Improving the integrated hydrological simulation on a data-scarce catchment with multi-objective calibration

Qianwen He and Frank Molkenthin

ABSTRACT

The process-based hydrological model Soil and Water Assessment Tool ensures the simulation's reliability by calibration. Compared to the commonly applied single-objective calibration, multi-objective calibration benefits the spatial parameterization and the simulation of specific processes. However, the requirements of additional observations and the practical procedure are among the reasons to prevent the wider application of the multi-objective calibration. This study proposes to consider three groups of objectives for the calibration: multisite, multi-objective function, and multi-metric. For the study catchment with limited observations like the Yuan River Catchment (YRC) in China, the three groups corresponded to discharge from three hydrometric stations, both Nash–Sutcliffe efficiency (NSE) and inversed NSE for discharge evaluation, and MODIS global terrestrial evapotranspiration product and baseflow filtered from discharge as metrics, respectively. The applicability of two multi-objective calibration approaches, the Euclidean distance and nondominated sorting genetic algorithm II, was analyzed to calibrate the above-mentioned objectives for the YRC. Results show that multi-objective calibration has simultaneously ensured the model's better performance in terms of the spatial parameterization, the magnitude of the output time series, and the water balance components, and it also reduces the parameter and prediction uncertainty. The study thus leads to a generalized, recommended procedure for catchments with data scarcity to perform the multi-objective calibration.

Key words | multi-objective calibration, NSGA-II, parameter uncertainty, SWAT

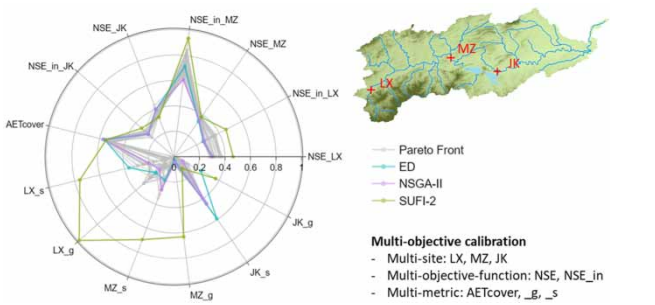
Qianwen He (corresponding author)
Frank Molkenthin
Chair of Hydrology,
Brandenburg University of Technology Cottbus-Senftenberg,
Platz der Deutschen Einheit 1, 03013 Cottbus,
Germany
E-mail: *qianwen.he@b-tu.de*

HIGHLIGHTS

- The application of the multi-objective calibration approach on a data-scarce catchment.
- Multiple objectives for calibration extracted for a data-scarce catchment from measured discharge and an open-access satellite-based dataset.
- The analysis of the applicability of the evolutionary algorithm, nondominated sorting genetic algorithm II, and the aggregation approach, the Euclidean distance.

doi: 10.2166/hydro.2021.132

GRAPHICAL ABSTRACT



INTRODUCTION

Integrated process-based hydrological models play an increasingly important role in supporting catchment management (van Griensven *et al.* 2006). The Soil and Water Assessment Tool (SWAT) (Neitsch *et al.* 2011) is one of the most commonly applied physically based hydrological models, which integrates various processes including hydrology, nutrients, erosion, etc. The complexity of the model system requires a sound calibration of the parametric assumption to derive the simulation with the least residue to the observations, which is essential to evaluate the model's reliability. However, the difficulty of calibrating the SWAT model lies in the validation of the spatially varied catchment characteristics and the interacted complex processes.

The classic calibration approach applied a regression-based summary statistic (Gupta *et al.* 2008) as the objective function, e.g. Nash–Sutcliffe efficiency (NSE), to optimize the goodness of fit of the output variable, e.g. the discharge at the catchment outlet. This approach is also referred to as single-objective calibration if only one output variable is evaluated and it is fundamental to perform the calibration procedure. However, the limitations of the single-objective approach are also obvious. It neglects the trade-off among interactive processes and responses at the interior location of the watershed (Yen *et al.* 2014), and the neglect of the catchment heterogeneity will lead to the parameters selected inconsistent with their physical meanings (Zhang *et al.* 2008). Also, applying only one objective function tends to derive a biased assessment, e.g. NSE is more sensitive to the peak values (Krause *et al.* 2005) and leads to a less reliable simulation in the low-flow period.

Multi-objective calibration was later proposed as a diagnostic approach (Gupta *et al.* 1998) that applied additional data, e.g. time-series or qualitative data, for validation (Yilmaz *et al.* 2008). The implementation of the additional observations in the calibration, e.g. mean annual water balances (Pfannerstill *et al.* 2017) and satellite-based evapotranspiration (ET) (Herman *et al.* 2018), improved the accuracy of the simulated water balance components. Multi-objective calibration was also applied to optimize the fitness of the magnitude of different flow periods (Pfannerstill *et al.* 2014) or the timing of the hydrograph (Zhang *et al.* 2016). Besides, the commonly applied multisite calibration helps to ensure reliable performance at the gauged locations (Zhang *et al.* 2008; Shrestha *et al.* 2016; Leta *et al.* 2017). Moreover, studies used extra objective functions to also tackle the low values in the time series, e.g. NSE with the inversed form to increase the weight of low flow (Pushpalatha *et al.* 2012).

Multi-objective calibration optimizes the parameters to meet the multiple criteria that control varied processes. Due to the existing conflicts between objectives in hydrological modeling, it is often not feasible to obtain the best performance of all the objectives simultaneously. The classic aggregation approach converts the objectives into a scalar function, e.g. the weighted average of all objectives (Zhang *et al.* 2016). The multi-objective evolutionary algorithms (MOEAs) (Schaffer 1984) combine the evolutionary

algorithms with the optimization of multiple objectives, e.g. the nondominated sorting genetic algorithm II (NSGA-II) (Deb *et al.* 2002).

The multi-objective calibration has not yet been widely adopted by the integrated hydrological models, such as the SWAT. The reasons include the requirements of additional observations, which are critical to catchments with limited measurements, and the lack of a practical procedure for implementation. Therefore, we proposed to overcome the data scarcity by utilizing satellite-based datasets and metrics extracted from the existing observations. The MODIS Global Terrestrial ET product (MOD16 ET) from NASA (Running *et al.* 2018) offers the global dataset of hydrological components and could be an option as an external metric (van Griensven *et al.* 2012; Abiodun *et al.* 2018). Metrics could also be extracted from the existing hydrography, e.g. the baseflow index filtered from hydrograph (Arnold *et al.* 1995; Ladson *et al.* 2013). Therefore, it is proposed in this study first to classify the objectives used in the multi-objective calibration procedure into three groups according to their potential effect: (1) multi-site, including the internal sites of observed discharge or water quality; (2) multi-objective function, evaluating both high and low values of the multisite objectives; and (3) multi-metric, including metrics extracted from hydrograph and external datasets, assessed by proper evaluation statistics.

To analyze the applicability of the multi-objective calibration on a catchment that only measured discharges were available for calibration, we conducted this study by utilizing the data from the MOD16 ET dataset and the existing measured discharge as the additional objectives. We implemented one representative aggregation approach, the Euclidean distance (ED), and one representative MOEA, the NSGA-II. ED is to aggregate the objectives (Gupta *et al.* 2009; Pfannerstill *et al.* 2017), and NSGA-II is a fast and efficient population-based optimization technique (Ercan & Goodall 2016). The research was designed to achieve the following aims: (1) to prove the applicability of the three objective groups with ED and NSGA-II, and (2) to demonstrate the advantages of ED and NSGA-II in comparison to the single-objective calibration.

METHODOLOGY

The study area and the input data

The Yuan River Catchment (YRC), displayed in Figure 1, is located in the middle west of Jiangxi province, China (113°50′–115°32′E; 27°22′–28°08′N). Its elevation varies from approximately 1,900 m in the southwest mountainous region to approximately 20 m in the northeast plain region. The catchment covers an area of 6,223.69 km^2 and is dominated by the subtropical monsoon climate. The long-term annual precipitation is approximately 1,668 mm, and the rain season, from April to June, contributes to 43.6% of the total precipitation. The average temperature ranges from 2.83 °C in January to 34.2 °C in July. The pan evaporation rate is approximately 1,228 mm (Fang 2011). Red and paddy soils dominate 50.3 and 30.7% of the YRC, respectively. Measured discharge of two hydrometric stations (LX and JK) was available. The discharges of MZ in the calibration period were extrapolated from the regression relation with LX in another period. The long-term average discharge was 11.5, 100.1, and 115.1 m^3/s in LX, MZ, and JK, respectively. The three stations were located from the upstream to the mid-downstream of the catchment; therefore, the YRC can only be calibrated by discharge at the upstream of the JK station. The calibration period was chosen from 2008 to 2010, and the validation period was from 2011 to 2014. The input data for the SWAT model are listed below in Table 1.

The SWAT model

The SWAT is an integrated process-based hydrological model to predict the long-term impact of land management practices on the water, sediment, and agricultural chemical yields on catchments (Neitsch *et al.* 2011). The computational unit of the SWAT is the hydrologic response unit (HRU) that consists of a certain land-use type, soil type, and slope in a sub-watershed. The hydrological cycle is simulated based on the water balance equation. The Soil Conservation Service (SCS) curve number method quantifies the daily surface runoff. Each soil layer is assumed to be

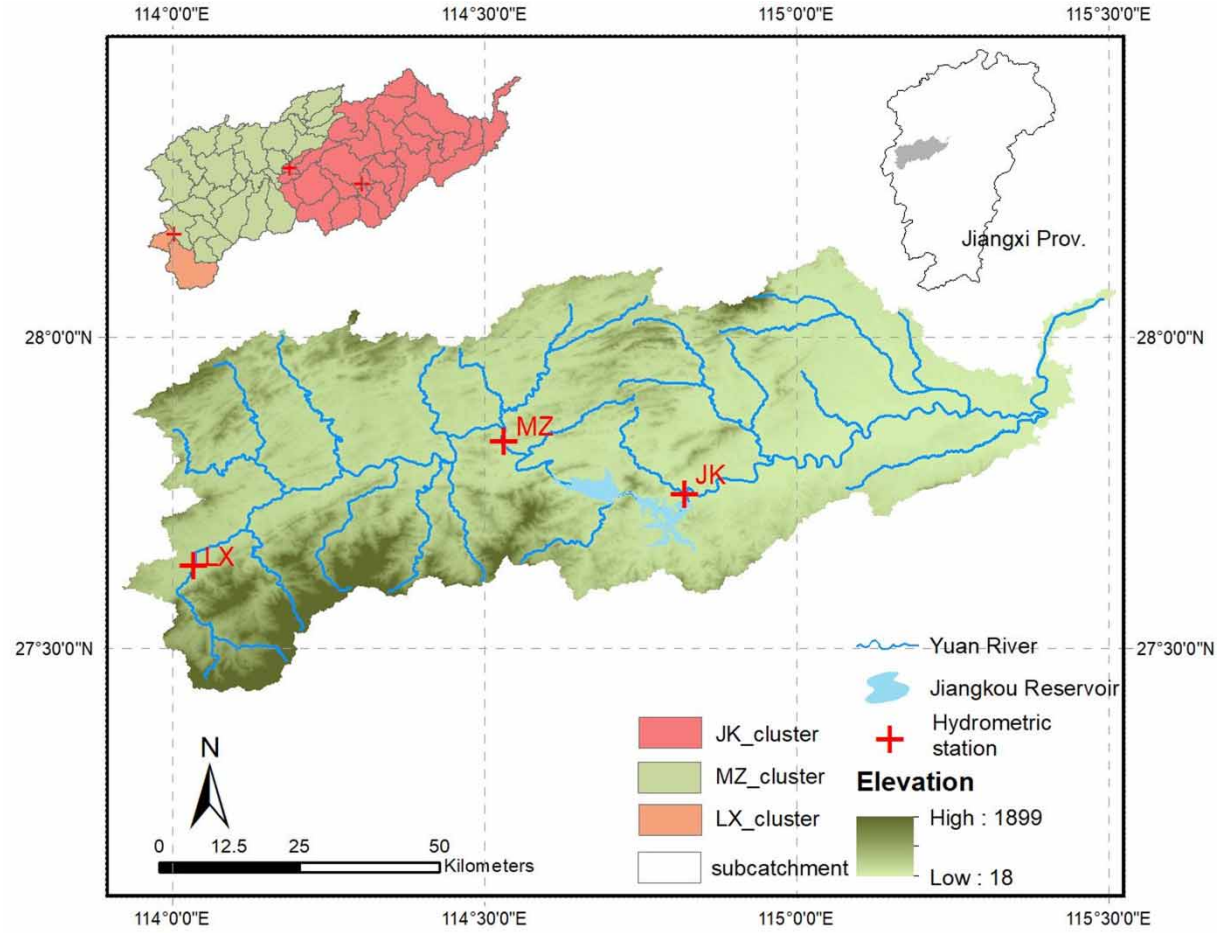


Figure 1 | The YRC with DEM, hydrometric stations, and subcatchment clusters.

Table 1 | SWAT model input data of the YRC

Data	Resolution or scale	Sources
DEM	30 m	Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Science
Yuan River network	1:200,000	
Land use in 2010	30 m	
Soil type and properties	1,000 m	Institute of Soil Science, Chinese Academy of Science
Precipitation[a] (2008–2014)	12 stations, daily	JK Hydropower Station
Climate data[b] (2008–2014)	1/3°, daily	China Meteorological Assimilation Driving Datasets for the SWAT Model, Chinese Academy of Science
Reservoir discharge (2008–2014)	Daily	JK Hydropower Station
Industry and domestic water use (2008–2014)	Annual average	Statistic Year Book of Jiangxi Prov., 2008–2014
Discharge (2008–2010 in LX; 2008–2014 in JK)	Two stations, daily	Yichun Hydrologic Bureau, JK reservoir power station

[a]It only included the region at the upstream of the JK hydrometric station.
[b]Precipitation data at the downstream were from the China Meteorological Assimilation Driving Dataset.

homogeneous horizontally. The Green–Ampt infiltration method and the storage routing method calculate the infiltration and percolation rates. The variable storage method was applied to compute the discharge in the stream. The groundwater system is simplified by the SWAT model to be a shallow aquifer and a deep aquifer.

The objectives for multi-objective calibration

Multisite

At the sequential order from upstream to downstream, the subcatchments were grouped into three subcatchment clusters based on the hydrometric stations (see Figure 1). Therefore, every parameter had a different value at each cluster, e.g. parameter CN2 (Table 4) was assigned to LX, MZ, and JK clusters, as CN2_LX, CN2_MZ, and CN2_JK, respectively. Subcatchments at the downstream of the JK station were ungauged and shared similarities of the identical land-use types (e.g. paddy field and forest) and soil types (e.g. red soil and paddy soil) with the subcatchments between MZ and JK; thus, the calibrated parameters in the JK cluster at the upstream of the JK station were assumed to be applicable at the ungauged subcatchments.

Multi-objective function

Two objective functions, NSE and inversed NSE (NSE_in) (Pushpalatha *et al.* 2012), were applied to evaluate the performance of the discharge simulation at hydrometric stations LX, MZ, and JK. When applying the NSE_in, 50 m^3/s was added to each value to avoid the high values obtained by the inversed form of very low-flow values. The forms of NSE and the adapted NSE_in are listed below.

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n} (Y_i^{\mathrm{obs}} - Y_i^{\mathrm{sim}})^2}{\sum_{i=1}^{n} (Y_i^{\mathrm{obs}} - Y_{\mathrm{mean}}^{\mathrm{obs}})^2} \quad (1)$$

$$\mathrm{NSE\_in} = 1 - \frac{\sum_{i=1}^{n} (1/(Y_i^{\mathrm{obs}} + 50) - 1/(Y_i^{\mathrm{sim}} + 50))^2}{\sum_{i=1}^{n} (1/(Y_i^{\mathrm{obs}} + 50) - 1/(Y_{\mathrm{mean}}^{\mathrm{obs}} + 50))^2} \quad (2)$$

where Y_i^{obs} and Y_i^{sim} are the observed or simulated values at time step i, respectively; $Y_{\mathrm{mean}}^{\mathrm{obs}}$ is the mean value over the period n.

Multi-metric

Actual ET. The observed actual ET (AET) was obtained from the MOD16 ET at a monthly time step from 2008 to 2014. A global average mean absolute error (MAE) of 24.1% was obtained by validating at 46 field-based eddy covariance flux towers (Running *et al.* 2018). Therefore, in this study, an uncertainty band with MAE at ±24.1% was applied to the MOD16 ET. The evaluation statistic of the AET simulation was computed as $\mathrm{AET}_{\mathrm{cover}}$, which is the coverage of the simulated AET by the uncertainty band, as listed below.

$$\mathrm{AET}_{\mathrm{cover}} = \frac{\mathrm{No\_of\_covered}}{N} \quad (3)$$

where N is the total number of simulated AET in a simulation, and No_of_covered is the number of simulated AET covered by the MOD16 ET uncertainty band.

Baseflow and surface runoff. The ratio of the baseflow to the discharge and the surface runoff to the discharge was derived from the separation of the hydrograph by applying the Lyne–Hollick baseflow filter, which was standardized by Ladson *et al.* (2013) and published as an R package (hydrostats).

The simulated monthly ratio of surface runoff (including lateral flow) or groundwater return flow to water yield was computed for each subcatchment cluster. Percent bias (PBIAS) was used to evaluate the deviation of the simulated ratio to the observed.

$$\mathrm{PBIAS} = \frac{\sum_{i=1}^{n} (Y_i^{\mathrm{sim}} - Y_i^{\mathrm{obs}}) \times 100}{\sum_{i=1}^{n} (Y_i^{\mathrm{obs}})} \quad (4)$$

The calibration approach

Single-objective calibration with Sequential Uncertainty Fitting ver. 2

SUFI-2 (Sequential Uncertainty Fitting ver. 2) (Abbaspour *et al.* 2004) was the single-objective calibration approach

applied in the study, and the only objective is the simulated discharge at JK evaluated by the NSE. The latin hypercube sampling (LHS) generated the parameter sets for every iteration, which is assumed to be a uniform distribution of the parameters in their selected value range. Accordingly, a Jacobian matrix was formed by all the parameter sets, and the Cremér–Rao theorem updated the parameter values for the next iteration. In this study, two iterations were performed for the hydrologic calibration and each iteration contained 2,000 simulations.

Multi-objective calibration with ED

The equation below listed the ED computed to aggregate the evaluation statistic for each objective.

$$\mathrm{ED} = \sqrt{\left[\sum_{i=1}^{N} (\mathrm{Obj}_i)^2\right]/N} \tag{5}$$

where Obj_i is the evaluation statistic of each objective, and N is the number of objectives.

Multi-objective calibration with NSGA-II

NSGA-II (Deb *et al.* 2002) evolves (updates) the parameter values generation (iteration) by generation. The population size is defined as the constant number of simulations to run in a generation. The nondominated sorting is to evaluate if a simulation dominates the other one, and it is defined that if a simulation $\vec{S1}$ dominates $\vec{S2}$, then (1) $\vec{S1}$ is no worse than $\vec{S2}$ in any objective and (2) $\vec{S1}$ has at least one objective better than $\vec{S2}$. Via nondominated sorting, the simulations that are not dominated by any others will form the Pareto Front.

The procedure of NSGA-II was narrated in detail by Deb *et al.* (2002). The procedure coupling the SWAT and NSGA-II is displayed in Figure 2. The SWAT model is first run based on the parameter sets obtained from the LHS technique as the initial iteration. The nondominated sorting is then performed to generate the Pareto Fronts, from Rank $= 1$ to m. The Pareto Front of Rank $= 1$ contains simulations that are not dominated by any others; the Pareto Front of Rank $= 2$ contains simulations that are not dominated by any others, except the ones from Rank $= 1$ and so on. n simulations (population size) are selected, starting from

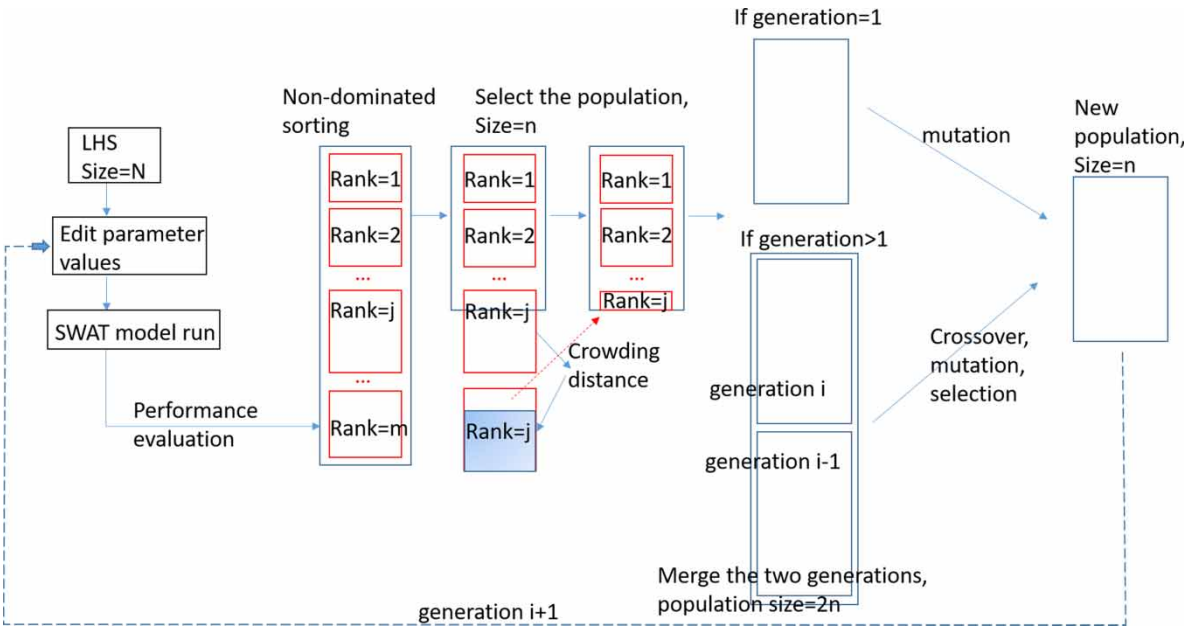


Figure 2 | The calibration procedure of NSGA-II for the SWAT model, adapted from Ercan & Goodall (2016).

the Rank = 1 on. When Rank = j can only be partially selected, the priority of the simulations in Rank = j will be determined by the crowding distance that simulations with distinct performances are preferred. After forming one generation, parameter sets, respectively, are updated via mutation and crossover. The diversity of the population is achieved by the crowding distance, and elitism is maintained by the merge of the current and previous generations for evolution. In this study, the NSGA-II was executed by R language, with the code adapted from Whittaker (2016) and Ercan & Goodall (2016).

The evaluation of the convergence. The convergence of the Pareto Front indicates how well the former generation can surpass the latter generation, which is determined by the population size and the number of generations. Therefore, the convergence was used in this study as a sign of whether more generations are needed for evolution. The convergence is quantified by the C function (Zitzler & Thiele 1999).

$$C(G_{n-1}, G_n) := \frac{|\vec{S'} \in G_n; \exists\, \vec{S} \in G_{n-1} : \vec{S'} \leq \vec{S}|}{|G_n|} \quad (6)$$

where G_n and G_{n-1} are n and $(n-1)$ generations; $\vec{S}$ or $\vec{S'}$ is a random simulation in G_n or G_{n-1}; $S' \leq S$ means that $\vec{S'}$ is dominated or equal to $\vec{S}$. $C(G_{n-1}, G_n)$ is a fraction, of which the denominator is the number of simulation in G_n and the nominator is the number of simulations in G_n that is dominated by or equal to simulations in G_{n-1}. When $C(G_{n-1}, G_n)$ equals 1, all simulations in G_n are dominated by or equivalent to G_{n-1}.

Table 2 | The standardization of the evaluation statistics

Evaluation statistics	Original value range	Standardization	Standardized value range
NSE	$(-\infty, 1)$	$1 - NSE$	$(0, +\infty)$
NSE_in	$(-\infty, 1)$	$1 - NSE\_in$	$(0, +\infty)$
AET_{cover}	(0, 1)	$1/AET_{cover}$	(0, 1)
PBIAS	$(-\infty, +\infty)$	|PBIAS|/100	$(0, +\infty)$

The standardized evaluation statistics

The original value ranges of the evaluation statistics differed, and they were thus standardized to enable the optimization by the nondominated sorting or the ED and to have the optimal value of 0, as listed in Table 2.

The summary of the metrics

To ensure the applicability and the efficiency of the nondominated sorting algorithm, it was proposed in this study to substitute the 13 individual objectives into three groups of categorized objectives, which are: (1) NSE for the three hydrometric stations; (2) NSE_in for the three stations; and (3) the water balance components. Each categorized objective is the ED of the individual objectives included. The categorized objectives are thus the objectives considered by ED and NSGA-II. All the objectives obtained for the YRC are summarized in Table 3.

The calibration procedure

Ten sensitive parameters (Table 4) were selected for the calibration and assigned with an individual value to each subcatchment cluster (see Figure 1), and each approach thus calibrated 30 parameters in total. The initial iteration

Table 3 | The summary of the objectives for multi- and single-objective calibrations

Categorized objectives	Individual objectives	Description
Multisite with NSE	NSE_LX; NSE_MZ; NSE_JK	**NSE_n_:** the NSE of discharge at the LX, MZ, or JK station
Multisite with NSE_in	NSE_in_LX; NSE_in_MZ; NSE_in_JK	**NSE_in_:** the inversed NSE at the LX, MZ, or JK station
Water balance components	LX_s; MZ_s; JK_s; LX_g; MZ_g; JK_g; AET_{cover}	**_s** and **_g:** the bias of the simulated surface runoff ratio and the baseflow ratio of the discharge at LX, MZ, and JK subcatchment clusters; **AET_{cover}**: the coverage of the AET by the MOD16 ET uncertainty band
Single objective	NSE_JK	NSE at the JK station

Table 4 | The 10 parameters selected for the calibration

Parameter	Description
OV_N	Manning's *n* value for overland flow
ESCO	Soil evaporation compensation factor
POT_FR	The fraction of HRU area that drains into the pothole
POT_VOL	The initial volume of water stored in the pothole (mm)
DEP_IMP	Depth to impervious layer for modeling perched water tables (mm)
DIS_STREAM	Average distance to stream (m)
CN2	SCS runoff curve number for moisture condition II
SOL_AWC	Available water capacity of the soil layer
GWQMN	Threshold depth of water in the shallow aquifer required for return flow to occur (mm)
CH_N2	Manning's *n* value for the main channel

included 2,000 simulations and applied to SUFI-2, ED, and NSGA-II. Each of the calibration approaches performed another 2,000 simulations as the next iteration or generations. Therefore, for each calibration approach, 4,000 simulations were performed in total. The procedure is displayed in Figure 3. The best simulation of SUFI-2 is the one with the lowest NSE (standardized) at JK, and ED is the one with the lowest ED value. NSGA-II derived the Pareto Front with optimal simulations, and thus the best simulation is defined as the one with the lowest ED value.

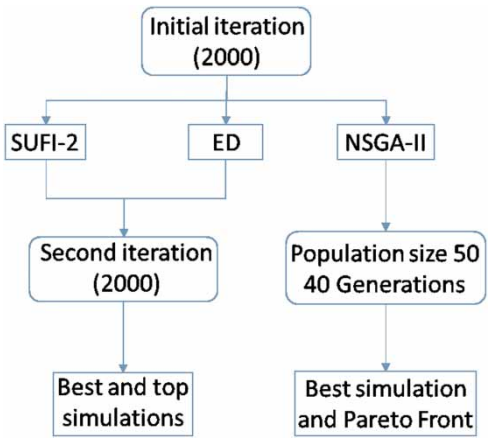


Figure 3 | The hydrological calibration procedure.

In this study, the Pareto Front contained 31 simulations. Therefore, to process the comparison, the first 31 simulations according to their performances were also obtained from SUFI-2 and ED.

RESULTS

The convergence of the NSGA-II

Due to the comparably long time required for the calibration, i.e. approximately 2 min/simulation (approximately 2.5 days/2,000 simulations) of NSGA-II and approximately 1.5 min/simulation (approximately 2 days/2,000 simulations) of ED or SUFI-2 for a standard PC of core i7, 4.20 GHz, and 16GB RAM, the number of simulations is a critical factor to determine the efficiency of the calibration approaches. Therefore, the population size and generations should be first analyzed. To be consistent with SUFI-2 and ED, the population size that reached the convergence within 2,000 simulations was preferred. Figure 4 displays the total simulations tested for population sizes of 50 (80 generations and 4,000 simulations), 100 (40 generations and 4,000 simulations), and 200 (10 generations and 2,000 simulations). Shafii & Smedt (2009) proposed to determine the convergence to be reached if the C function is 1 in consecutive 10 generations. In this study, due to the preferred lower number of simulations, the convergence was defined to be reached if C function values were always above 0.8 after this generation.

In Figure 4, though the C function values are fluctuating, as the generation increases, population sizes of 50 and 100

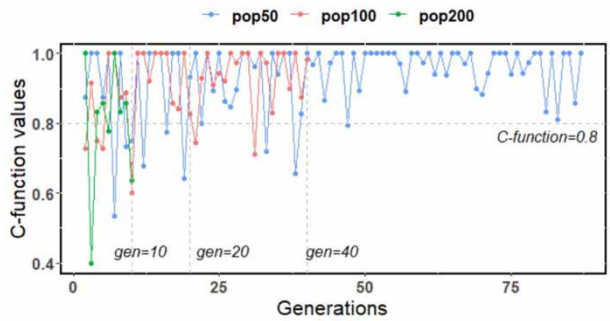


Figure 4 | The convergence evaluated by the C function of population sizes of 50, 100, and 200.

can reach the convergence within 2,000 simulations. With a population size of 50, when the generations exceeded 40, its C function value all reached 0.8 with only one exception. With a population size of 100, when the generation exceeded 20, two exceptions under 0.8 can be observed. However, within 10 generations, a population size of 200 was not able to reach the convergence. Therefore, independent from the number of simulations, the larger the number of generations was, the higher the possibility of the convergence was reached. A population size of 50 was thus selected for the study.

The parameters for calibration

Figure 5 lists the parameter values of the best and the 31 simulations of SUFI-2, NSGA-II, or ED. The parameter values were standardized to the range of [0, 1] (parameters with an absolute value change, e.g. GWQMN) or [−1, 1] (parameters with a relative value change, e.g. CN2) for displaying.

The analysis of the standardized parameter values was focused on interpreting the pattern of the value distribution

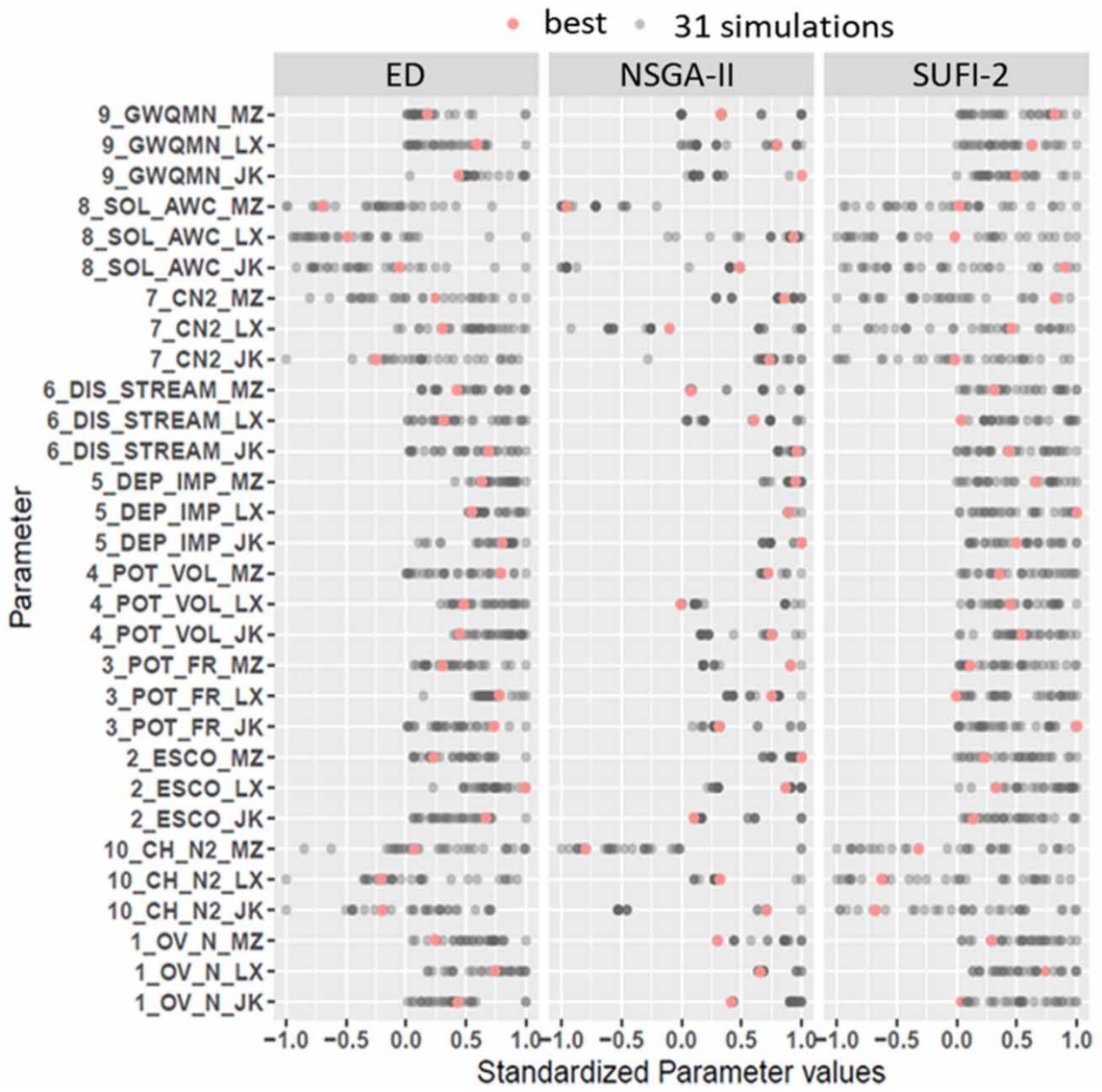


Figure 5 | The distribution of the standardized parameter values after calibration.

as shown in Figure 5. Among the selected 31 simulations, none of the parameters was constrained by any calibration approach into one value; instead, they were all scattered distributed, indicating that the equifinality resulted from the integration of multiple processes in one model. However, certain unique patterns could still be observed among the approaches. SUFI-2 had the most obvious uniform distribution, and the majority of the parameters still possessed the original value ranges. On the contrary, the unique pattern was derived by NSGA-II, in which most parameters were constrained to certain values, e.g. GWQMN_MZ, OV_N_JK, and OV_N_LX. Other parameters could also be observed with the narrowest value range, e.g. DIS_STREAM_JK, DEP_IMP, and POL_VOL_LX. When analyzing ED, the parameter uncertainty range was between NSGA-II and SUFI-2.

The performance of the objectives

The multisite objectives evaluated by NSE and NSE_in

Figure 6 indicates the prediction uncertainty of all the 13 individual objectives of the 31 simulations in the calibration period by a violin plot. In terms of evaluating the discharge at the three hydrometric stations, SUFI-2, derived the lowest NSE_JK value (NSE of the simulated discharge at the JK

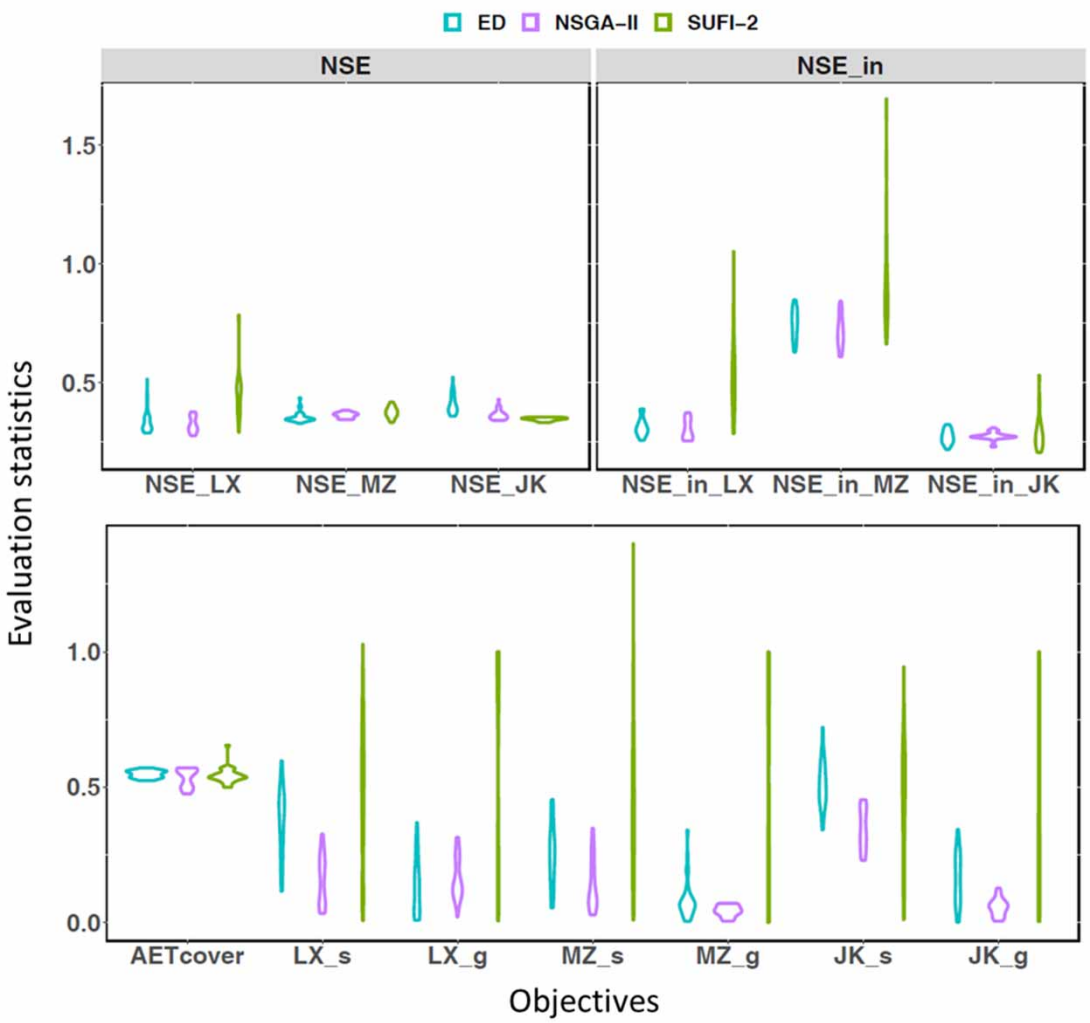


Figure 6 | The violin plot of the evaluation statistics of the objectives.

station) among the 4,000 simulations, ranged from 0.33 to 0.35 (a standardized value, see Table 2). NSE_MZ (at the MZ station) of SUFI-2 also showed a relatively low uncertainty. However, NSE_LX of SUFI-2 had an obvious larger uncertainty range, from approximately 0.29 to 0.78, whereas ED and NSGA-II had values from approximately 0.29 to 0.51 and a denser distribution can be found at the low values. If simulations with an NSE (standardized value) lower than 0.5 are assessed as satisfactory (Moriasi *et al.* 2015), SUFI-2 had an unacceptable performance at LX. ED also had unacceptable simulations at both LX and JK, however, with a smaller number than SUFI-2, whereas NSGA-II derived an acceptable NSE at all the three stations. The NSE_in at the three stations of SUFI-2 also derived a significantly larger uncertainty than NSE. On the contrary, ED and NSGA-II obtained a similar and much narrower uncertainty range of NSE_in at all stations. However, the NSE_in_MZ of ED and NSGA-II ranged from 0.61 to 0.85. Though they have a lower uncertainty than the NSE_in_MZ of SUFI-2, the performance was not comparable to NSE_in_LX and NSE_in_JK.

The performance of the multi-metric objectives

Consistent with the prediction uncertainty of the discharge, SUFI-2 also derived a larger uncertainty range of all metrics (see Figure 6), especially the PBIAS of the surface runoff ratio (LX_s, MZ_s, and JK_s) and the baseflow (LX_g, MZ_g, and JK_g) ratio, where the 31 simulations had a rather scattered distribution of the PBIAS values (the value ranged from approximately 0 to larger than 1). NSGA-II derived an overall lower uncertainty at most metrics, except AET_{cover}, and it also generated the best performance of most metrics, except LX_g. AET_{cover} demonstrated a more concentrated distribution of all the three approaches that the value ranged from 0.5 to 0.6.

When analyzing the monthly baseflow filtered from discharge at the JK station in Figure 7, SUFI-2 underestimated the baseflow of the entire period, whereas ED and NSGA-II had both overestimation and underestimation periods. The baseflow index of the observed discharge at JK was 0.48, and NSGA-II had the lowest bias of baseflow index, which was 0.57. The baseflow index of ED was 0.65, and SUFI-2 was 0.36. The daily discharge has shown that SUFI-2

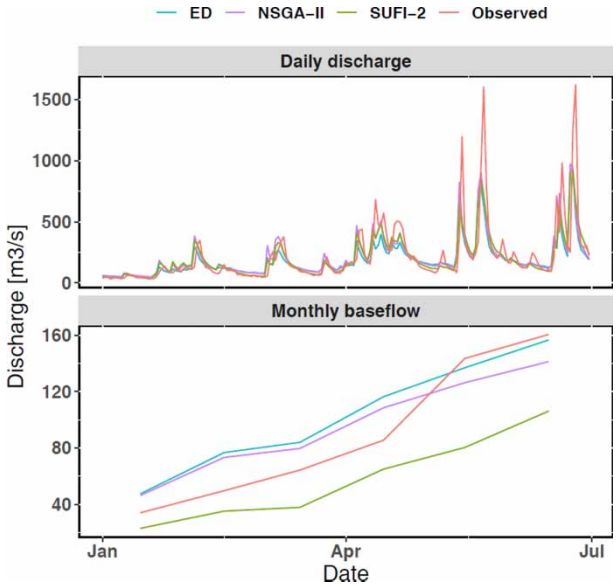


Figure 7 | Daily discharge and the monthly baseflow filtered from hydrograph at the JK station in 2010 from January to June.

derived, except at the peak flow period, lower simulated values than ED and NSGA-II. When analyzing the discrepancy of the simulated discharge to the observed discharge, none of the calibration approaches were able to capture the magnitude of the peak flow after April. However, NSGA-II still derived the least discrepancy during the most peak flow periods.

The hydrological components were validated from 2011 to 2014. The selected 31 simulations are displayed in Figure 8 on five objectives, due to the reason that only discharge at JK was available from 2011 to 2014. The simulated discharge evaluated by NSE or NSE_in was consistent with the calibration period that NSGA-II derived the

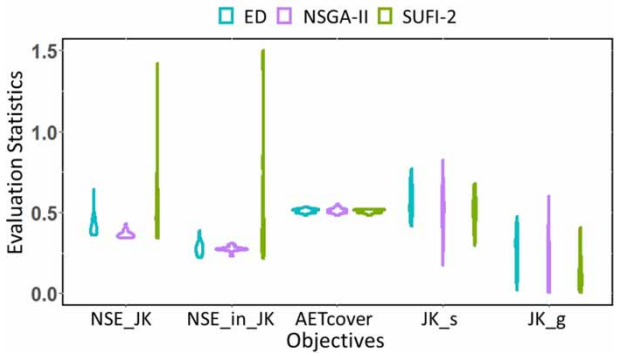


Figure 8 | The prediction uncertainty of the individual objectives of the validation.

least uncertainty range (approximately 0.4–0.5) and followed by ED (approximately 0.4–0.7). For JK_g and JK_s, ED displayed an equivalent uncertainty range as in the calibration period, whereas NSGA-II indicated a larger range of the prediction uncertainty. On the contrary, SUFI-2 obtained a lower uncertainty range of JK_g and JK_s in the validation period than ED or NSGA-II. AET_{cover} can still be observed with a smaller uncertainty of all calibration approaches at the validation period.

The best simulation comparison

The radar chart (Figure 9) displayed the best simulation of SUFI-2, ED, and NSGA-II. It is obvious to visualize the biased results obtained from SUFI-2 that larger values of most objectives were obtained. In contrast, though ED and NSGA-II had higher NSE_JK, JK_s, and AET_{cover} values than SUFI-2, the remaining objectives showed a significantly better performance. ED and NSGA-II generated an equivalent simulation of the discharge at all stations, and NSGA-II had a lower bias of baseflow and surface runoff ratios at LX and JK.

The water balance components were generated at each subcatchment cluster (see Figure 10). The components were the annual average value in mmH_2O at the HRU level in the calibration period. The simulated AET was approximately 500 mm without a significant difference among clusters or calibration approaches. However, the annual average AET of the MOD16 ET dataset in the YRC was 839 mm, and the uncertainty ranged from 637 to 1,041 mm, which showed the model's underestimation of AET. Beside AET, the lateral flow also obtained a low variation among calibration approaches. In MZ and JK clusters, the value ranged from 78 to 190 mm, where the highest lateral flow was derived from ED and the lowest from NSGA-II; at LX, the lateral flow was from 177 to 242 mm, and NSGA-II derived the highest values.

In JK and MZ clusters, a larger difference of percolation can be observed among calibration approaches, S_runoff (surface runoff) and GW_returnflow (baseflow). SUFI-2 obtained a much higher surface runoff at all the stations than baseflow. On the contrary, in the MZ cluster, both NSGA-II and ED indicated a comparable surface runoff and the baseflow rate. However, in the JK cluster, NSGA-II obtained a much higher surface runoff and a lower baseflow rate than ED. The variation of the water balance components was consistent with the variation of the parameter values (see Figure 5) among the stations and the calibration approaches. Parameter GWQMN was comparably higher of SUFI-2 at all clusters. The direct impact was

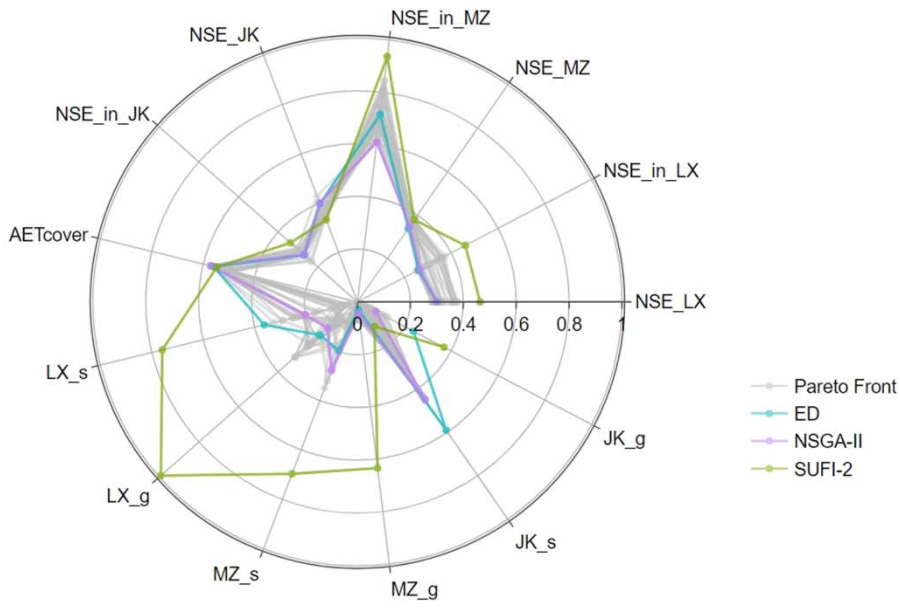


Figure 9 | The radar chart of the best simulation of SUFI-2, ED, and NSGA-II.

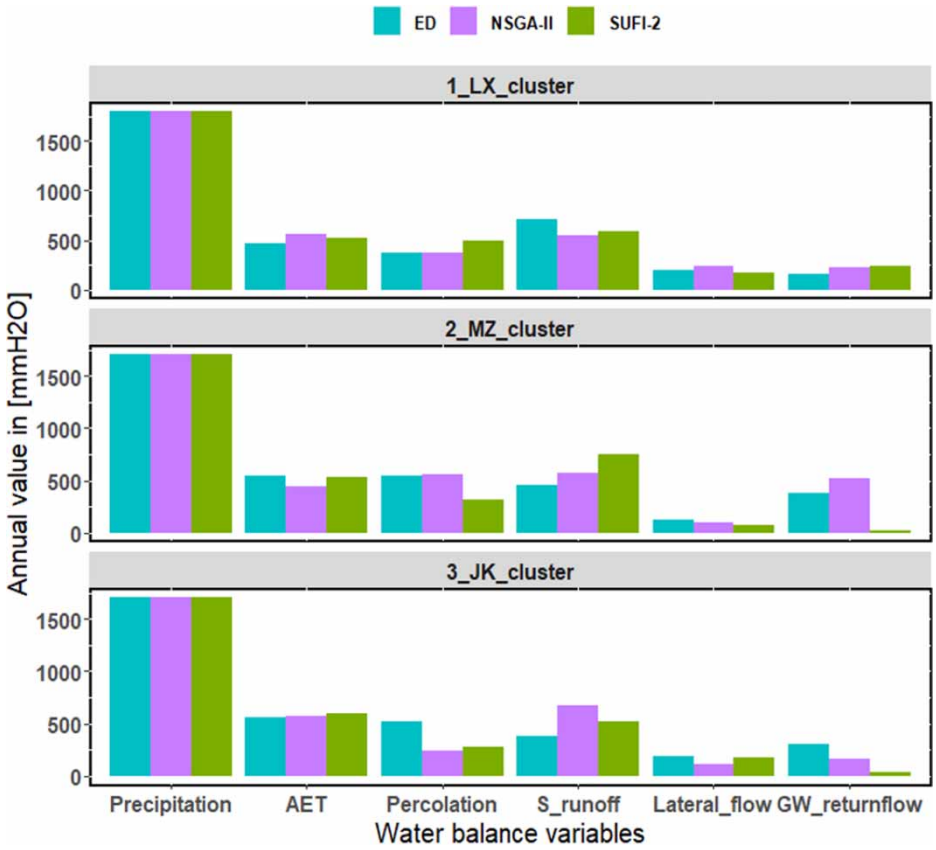


Figure 10 | The catchment average annual water balance components.

thus the higher threshold for baseflow to occur and, consequently, a lower simulated baseflow than the other two approaches. A higher POT_VOL and CN2 at JK of NSGA-II would also contribute to the higher surface runoff rate.

DISCUSSION

The evaluation of the objectives

The multi-metric objectives, excluding AET, were derived from the existing measured discharge. Multi-objective function is applied to ensure both low and peak values of the time series. These two categories were applied in the calibration without the effort of collecting additional observations. Therefore, the fundamental observations, e.g. discharge, are likely sources to obtain objectives for catchments without additional observations.

The AET obtained from the MOD16 ET was applied previously as a reference to assess the model's performance, and van Griensven *et al.* (2012) have concluded an underestimation of the AET by the SWAT. The MOD16 ET was only validated in limited locations globally, and the accuracy at the study area can thus not be fully guaranteed. Therefore, the percent of the coverage by the uncertainty band of the MOD16 ET was used for evaluation, aiming at reducing the impact of the observation uncertainty. However, only 52% of the simulated AET was covered by the uncertainty band. Moreover, compared to other objectives, the narrower prediction uncertainty of AET also indicated the insensitivity of the parameters to the relevant processes related to the AET. Besides, the baseflow and surface runoff ratios were only the estimation based on an empirical approach and cannot fully represent the real condition. The metrics were thus evaluated at a monthly step by the PBIAS to reduce the impact of the

uncertainty. Therefore, when evaluating the calibration results, higher tolerance of the multi-metric in terms of the uncertainty and the performance should be considered.

The evaluation of the three calibration approaches

The trade-off among objectives

The trade-off of the single-objective calibration behaved to optimize the objective of NSE_JK at the sacrifice of the performance at the other sites and metrics. It only ensured the good performance of NSE_JK and NSE_MZ, which might be due to the relatively shorter distance and similar magnitude of the observed discharge at JK and MZ. For single-objective calibration, sensitive processes included the surface runoff, which contributed the highest portion to the discharge in the YRC. As a result, this process was mathematically optimized to fit the observations and especially to the peak values. Meanwhile, the lateral flow and groundwater return flow were relatively low when compared to the result from multi-objective calibration at both MZ and JK. The multi-objective calibration set extra conditions to ensure the reliability of additional specific processes computed by the SWAT model, therefore, to avoid unrealistic trade-off among the objectives, which can be observed at the varied surface runoff and baseflow values in the three subcatchments (Figure 10). Though the trade-off is still unavoidable at the processes without outputs calibrated, compared to single-objective calibration, multiple objectives showed its advantages by deriving the acceptable performance of the simulated discharge in terms of the magnitude and the water balance components at the three catchment clusters.

The applicability of the NSGA-II

The efficiency of deriving the Pareto Front by the nondominated sorting was determined by the number of the objectives. The larger the number is, the less likely a simulation can dominate the others on all the objectives to generate the Pareto Front. Her & Seong (2018) also pointed out no significant effect to decrease the parameter uncertainty if the objective functions were more than four. Therefore, in this study, three categorized objectives were proposed to substitute the 13 individual objectives. The applicability of the substitution can be indicated by the results that NSGA-II derived a reliable simulation with a narrower uncertainty band of most objectives.

Results showed that in comparison to the number of generations, the convergence was more independent from the population size. Whether the population size was 50, 100, or 200, the convergence cannot be reached within 10 generations. The more generations there are, the more likely the convergence can be reached. Therefore, if the computational effort is critical, as in this study, a smaller population should be considered, which enables more generations to execute. However, since the population size determines the diversity of the parameter set values, a smaller population size will also likely to derive a smaller simulation size in the Pareto Front.

Sorting algorithm comparison

When compared to NSGA-II that the parameters were optimized after each generation, ED is a post-processing procedure in this study. The value range of each objective is not identical in this study. Therefore, in comparison to nondominated sorting, ED is a more subjective approach. The scalar function of ED is computed based on the absolute value of the evaluation statistic. However, if the prediction uncertainty ranges of the individual objectives were too distinct from each other, the emphasis of the sorting will be put naturally on the objectives with the larger uncertainty range. In comparison, nondominated sorting only considers whether a value is surpassing the other instead of the magnitude of the surpassing, and thus the sorting algorithm is not impacted by the absolute value of an object. The best simulation derived by ED is the simulation with the lowest ED. However, NSGA-II with nondominated sorting would derive multiple simulations with equivalent performances. Therefore, if only one simulation is to be selected as the calibration result, criteria should still be set to distinguish simulations at the Pareto Front. Moreover, the processing time to implement the nondominated sorting is proportional to the population size, and the applicability is decreasing as the number of objectives is increasing. However, when applying ED, the number of objectives is no longer a limitation to either

the algorithm or the computational effort, and the computational effort was equivalent to that of single-objective calibration.

The parameter uncertainty

Due to equifinality, multiple parameter sets would generate simulations that meet the requirement of the evaluation criteria. However, the higher the number of the objectives is, the fewer parameter sets could theoretically be qualified. Therefore, it resulted in a shorter range of parameter values of ED and NSGA-II, which might explain the less reliable validation period. ED screened the simulations among all simulations, and multiple parameters have kept the uniform distribution but with a narrower value range. However, the initial parameter values used by NSGA-II were only based on the population size. When selecting the population for the first generation, the value range of the parameters has already been constrained to the values that have a good performance in the first iteration (2,000 simulations) in the study, and the parameter values of the subsequent generations were evolved based on the first generation. Though mutation and crossover were performed, the diversity of the parameter values was not comparable to SUFI-2 or ED. However, the larger the population size is, the more diverted the parameter values can be selected. However, to guarantee the convergence and the stability of the performance of the last generation within 2,000 simulations, a population size of 50 was applied. Therefore, comparably more constrained parameter value ranges are displayed, and a narrower prediction uncertainty was also obtained by NSGA-II. Meanwhile, it should be mentioned that, though the multi-objective calibration leads to a reduced parameter and, thus, a prediction uncertainty, it cannot overcome the uncertainty introduced by the input data scarcity, e.g. the accuracy of the rainfall data or the measured hydrometric data, and the model itself, e.g. the limitation of the empirical SCS curve number method to capture the surface runoff, as displayed by the relatively large discrepancy of the simulated peak values in Figure 7, which also indicates the limits of the calibration approaches in terms of improving the model's performance.

CONCLUSION

In terms of the comparison between single-objective calibration and multi-objective calibration, the key findings are summarized as follows:

1. The trade-off is unavoidable among objectives. However, multi-objective calibration can add extra constraints to the model to ensure the reliability of the critical internal processes for model application.
2. The applicability of NSGA-II is also determined by the proper size of population and generation. If the computational effort is critical, the smaller size of the population is preferred.
3. ED and NSGA-II are both suitable multi-objective calibration approaches, represented by a constrained parameter uncertainty and prediction uncertainty.

Therefore, it is highly recommended to apply the multi-objective calibration to the process-based hydrological model, like SWAT, even to the catchment with limited observations. The multi-objective calibration framework applied in the study is also provided a practical procedure for the calibration of catchments similar to the YRC. The key points include the following:

1. Preparing the observations according to the multisite, multi-objective function, and multi-metric with the best use of already obtained datasets, e.g. measured discharge, but not necessarily the objectives applied in the study.
2. If the ED or NSGA-II is considered as the calibration approach, their applicability and suitability should be analyzed according to the processing time, the number of objectives, the objectivity of the dominance, and the equifinality.

The advantage of the multi-objective calibration applied for the process-based hydrological model is clearly illustrated in the study, which would be a great benefit for future application of the model, e.g. an expected improvement of the nutrient load simulation at the inner stations due to a more reliable simulated discharge. However, multi-objective calibration cannot overcome the uncertainty introduced by the model's approach and the input data;

therefore, further research contributed to these two aspects is also highly promoted.

ACKNOWLEDGEMENTS

The first author is financially supported by the Chinese Scholarship Council (CSC).

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

REFERENCES

Abbaspour, K. C., Johnson, C. A. & van Genuchten, M. T. 2004 Estimating uncertain flow and transport parameters using a sequential uncertainty fitting procedure. *Vadose Zone Journal* **3** (4), 1340–1352.

Abiodun, O. O., Guan, H., Post, V. E. A. & Batelaan, O. 2017 Comparison of MODIS and SWAT evapotranspiration over a complex terrain at different spatial scales. *Hydrology and Earth System Sciences* **22**, 2775–2794.

Arnold, J. G., Allen, P. M., Muttiah, R. & Bernhardt, G. 1995 Automated base flow separation and recession analysis techniques. *Groundwater* **33** (6), 1010–1018.

Deb, K., Pratap, A., Agarwal, S. & Meyarivan, T. 2002 A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* **6** (2), 182–197.

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.

Fang, Z. 2011 The Report of Integrated Planning of Yuan River Catchment, Jiangxi Prov. [in Chinese]. Jiangxi Water Conservancy Planning Design Institute, Nanchang.

Gupta, H. V., Sorooshian, S. & Yapo, P. O. 1998 Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information. *Water Resources Research* **34** (4), 751–763.

Gupta, H. V., Wagener, T. & Liu, Y. 2008 Reconciling theory with observations: elements of a diagnostic approach to model evaluation. *Hydrological Processes* **22** (18), 3802–3813.

Gupta, H. V., Kling, H., Yilmaz, K. K. & Martinez, G. F. 2009 Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. *Journal of Hydrology* **377** (1–2), 80–91.

Her, Y. & Seong, C. 2018 Responses of hydrological model equifinality, uncertainty, and performance to multi-objective parameter calibration. *Journal of Hydroinformatics* **20** (4), 864–885.

Herman, M. R., Nejadhashemi, A. P., Abouali, M., Hernandez-Suarez, J. S., Daneshvar, F., Zhang, Z., Anderson, M. C., Sadeghi, A. M., Hain, C. R. & Sharifi, A. 2018 Evaluating the role of evapotranspiration remote sensing data in improving hydrological modeling predictability. *Journal of Hydrology* **556**, 39–49.

Krause, P., Boyle, D. P. & Bäse, F. 2005 Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences* **5**, 89–97.

Ladson, A. R., Brown, R., Neal, B. & Nathan, R. 2013 A standard approach to baseflow separation using the Lyne and Hollick filter. *Australian Journal of Water Resources* **17**, 1.

Leta, O. T., van Griensven, A. & Bauwens, W. 2017 Effect of single and multisite calibration techniques on the parameter estimation, performance, and output of a SWAT model of a spatially heterogeneous catchment. *Journal of Hydrologic Engineering* **22** (3).

Moriasi, D. N., Gitau, M. W., Pai, N. & Daggupati, P. 2015 Hydrologic and water quality models. Performance measures and evaluation criteria. *Transactions of the ASABE* **58** (6), 1763–1785.

Neitsch, S. L., Arnold, J. G., Kiniry, J. R. & Williams, J. R. 2011 *Soil and Water Assessment Tool, Theoretical Documentation: Version 2009*. Texas A&M University, Temple.

Pfannerstill, M., Guse, B. & Fohrer, N. 2014 Smart low flow signature metrics for an improved overall performance evaluation of hydrological models. *Journal of Hydrology* **510**, 447–458.

Pfannerstill, M., Bieger, K., Guse, B., Bosch, D. D., Fohrer, N. & Arnold, J. G. 2017 How to constrain multi-objective calibrations of the SWAT model using water balance components. *JAWRA Journal of the American Water Resources Association* **53** (3), 532–546.

Pushpalatha, R., Perrin, C., Le Moine, N. & Andréassian, V. 2012 A review of efficiency criteria suitable for evaluating low-flow simulations. *Journal of Hydrology* **420–421**, 171–182.

Running, S. W., Mu, Q., Zhao, M. & Moreno, A. 2018 *User's Guide: MODIS Global Terrestrial Evapotranspiration (ET) Product (NASA MOD16A2/A3): NASA Earth Observing System MODIS Land Algorithm*. NTSG University of Montana, Missoula.

Schaffer, J. 1984 *Some Experiments in Machine Learning Using Vector Evaluated Genetic Algorithms*. PhD Thesis, Vanderbilt University, Nashville, TN, USA.

Shafii, M. & Smedt, F. d. 2009 Multi-objective calibration of a distributed hydrological model (WetSpa) using a genetic algorithm. *Hydrology and Earth System Sciences* **13** (11), 2137–2149.

Shrestha, M. K., Recknagel, F., Frizenschaf, J. & Meyer, W. 2016 Assessing SWAT models based on single and multi-site calibration for the simulation of flow and nutrient loads in

the semi-arid Onkaparinga catchment in South Australia. *Agricultural Water Management* **175**, 61–71.

van Griensven, A., Breuer, L., Di Luzio, M., Vandenberghe, V., Goethals, P., Meixner, T., Arnold, J. & Srinivasan, R. 2006 Environmental and ecological hydroinformatics to support the implementation of the European Water Framework Directive for river basin management. *Journal of Hydroinformatics* **8** (4), 239–252.

Van Griensven, A., Maskey, S. & Stefanova, A. 2012 The use of satellite images for evaluating a SWAT model: Application on the Vit basin, Bulgaria. In: iEMSs 2012 – Managing Resources of a Limited Planet: Proceedings of the 6th Biennial Meeting of the International Environmental Modelling and Software Society, pp. 3030–3037.

Whittaker, G. 2016 *Non-dominated Sorting Genetic Algorithm-II (NSGA-II) in R, USDA-ARS*. https://www.ars.usda.gov/research/software/download/?softwareid=393&modecode=20-72-05-00, accessed 2 May 2016.

Yen, H., Bailey, R. T., Arabi, M., Ahmadi, M., White, M. J. & Arnold, J. G. 2014 The role of interior watershed processes in improving parameter estimation and performance of watershed models. *Journal of Environment Quality* **43** (5), 1601–1613.

Yilmaz, K. K., Gupta, H. V. & Wagener, T. 2008 A process-based diagnostic approach to model evaluation: application to the NWS distributed hydrologic model. *Water Resources Research* **44** (9), 103.

Zhang, X., Srinivasan, R. & van Liew, M. 2008 Multi-site calibration of the SWAT model for hydrologic modeling. *Transactions of the ASABE* **51** (6), 2039–2049.

Zhang, Y., Shao, Q., Zhang, S., Zhai, X. & She, D. 2016 Multi-metric calibration of hydrological model to capture overall flow regimes. *Journal of Hydrology* **539**, 525–538.

Zitzler, E. & Thiele, L. 1999 Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. *IEEE Transactions on Evolutionary Computation* **3** (4), 257–271.

First received 18 August 2020; accepted in revised form 8 January 2021. Available online 29 January 2021