adv. Water Resour. 32 (8), 1154–1169. https://doi.org/10.1016/j.advwatres.2009.03.002.

Villa-Vialaneix, N., Follador, M., Ratto, M., Leip, A., 2012. A comparison of eight metamodeling techniques for the simulation of N_2O fluxes and N leaching from corn crops. Environ. Modell. Softw. 34, 51–66. https://doi.org/10.1016/j.envsoft.2011.05.003.

Vrugt, J.A., Gupta, H.V., Bastidas, L.A., Bouten, W., Sorooshian, S., 2003. Effective and efficient algorithm for multiobjective optimization of hydrologic models. Water Resour. Res. 39 (8), 1214. https://doi.org/10.1029/2002WR001746.

Wagener, T., McIntyre, N., Lees, M.J., Wheater, H.S., Gupta, H.V., 2003. Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability analysis. Hydrol. Process. 17 (2), 455–476. https://doi.org/10.1002/hyp.1135.

Wang, C., Duan, Q., Gong, W., Ye, A., Di, Z., Miao, C., 2014. An evaluation of adaptive surrogate modeling based optimization with two benchmark problems. Environ. Modell. Softw. 60, 167–179. https://doi.org/10.1016/j.envsoft.2014.05.026.

Wang, J., et al., 2011. The coupled routing and excess storage (CREST) distributed hydrological model. Hydrolog. Sci. J. 56 (1), 84–98. https://doi.org/10.1080/02626667.2010.543087.

Wu, H., Adler, R.F., Hong, Y., Tian, Y., Policelli, F., 2012. Evaluation of global flood detection using satellite-based rainfall and a hydrologic model. J. Hydrometeorol. 13 (4), 1268–1284. https://doi.org/10.1175/JHM-D-11-087.1.

Xue, X., Hong, Y., Limaye, A.S., Gourley, J.J., Huffman, G.J., Khan, S.I., Dorji, C., Chen, S., 2013. Statistical and hydrological evaluation of TRMM-based Multi-satellite Precipitation Analysis over the Wangchu Basin of Bhutan: are the latest satellite precipitation products 3B42V7 ready for use in ungauged basins? J. Hydrol. 499, 91–99. https://doi.org/10.1016/j.jhydrol.2013.06.042.

Xue, X., Hong, Y., Zhang, K., 2015. In: CREST – Coupled Routing and Excess STorage user manual, version 2.1. University of Oklahoma, Norman, OK, pp. 56.

Zhang, X., Srinivasan, R., Bosch, D., 2009a. Calibration and uncertainty analysis of the SWAT model using Genetic Algorithms and Bayesian Model Averaging. J. Hydrol. 374 (3), 307–317. https://doi.org/10.1016/j.jhydrol.2009.06.023.

Zhang, X., Srinivasan, R., Van Liew, M., 2009b. Approximating SWAT model using artificial neural network and support vector machine. J. Am. Water Resour. As. 45 (2), 460–474. https://doi.org/10.1111/j.1752-1688.2009.00302.x.