Engineering 28 (2023) 93–104

Contents lists available at ScienceDirect

Engineering

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

Research
Hydraulic Engineering—Article

Runoff Modeling in Ungauged Catchments Using Machine Learning Algorithm-Based Model Parameters Regionalization Methodology

Houfa Wu [a,b,c], Jianyun Zhang [b,c], Zhenxin Bao [b,c,*], Guoqing Wang [b,c], Wensheng Wang [a], Yanqing Yang [a,b,c], Jie Wang [b,c]

[a] State Key Laboratory of Hydraulics and Mountain River Engineering, College of Water Resource and Hydropower, Sichuan University, Chengdu 610065, China
[b] State Key Laboratory of Hydrology–Water Resources and Hydraulic Engineering, Nanjing Hydraulic Research Institute, Nanjing 210029, China
[c] Research Center for Climate Change, Ministry of Water Resources, Nanjing 210029, China

ARTICLE INFO

Article history:
Received 21 June 2021
Revised 22 December 2021
Accepted 26 December 2021
Available online 18 February 2022

Keywords:
Parameters estimation
Ungauged catchments
Regionalization scheme
Machine learning algorithms
Soil and water assessment tool model

ABSTRACT

Model parameters estimation is a pivotal issue for runoff modeling in ungauged catchments. The nonlinear relationship between model parameters and catchment descriptors is a major obstacle for parameter regionalization, which is the most widely used approach. Runoff modeling was studied in 38 catchments located in the Yellow–Huai–Hai River Basin (YHHRB). The values of the Nash–Sutcliffe efficiency coefficient (NSE), coefficient of determination (R^2), and percent bias (PBIAS) indicated the acceptable performance of the soil and water assessment tool (SWAT) model in the YHHRB. Nine descriptors belonging to the categories of climate, soil, vegetation, and topography were used to express the catchment characteristics related to the hydrological processes. The quantitative relationships between the parameters of the SWAT model and the catchment descriptors were analyzed by six regression-based models, including linear regression (LR) equations, support vector regression (SVR), random forest (RF), *k*-nearest neighbor (*k*NN), decision tree (DT), and radial basis function (RBF). Each of the 38 catchments was assumed to be an ungauged catchment in turn. Then, the parameters in each target catchment were estimated by the constructed regression models based on the remaining 37 donor catchments. Furthermore, the similarity-based regionalization scheme was used for comparison with the regression-based approach. The results indicated that the runoff with the highest accuracy was modeled by the SVR-based scheme in ungauged catchments. Compared with the traditional LR-based approach, the accuracy of the runoff modeling in ungauged catchments was improved by the machine learning algorithms because of the outstanding capability to deal with nonlinear relationships. The performances of different approaches were similar in humid regions, while the advantages of the machine learning techniques were more evident in arid regions. When the study area contained nested catchments, the best result was calculated with the similarity-based parameter regionalization scheme because of the high catchment density and short spatial distance. The new findings could improve flood forecasting and water resources planning in regions that lack observed data.

1. Introduction

Hydrological models are popular tools for hydrological process modeling, and these models have been extensively applied in flood forecasting, water resources management, and the assessment of climate change impact in recent decades [1,2]. With the improvement of computer technology and the application of multiple interdisciplinary subjects, hydrological models can now describe hydrological processes more accurately. Hydrological models have developed from the original conceptual models (Tank and Sacramento) and centralized models (Xin'anjiang and simplified hydrology model (SIMHYD)) to the current popular semi-distributed models (TOPMODEL; soil and water assessment tool (SWAT)); and distributed model (variable infiltration capacity (VIC)) [3–5]. The accurate estimation of model parameters directly influences the accuracy of runoff simulation. Generally, the model parameters are optimized and calibrated by observed streamflow data at the outlet of a basin. However, numerous catchments are limited by

[*] Corresponding author.
E-mail address: zxbao@nhri.cn (Z. Bao).

https://doi.org/10.1016/j.eng.2021.12.014

geographical or economic conditions and lack adequate observed data to calibrate model parameters [1]. Therefore, runoff modeling in ungauged catchments has become a focus for researchers [6,7]. This problem is termed the "prediction in ungauged basins" (PUB) in hydrology. To tackle the PUB issues, various regionalization approaches are widely used to simulate runoff in ungauged catchments by transferring the model parameters from similar catchments to ungauged ones [8,9].

The three widely used parameter regionalization approaches are regression-based, physical similarity-based, and spatial proximity-based. Regression analysis method is the most popular and widely studied approach [10,11]. The key steps are to establish regression equations between model parameters and catchment descriptors in gauged catchments, and to estimate model parameters in ungauged catchments with the constructed regression relationship [12,13]. However, some studies have reported that the relationships between model parameters and catchment descriptors are often complex, and estimation in ungauged catchments usually leads to large errors [14].

The physical similarity approach assumes that catchments with the same physical attributes (such as climate, vegetation, and topography) have similar modes of runoff generation and confluence processes [1,15]. The spatial proximity method selects the donor catchments according to the spatial distance between the neighboring observed and ungauged catchments, and the parameters of the donor catchment are transferred to the target catchment [16]. The advantage of the above two approaches over the regression analysis method is that they do not make linear assumptions, and these methods have been widely used in recent years [17,18]. However, the spatial proximity approach is not suitable for the large spatial variation of adjacent basins [19], and the physical similarity approach is limited by the rationality of selecting catchment characteristics [20]. Some researchers have compared and evaluated the three approaches. In most cases, the spatial proximity and the physical similarity approaches are the most effective [21]. In addition, some researchers have combined the physical similarity and spatial proximity approaches to estimate the model parameters of ungauged catchments. The results found that the integrated similarity-based approach performed slightly better than spatial proximity-based or physical similarity-based alone [22].

Catchment descriptors and model parameters are interdependent, and their relationship may be nonlinear [23,24]. Furthermore, a hydrological model is a generalized description of the catchment hydrological process, and it will inevitably have the phenomenon of equifinality, making it challenging to obtain the only optimal solution of the model parameters through calibration. Estimating the model parameters with the traditional multiple regression scheme may result in large errors. With the development of data mining and artificial intelligence technology, the machine learning technique has been successfully applied in flood forecasting, earth science modeling, and remote sensing due to its good performance in dealing with nonlinear relationships [25]. In the last decade, some machine learning models have received increasing attention in the field of model parameters regionalization, including support vector machine (SVM), random forest (RF), and decision tree (DT). For example, Saadi et al. [23] investigated the potential of RF algorithms in the regionalization of the hourly hydrological model parameters. Hao et al. [26] used an RF model to regionalize the parameters of the mountain flood prediction model. Jafarzadegan et al. [14] estimated the parameters of an environmental model in data-scarce regions with the SVM technique. Ragettli et al. [27] used the splitting rules of classification and regression trees (CART) to regionalize the parameters of 35 catchments in China. The results showed that machine learning algorithms could present accurate predictions in general. However, most existing research focuses on the comparative analysis between a single machine learning algorithm and the traditional regionalization approach. The applicability of different machine learning algorithms in parameter regionalization has not been assessed.

The main objective of this paper is to evaluate different machine learning techniques for parameter regionalization in the Yellow–Huai–Hai River Basin (YHHRB), analyzing their advantages and limitations. The performances of the five classical machine learning-based approaches were compared with linear regression (LR)-based and similarity-based schemes (combining the physical similarity and spatial proximity). The performances of the different parameter regionalization approaches in various climate regions were further compared. The sections of this paper are organized as follows: Section 2 describes the study area and the datasets; Section 3 introduces the methodology used; Sections 4 and 5 describe and discuss the regionalization results, and the conclusions are summarized in Section 6.

2. Study area and datasets

Located in northern China (95°E–123°E, 30°N–43°N), the YHHRB is the general name of the three first-class basins (Yellow River Basin, Huai River Basin, and Hai River Basin) in China. The YHHRB covers 16 provinces with a total area of 1 445 000 km^2. The Yellow River Basin, Huai River Basin, and Hai River Basin have drainage areas of 795 000 km^2, 330 000 km^2, and 320 000 km^2, respectively. The population and the gross domestic product (GDP) of the YHHRB account for about 35% and 32% of the national total, respectively. The eastern plains in the YHHRB are a substantial agricultural production base in China, and the areas of cultivated land and grain output account for 20.4% and 23.6% of the country's total, respectively [28]. Thirty-eight typical catchments with different hydrologic and climatic conditions in the YHHRB were selected as the study areas in this study (Fig. 1(a)), including 22 catchments in relatively humid regions (aridity index $\varphi < 1.7$), and 16 catchments in relatively arid regions ($\varphi > 1.7$). The detailed information for the 38 catchments is presented in Table 1.

The monthly mean streamflow of the 38 catchments was obtained from *China's Hydrological Yearbook*, published by the Hydrological Bureau of the Ministry of Water Resources, China. The daily data of the rainfall, temperature, wind speed, relative humidity, and solar radiation during 1961–2015 were extracted from the gridded daily observation dataset over the China region (CN05.1), published by the National Climate Center of the China Meteorological Administration [29]. The digital elevation model (DEM) of the YHHRB was derived from the Shuttle Radar Topography Mission (SRTM) data provided by the Geospatial Data Cloud Platform[†], with a resolution of 30 m, and these data were used to generate the river network of the hydrological model. The land use data in 1980 with a spatial resolution of 1 km were obtained from the Resources and Environment Data Cloud Platform[‡] of the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences (Fig. 1(b)). The soil data was extracted from the Harmonized World Soil Database (HWSD), constructed by the Food and Agriculture Organization (FAO) and the International Institute for Applied Systems Analysis (IIASA), with a spatial resolution of 1 km (Fig. 1(c)).

3. Methodology

The observed streamflow data were used to calibrate the SWAT model parameters of 38 typical catchments, and the sensitivity and

[†] http://www.gscloud.cn/home

[‡] http://www.resdc.cn

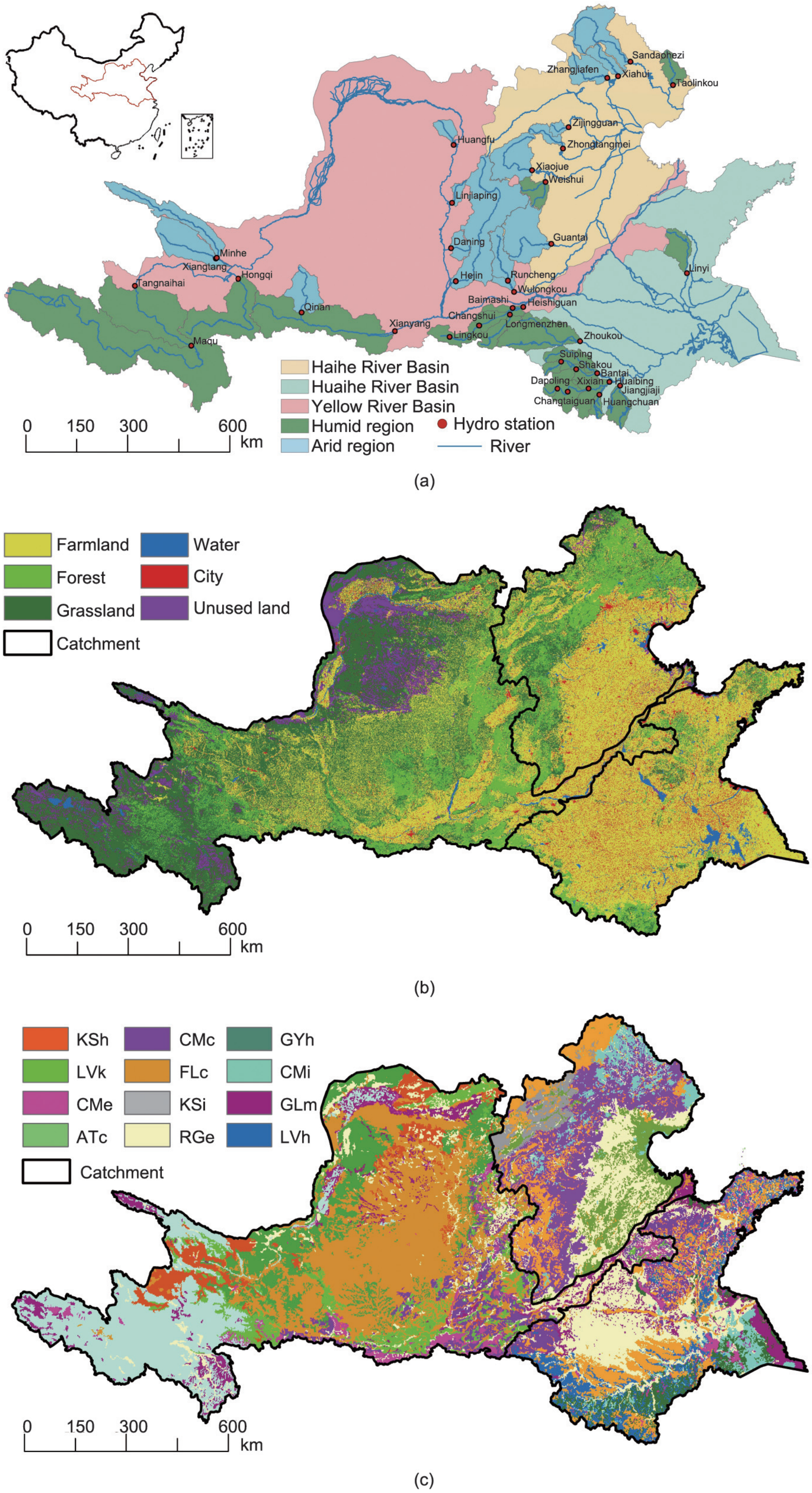

Fig. 1. The basic information of the YHHRB. (a) Location of the study area; (b) landuse type; (c) soil type. KSh: haplic kastanozems; LVk: calcic luvisols; CMe: eutric cambisols; ATc: cumulic anthrosols; CMc: calcaric cambisols; FLc: calcaric fluvisols; KSi: luvic kastanozems; RGe: eutric regosols; GYh: haplic gypsisols; CMi: gelic cambisols; GLm: mollic gleysols; LVh: haplic luvisols.

applicability of these parameters were analyzed. Then, each of the 38 catchments was assumed to be an ungauged catchment in turn, that is, the target catchment. The model parameters of the target catchment were estimated with machine learning techniques, LR-based method, and similarity-based approach. Finally, the estimated model parameters were input into the SWAT model to sim-

Table 1
General information for the 38 catchments in the YHHRB.

Basin	Station	Available data	Longitude (°E)	Latitude (°N)	Area (km^2)	P (mm)	T (°C)	NDVI	φ
Yellow River	Minhe	1962–1980	102.80	36.33	15 296.21	450.05	0.84	0.32	2.15
	Xiangtang	1961–1973	102.83	36.35	13 833.36	460.42	−2.80	0.34	2.13
	Hongqi	1961–1975	103.57	35.80	23 754.83	633.48	1.36	0.41	1.33
	Tangnaihai	1961–1976	100.15	35.50	114 132.04	508.86	−3.11	0.34	1.52
	Huangfu	1961–1979	111.08	39.28	2 995.68	420.87	6.66	0.19	2.47
	Hejin	1961–1971	110.80	35.57	36 753.48	583.02	8.09	0.36	2.03
	Daning	1964–1980	110.72	36.47	3 668.02	559.48	8.66	0.35	1.85
	Qinan	1961–1985	105.67	34.90	9 274.43	539.07	6.13	0.26	1.76
	Wulongkou	1961–1975	112.68	35.15	8 627.86	643.59	8.91	0.45	1.82
	Longmenzhen	1961–1970	112.47	34.55	5 053.27	809.80	11.77	0.49	1.53
	Baimashi	1961–1969	112.58	34.72	9 728.62	750.97	11.15	0.46	1.59
	Changshui	1961–1970	111.44	34.33	5 931.04	765.35	10.40	0.50	1.52
	Heishiguan	1961–1969	112.93	34.72	17 288.82	745.30	11.86	0.46	1.61
	Linjiaping	1961–1970	110.87	37.70	1 765.52	428.10	7.77	0.28	2.34
	Maqu	1963–1976	102.08	33.96	82 808.09	504.34	−3.22	0.34	1.46
	Runcheng	1963–1978	112.51	35.47	6 811.32	640.60	8.68	0.45	1.83
	Xianyang	1964–1980	108.70	34.32	43 108.76	604.57	7.05	0.38	1.53
	Lingkou	1961–1980	110.47	34.08	2 357.79	727.11	10.13	0.53	1.49
Huai River	Bantai	1961–1975	115.06	32.71	11 280.02	870.91	14.62	0.48	1.18
	Jiangjiaji	1968–1980	115.73	32.30	5 634.37	1 305.43	14.37	0.52	0.85
	Linyi	1965–1976	118.40	35.02	6 280.05	731.15	11.91	0.38	1.32
	Xixian	1966–1980	114.73	32.33	6 547.73	1 024.95	14.54	0.49	1.00
	Zhoukou	1961–1975	114.65	33.63	17 100.01	742.70	13.73	0.48	1.56
	Changtaiguan	1965–1980	114.07	32.31	2 940.93	1 032.69	14.44	0.49	1.00
	Huaibin	1965–1980	115.41	32.44	14 645.73	1 047.39	14.73	0.48	0.98
	Huangchuan	1965–1980	115.04	32.13	1 909.01	1 130.45	14.74	0.49	0.90
	Shakou	1970–1981	114.42	32.88	1 878.62	934.07	14.46	0.49	1.14
	Suiping	1965–1980	113.97	33.14	1 673.29	885.96	14.19	0.44	1.20
	Dapoling	1965–1980	113.75	32.43	1 573.70	1 032.86	14.30	0.50	1.01
Hai River	Zhongtangmei	1964–1980	114.88	38.88	3 190.59	501.64	6.41	0.34	1.93
	Xiahui	1961–1980	117.17	40.62	4 132.02	549.64	5.67	0.44	1.81
	Taolinkou	1961–1980	119.05	40.13	4 502.12	636.02	7.74	0.42	1.39
	Guantai	1961–1973	114.08	36.33	16 640.03	643.61	8.62	0.38	1.77
	Weishui	1961–1970	114.13	38.03	4 907.88	695.82	9.18	0.36	1.65
	Zhangjiafen	1961–1980	116.78	40.62	8 031.28	507.13	4.52	0.43	2.11
	Xiaojue	1961–1980	113.71	38.39	13 051.19	514.22	6.23	0.35	1.84
	Zijingguan	1961–1970	115.17	39.43	1 639.49	557.36	6.40	0.40	1.97
	Sandaohezi	1961–1967	117.70	40.97	15 329.89	443.62	2.51	0.36	2.10

P: annual precipitation; T: annual mean temperature; NDVI: normalized difference vegetation index.

ulate the runoff process in the target catchment. Based on the results of the parameter regionalization and the runoff simulation in the 38 catchments, various regionalization approaches performance was evaluated.

3.1. SWAT model

The SWAT model is a physically-based, semi-distributed, and continuous hydrological model developed by the Agricultural Research Service of the US Department of Agriculture (USDA) [30]. The model can simultaneously consider meteorological conditions, soil types, land use patterns, and various water conservancy engineering conditions, and it has been widely used to simulate the hydrological change process at the watershed scale [31]. The detailed steps of model construction can be found in the literature [32]. The parameters of the SWAT model need to be calibrated to achieve the optimal simulation effect after the model is built and run. The sequential uncertainty fitting version 2 (SUFI2) algorithm in SWAT calibration and uncertainty programs (e.g., SWAT-CUP) software is used to calibrate and validate the model parameters, and the effect of the simulation results is evaluated with three indexes: Nash–Sutcliffe efficiency coefficient (NSE), coefficient of determination (R^2), and percent bias (PBIAS), which can be expressed as follows:

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}(Q_s - Q_o)^2}{\sum_{i=1}^{n}\left(Q_o - \overline{Q_o}\right)^2} \tag{1}$$

$$R^2 = \frac{\left[\sum_{i=1}^{n}(Q_o - \overline{Q_o})(Q_s - \overline{Q_s})\right]^2}{\sum_{i=1}^{n}(Q_o - \overline{Q_o})^2\sum_{i=1}^{n}(Q_s - \overline{Q_s})^2} \tag{2}$$

$$\mathrm{PBIAS} = \frac{\sum_{i=1}^{n}(Q_o - Q_s)}{\sum_{i=1}^{n}Q_o} \tag{3}$$

where Q_o and Q_s are the observed and simulated streamflow (m^3·s^{-1}), $\overline{Q_o}$ and $\overline{Q_s}$ are mean observed and simulated streamflow (m^3·s^{-1}), and n is the amount of measured data. Existing studies have demonstrated that the model simulation results are credible when NSE > 0.5, R^2 > 0.5, −25% < PBIAS <25%, and simulation results with NSE above 0.75 are considered to be very good [33].

3.2. Regression-based methods

Six regression-based models were introduced to estimate model parameters, including the LR equations, support vector regression (SVR), RF, k-nearest neighbor (kNN), DT, and radial basis function (RBF). Based on the constructed models, the model parameters were modeled with nine catchment descriptors, including the catchment area (Area), mean catchment elevation (Ele), mean catchment slope (Slope), soil sand content (Sand), soil clay content (Clay), annual precipitation (P), annual mean temperature (T), normalized difference vegetation index (NDVI), and φ. The regression model can be expressed as follows:

$$y = f(x, \boldsymbol{u}) \tag{4}$$

where y and x are the model parameters and the catchment characteristic values, respectively, and $\boldsymbol{u}$ is the vector of the model parameters.

Since the LR analysis cannot describe the nonlinear relationship between the model parameters and the catchment descriptors, the more complex algorithms were used, including SVR, RF, *k*NN, DT, and RBF. As a supervised learning method, the SVR can describe the nonlinear relationship between variables by mapping the kernel function to the high dimensional space [34,35]. The RF is stable and insensitive to overfitting because some training samples are randomly selected from the regression tree. It also has good robustness compared with other algorithms [36]. The *k*NN is a non-parametric estimation method that is fitted by calculating the distance between different eigenvalues of samples, and it does not require making assumptions about data input [37]. The DT does not depend on the distribution of the sample data in the model construction and sample prediction, making the estimate results more stable [27]. The RBF is a type of feedforward neural network with wide application, and it can approximate any arbitrary nonlinear function with unlimited accuracy [38]. Based on these advantages, the above five classical machine learning algorithms were applied to the parameter regionalization as a supplement to the traditional LR approach.

A Taylor diagram was used to quantify the similarity of the model parameters between two patterns (calibration and estimation). It contains three indicators: standard deviations (STDs), root mean squared error (RMSE), and correlation coefficient (r) [39].

3.3. Similarity-based method

The similarity-based approach integrated consideration of both the physical similarity and the spatial proximity, which were combined according to their respective weights. Two options were considered to combine the information from the donor catchments: parameter weighted averaging (PA) and output weighted averaging (OA). The PA method involved combining the model parameters of the donor catchments according to their corresponding weights, and then substituting the integrated parameters into the SWAT model to simulate the runoff of the target catchment (Q_{1j}).

$$Q_{1j} = Q\left(j,\ \sum_{i=1}^{k}(w_i \times X_i)\right) \tag{5}$$

where k is the number of donor catchment; and j is the time step.

In the OA method, the model parameters of the donor catchment were substituted into the SWAT model to simulate the runoff, and then the simulation results were combined according to their corresponding weights to estimate the runoff of the target catchment (Q_{2j}).

$$Q_{2j} = \sum_{i=1}^{k} w_i \times Q(j,\ X_i) \tag{6}$$

where X_i is the model parameters of the donor catchment, and w_i is the integrated weights of the spatial proximity and the physical similarity methods. The calculation method of w_i can be found in the literature [40].

4. Results

4.1. Runoff modeling and parameter sensitivity analysis of typical catchment

Eleven parameters related to runoff in the SWAT model were selected for calibration and sensitivity testing. The physical meaning and the original range of the parameters are summarized in Table 2. These parameters could be divided into four groups: parameters that control water movement between soil aquifers (ALPHA_BF, GW_DELAY, GWQMN, GW_REVAP, and REVAPMN), soil hydraulic characteristics (SOL_AWC, SOL_K, and ESCO), hydraulic channel parameters (CH_K2 and ALPHA_BNK) and the Soil Conservation Service (SCS) curve number (CN2). The *t*-stat and the *p*-value were used to represent the sensitivity of the model parameter. The higher the absolute value of the *t*-stat was and the lower the *p*-value was, the more sensitive the parameter was. Generally, the most sensitive parameters were CN2, ALPHA_BNK, ESCO, and GWQMN. The result was generally consistent with previous studies [41,42].

The runoff simulation accuracy of the SWAT model in the 38 catchments, indicated by NSE, R^2, and PBIAS, was quantitatively assessed on the monthly scale. To reduce the influence of the initial conditions of operation, the first year of the calibration period was used as the warm-up period of the model. The calibration and validation results of the SWAT model are shown in Fig. 2. During the simulation periods, the values of NSE and R^2 for all catchments were greater than 0.5, and the PBIAS values were less than 25%. The values of NSE and R^2 were not as high as expected during the calibration and validation periods, mainly because the constructed model was not perfect for some basins. Because the efficiency in the simulation period was acceptable, this part of the error was believed to be acceptable. The efficiency of the model in the validation period was usually inferior to that of the calibration period, because the model parameters were not adjusted to match the observed data during the validation period [5]. The SWAT model performed better in humid regions than in arid regions. For example, the 50th percentile values of NSE (R^2) in the humid and arid regions in the calibration period were 0.85 (0.87) and 0.78 (0.81), respectively. The runoff simulation in arid areas was still a challenge for hydrology [43]. The performance of the SWAT model in the simulation period indicated that the constructed model had good applicability in the study area, and its calibrated parameters were reliable for parameter regionalization.

Table 2
The parameters information of the SWAT model.

Parameter name	Parameter definition	Default range	*t*-stat*	*p*-value*	Sensitivity order
CN2	SCS runoff curve number	−0.2–0.2	12.25	0	1
SOL_AWC	Available water capacity of the soil layer	−0.2–0.4	3.29	0	6
SOL_K	Saturated hydraulic conductivity	−0.8–0.8	6.55	0	5
ALPHA_BF	Baseflow alpha factor	0–1	0.66	0.51	10
GW_DELAY	Groundwater delay time	0–500	1.98	0.05	8
GWQMN	Threshold depth of water in the shallow aquifer required for return flow to occur	0–5000	6.68	0	4
ESCO	Soil evaporation compensation factor	0–1	8.24	0	3
GW_REVAP	Groundwater re-evaporation coefficient	0–0.25	2.78	0.01	7
REVAPMN	Threshold depth of water in the shallow aquifer for re-evaporation to occur	0–600	0.51	0.61	11
CH_K2	Effective hydraulic conductivity in the main channel alluvium	0–500	1.85	0.06	9
ALPHA_BNK	Base flow alpha factor for bank storage	0–1	10.45	0	2

t-stat* and *p*-value* are the median of the *t*-stat and *p*-values of the 38 catchments.

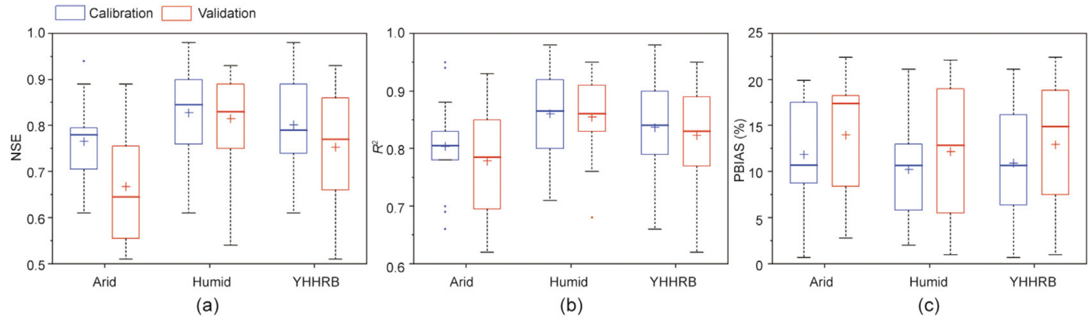

Fig. 2. The boxplot of (a) NSE, (b) R^2, and (c) PBIAS of calibration and validation periods. The boxes indicate the 25th and 75th percentiles; whiskers represent the lowest and highest value; the red and blue central lines indicate the 50th percentile; "+" represents the mean value.

4.2. Model parameters regionalization based on regression-based method

The degree of correlation between the model parameters and the catchment descriptors is illustrated in Fig. 3(a). The result indicated that CN2, SOL_K, ESCO, and GW_REVAP were correlated with multiple descriptors. Taking CN2 as an example, its absolute value of correlation coefficients with Slope, Clay, P, and φ were all greater than 0.5, indicating that these descriptors were relatively crucial to CN2. The sensitivity of REVAPMN, CH_K2, and ALPHA_BF in the calibration period was low (Table 2), and these parameters were difficult to obtain the optimal solution, resulting in the low correlation between the parameters and the catchment descriptors [40]. Although ALPHA_BNK and GWQMN were the sensitivity parameters of the model, the correlation coefficients between these parameters and catchment descriptors were low, mainly because the physical meaning of the model parameters had little correlation with the descriptors.

The heat map of the correlation coefficients among the nine catchment descriptors is plotted in Fig. 3(b). The result indicated that Area and Ele, T and Ele, P and NDVI, φ and NDVI, T and P, φ and P had strong correlations, and their absolute value of correlation coefficients were greater than 0.7, thus indicating the poor independence of the variables. The variance inflation factor (VIF) values of Ele, P, and T were greater than 10 (Fig. 4), indicating statistically significant multicollinearity between the catchment descriptors. Hence, the principal component analysis (PCA) method was used to reduce the dimensions of nine descriptors to solve the collinearity problem. Based on the principle that the cumulative variance contribution rate was greater than 85%, four principal components were selected, and the final variables were calculated according to the principal component coefficients.

The four principal component variables were identified as the input of six regression-based models, and the model was evaluated with the leave-one-out method. The correlation diagrams of the estimated and calibrated high sensitivity parameters are shown in Fig. 5, including CN2, SOL_AWC, SOL_K, GW_DELAY, ESCO, and GW_REVAP. The correlation coefficients between the estimated values of CN2 and SOL_K and the calibration values were greater than 0.5, indicating the high estimation accuracy. The remaining

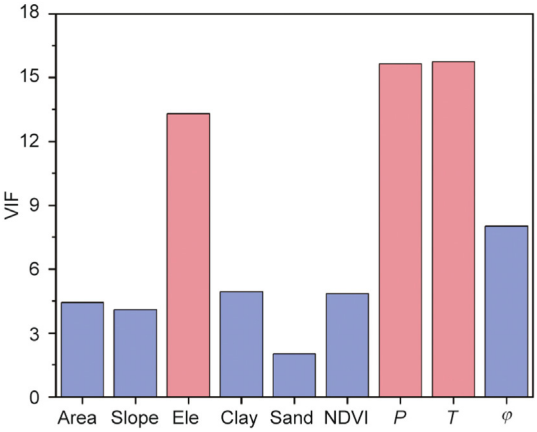

Fig. 4. The VIF test values among nine catchment descriptors.

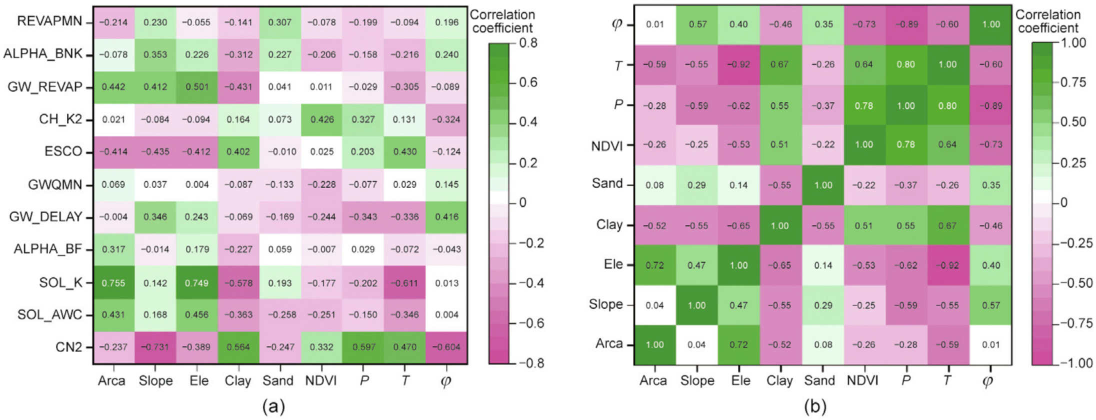

Fig. 3. The heat map of the correlation coefficients: (a) calibrated model parameters versus catchment descriptors; and (b) catchment descriptors versus catchment descriptors.

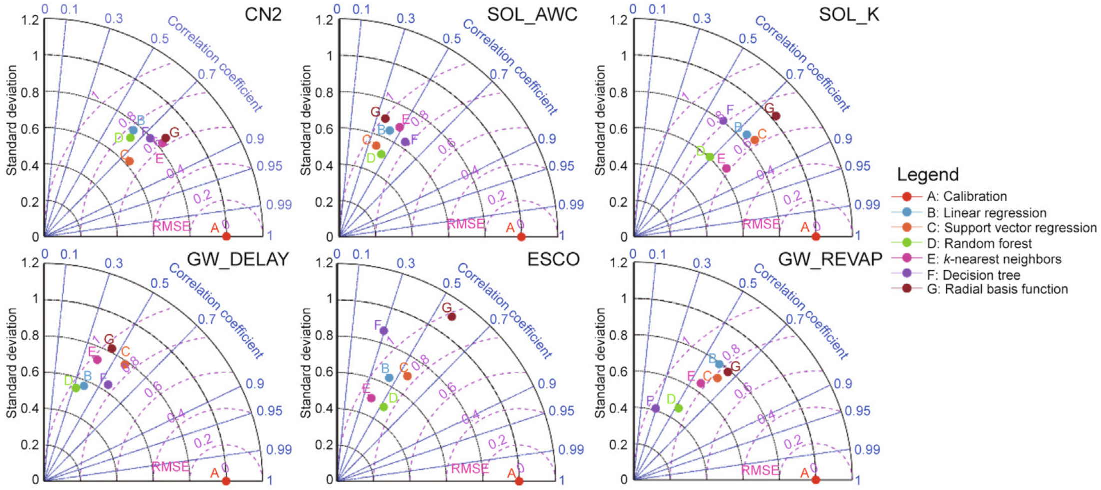

Fig. 5. Taylor diagrams of model parameters estimated by six regression-based models.

five model calibration parameters were difficult to estimate using the catchment descriptors due to their low correlation with the descriptors and low sensitivity. In comparing the estimation results of the six regression-based models, SVR performed better than the other models, and the estimation effect of DT was relatively poor. After the estimated values of the model parameters were obtained, which were input into the SWAT model of the ungauged catchment for runoff simulation.

4.3. Model parameters regionalization based on similarity-based method

The number of donor catchments directly affects the simulation accuracy of the target catchment for the similarity-based approach. Therefore, donor catchment numbers from 1 to 38 were tested, and the relationship between the number of donor catchments and the model evaluation criterion (NSE and PBIAS) was analyzed (Fig. 6). The results indicated that one donor catchment was the most suitable when the PA and the OA methods were adopted. For example, the 50th percentile values of NSE were the highest when one donor catchment was used, and the 50th percentile values of PBIAS were low. One donor catchment meant that the catchment closest to the target catchment was used. In this case, the results of the two methods were consistent. The number of donor catchments obtained was smaller than that from Bao et al. [40] and Oudin et al. [44] used. Compared with these studies, multiple nested catchments were used in this study (Fig. 1(a)), including four catchments within the Heishiguan basin, seven catchments within the Huaibin basin, and one catchment within the Xianyang basin. The hydrometeorological conditions in the nested catchments were similar, leading to the excellent performance of the given regionalization methodology. In order to investigate the suitable number of donor catchments after the nested catchments were excluded, numbers of donor catchments from 1 to 26 were tested. In this case, one donor catchment was the most suitable when the PA method was used. When the OA method was used, three donor catchments were the most suitable. The regionalization performance decreased significantly when the nested catchments were excluded.

4.4. Results of regionalization approaches

Based on the calibration results indicated by NSE and PBIAS, the runoff simulation accuracy of the assumed ungauged catchment under the regression-based schemes and similarity-based approach was compared (Fig. 7). In 38 catchments, the simulation

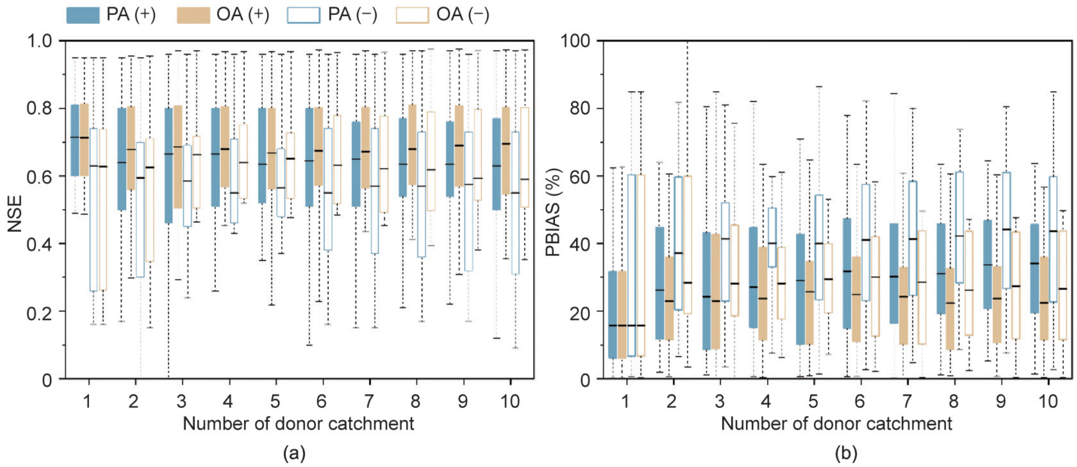

Fig. 6. Impact on model accuracy of the number of donor catchments used for ungauged catchment simulation, including PA and OA methods: (a) NSE and (b) PBIAS. "+" represents the nested catchments are contained; "−" represents nested catchments are excluded; the boxes indicate the 25th and 75th percentiles; whiskers represent the lowest and highest values; the black central lines represent the 50th percentile.

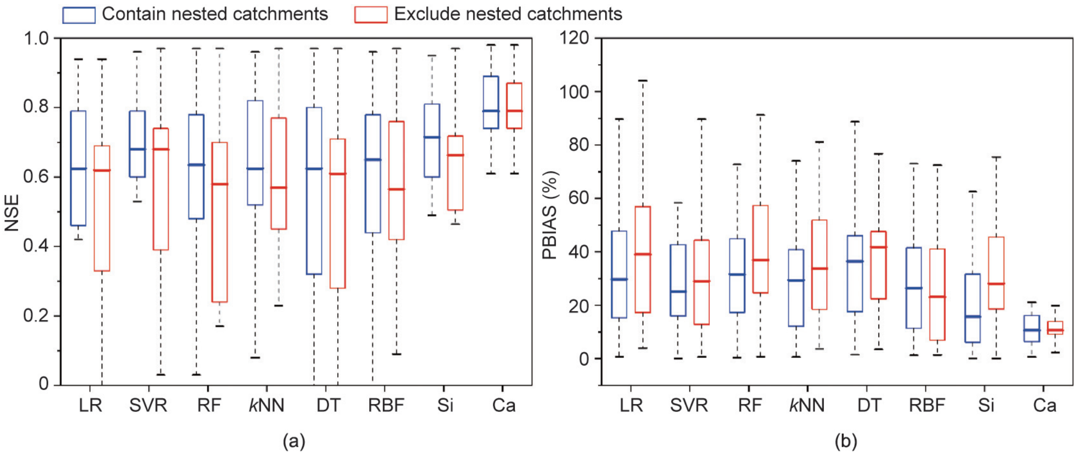

Fig. 7. Comparison of calibrated and regionalized results: (a) NSE and (b) PBIAS. Ca: calibration; Si: similarity-based. The boxes indicate the 25th and 75th percentiles; whiskers represent the lowest and highest values; and the red and blue central lines indicate the 50th percentile.

accuracy of SVR-based and RBF-based regionalization approaches was higher than that of the LR-based method. However, the runoff simulation result based on the similarity-based approach (Si) was more accurate than the six regression-based methods. The 50th percentile values of NSE and PBIAS of Si method were 0.71 and 12.5%, respectively, and their accuracy was significantly higher than that of the regression-based approaches. The reason for this phenomenon might have been the impact of the nested catchments. Therefore, the nested catchments in the study area were excluded, and the accuracy of different regionalization methods was compared in the remaining 26 catchments (Fig. 7). In this case, the 50th percentile values of NSE (PBIAS) were 0.68 (29%), 0.66 (28.15%), and 0.62 (39.05%) for the SVR, Si, and LR, respectively. The results indicated that the runoff simulation accuracy of the SVR-based approach was higher than that of Si and LR-based methods in the ungauged catchments.

According to Fig. 8, the most successful regression-based schemes were distributed differently in the 38 catchments. The LR-based method performed poorly, there were only 2 out of 38 catchments in which the NSE values were higher than that for other machine learning techniques when the LR-based method was used, and only 6 out of 38 catchments had the lowest PBIAS values. The number of the three best performing regionalization approaches on each catchment was investigated. The results showed that the performances of SVR, RBF, and *k*NN were significantly better than those of the other methods in the 38 catchments.

As presented in Fig. 9, all regionalization approaches performed better in humid regions than in arid regions. In humid regions, the 50th percentile values of NSE for different methods were all greater than 0.7. In contrast, the 50th percentile values of PBIAS varied in these regions. Generally, the regionalization effect order in humid regions from good to poor was *k*NN, Si, RF, SVR, RBF, DT, and LR. In arid regions, the 50th percentile values of NSE for different methodologies varied greatly. The 50th percentile values of PBIAS with the regression-based schemes were greater than 30%. Generally, the accuracy of the regionalization methods in arid regions, from high to low, was in the following order: Si, SVR, RF, RBF, *k*NN, LR, and DT.

5. Discussion

5.1. Application of machine learning algorithm in parameter regionalization

Given that catchment descriptors and model parameters are interdependent, and their relationship is complex and nonlinear. The machine learning technique is an interesting modeling structure for parameter regionalization, which can accurately capture

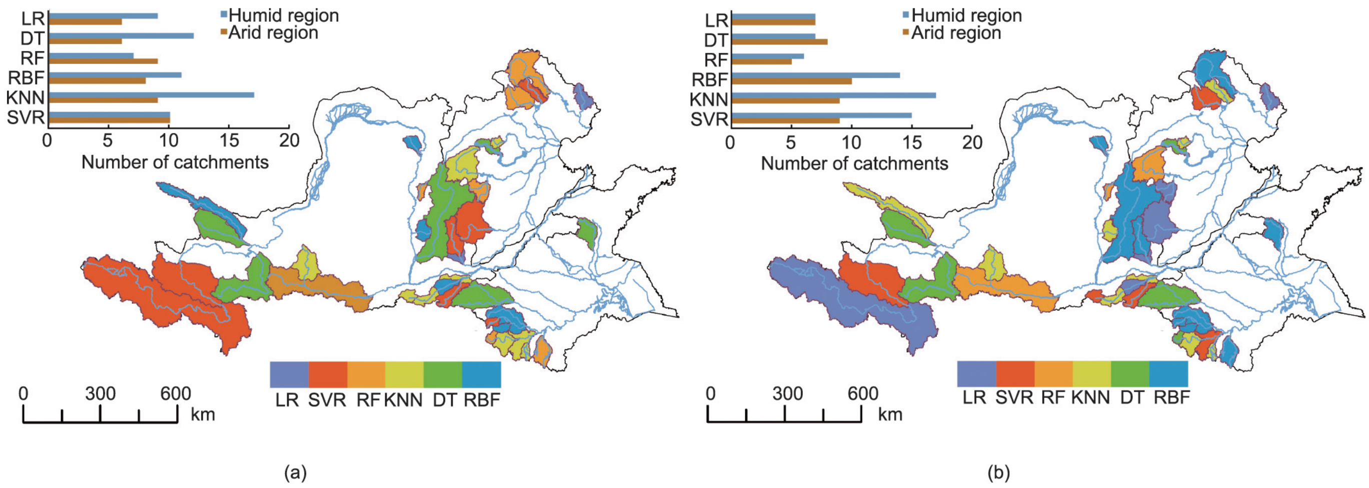

Fig. 8. Spatial distribution of the most successful regionalization methods: (a) method corresponding to the maximum NSE values; and (b) method corresponding to the minimum absolute values of PBIAS. The bar diagram represents the statistics of the three best performing regression-based regionalization methods on each catchment.

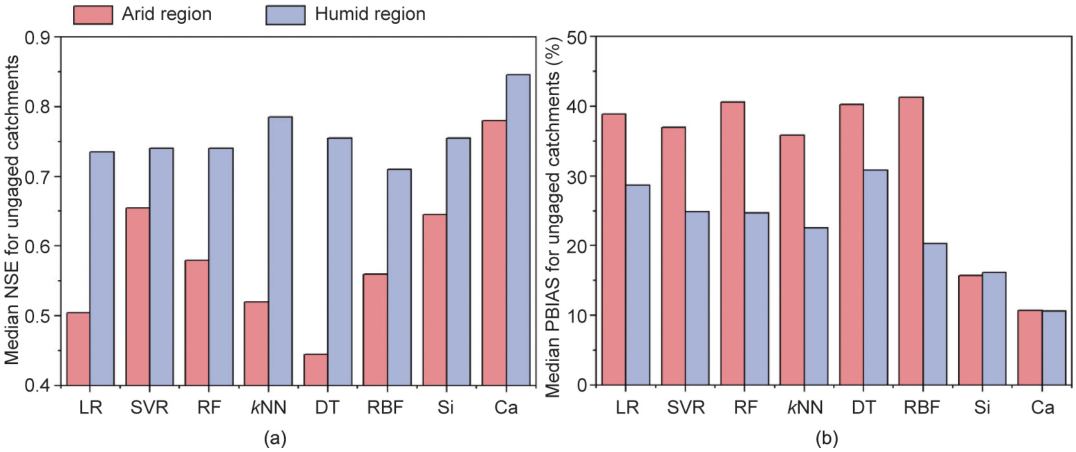

Fig. 9. Comparison of regionalized results estimated by calibration (Ca), similarity-based approaches (Si), and six regression-based approaches (LR, SVR, RF, DT, *k*NN, RBF) in two types of climatic zones: median (a) NSE and (b) PBIAS for ungaged catchments.

the intrinsic relationships between the input and output variables, regardless of their internal physical links. This might be a reliable and robust solution to the PUB issues. Booker and Snelder [45] and Golian et al. [8] also found that complex modeling techniques were superior to the linear method in predicting hydrologic properties. These technologies produced improved performances and a high degree of flexibility in capturing nonlinear and complex relationships between the model parameters and catchment descriptors [46,47]. Unlike the single machine learning algorithm used in previous research, the potential of multiple methods in regionalization application was compared in this study, and the result showed that the SVR-based method performed better than other algorithms. SVR sought to minimize structural risk in the modeling process, giving it additional generalization capabilities. Patel and Ramachandran [48] also found that SVR provided superior performance in modeling the discharge time series data. The performance of different machine learning algorithms varied significantly in different climate regions. Different data input might have had a greater impact on some model performance than the algorithm itself. It was difficult to determine whether the machine learning model was the best solution to all problems.

The parameter regionalization error of the regression-based schemes was larger than that of the similarity-based approach, because there were fewer training samples and the nonlinear relationship between the catchment descriptors and the model parameters was not fully learned. When the parameter regionalization results of the various methods were compared with the calibration results, the performances of the six regression-based schemes were significantly inferior to that of the calibration method. Regardless of the strength of the correlation between the model parameters and the catchment descriptors, with the use of the descriptors alone, estimating the model parameters would lead to the decline of parameter regionalization performance, indicating that there was still considerable room for improvement in the problem of model parameterization.

5.2. Donor catchment selection

As an important parameter regionalization methodology, the similarity-based approach involved the selection of donor catchment using the spatial distance and the physical similarity of the catchment. The number of donor catchments was related to the study area, basin density, and the approach (physical similarity or spatial proximity) used [1]. The regionalization performance was best when one donor catchment was selected in this study, whether the PA method or the OA method was adopted (Fig. 6). One reason for this was that the 38 catchments included multiple nested catchments. The similarity degree of the hydrometeorological conditions in the nested catchments was relatively high, resulting in a better performance of parameter regionalization. When the donor catchments did not contain nested catchments, the accuracy of the parameter regionalization decreased significantly. In addition to the impact of the nested catchments, the hydrologic and climatic conditions of some typical catchments had a large span, resulting in low similarity among catchments. For example, the Huangfu, Linjiaping, Daning, and Hejin catchments were significantly different from other catchments regardless of the spatial distance or physical similarity. In terms of the physical attributes, the existence of a gauged catchment adjacent to ungauged catchments was more important than the similarity between gauged and ungauged catchments [49].

Whether or not the study area contained nested catchments, the regionalization performance of the OA method was significantly better than that of the PA method (Fig. 6). The OA method was used to directly apply the model parameters from the donor catchment to the ungauged basin without modification, and the method involved the use of all information for the calibrated model parameters. However, the PA method was used to weigh and average the model parameters of the donor catchment, and then apply these parameters to the unmeasured catchment. There is strong interdependence among hydrological model parameters, which is weakened when the parameters are averaged [44]. Therefore, the PA method is commonly used when the correlation between the hydrological model parameters is small.

5.3. Descriptor importance in parameter regionalization

The application of regression-based schemes to parameter regionalization assumed that the selected catchment descriptors could describe the hydrological behavior of a basin well. Therefore, the selection of descriptors is crucial to the success of parameter regionalization. Although the selection of suitable catchment descriptors in hydrological parameterization studies has been widely discussed, no universally accepted selection criteria exist [14,50]. Merz and Blöschl [51] mentioned that the selected catchment descriptors should be the influence factors that can drive the watershed hydrological response. Mwakalila [52] proposed that the catchment descriptors should have both geographical and parameter spatial significance. In previous studies, the selection of catchment descriptors was mainly based on geography, meteorology, hydrology, and soil [53]. Other descriptors have occasionally been used, such as land use, drainage density [49], and other meteorological data (mean annual evaporation) [47]. In addition, the selection of appropriate catchment descriptors also depends on the physical significance of the model parameters. For example, the CN2 in the SWAT model depends on the soil and land use characteristics of a catchment [18], and the catchment descriptor in this respect should be considered. The regional climate, soil types, and vegetation make regionalization special, and the SWAT model parameters are mostly related to these factors [49]. The nine catchment descriptors selected in this study covered these factors, but statistically significant multicollinearity problems existed among the descriptors. Saadi et al. [23] only retained the catchment descriptors with low correlation values to each other. Penas et al. [47] selected predictor variables based on the combination of scatter plots (hydrological indices versus environmental variables) and parametric correlations. The PCA method used in this study reduced the dimensions of the catchment descriptors. With the reduction of descriptors dimensions, the information regarding the original variables was retained to the greatest extent.

5.4. Influence of hydrologic and climatic conditions on regionalization

The study area included two climatic regions, namely, a relatively humid region and a relatively arid region. The humid areas were mainly located in the Huai River Basin, where the mean annual precipitation was 976.23 mm (Table 1), and the mean value of NSE in the calibration period was 0.82 (Fig. 2). The arid regions were distributed in the Yellow River Basin and the Hai River Basin, where the mean annual precipitation was 561.01 mm (Table 1), and the mean value of NSE in the calibration period was 0.77 (Fig. 2). Given the influence of the precipitation distribution, the runoff simulation accuracy of the SWAT model in the humid regions was better than that in the arid regions. According to the results of the parameter regionalization (Fig. 9), the regionalization performance in the humid areas was better than that in the arid areas, which was consistent with the runoff simulation results. Therefore, the hydrological model parameters regionalization results largely depended on the accuracy of the runoff simulation in the donor catchment. Only when the parameters with sufficient accuracy were obtained, could the simulation results in the ungauged catchments be obtained with parameter regionalization. In arid regions, the economy was usually undeveloped, the monitoring stations were few, and the hydrological data were relatively scarce. Moreover, the runoff simulation was more sensitive to the model parameters in these regions than in the humid regions. For the same parameter error, the deviation of the simulation results in arid regions was greater than that in humid regions [40]. Parajka et al. [54] and Yang et al. [55] also pointed out the impact of climate conditions on regionalization performance.

The most successful regionalization methodology in humid catchments may be differ from those in arid catchments. The performance of a particular approach varies between different studies more often than between methods tested in a single study [53,56]. As shown in Fig. 9, the SVR showed better regionalization performance in arid areas, while the *k*NN had higher regionalization accuracy in humid areas. Different data inputs may have a greater impact on the performance of some models than the algorithm itself, and determining the machine learning model to use as the best solution to the problem was difficult [57]. To improve the accuracy of parameter regionalization, the next challenge to consider is the introduction of more machine learning techniques.

5.5. Uncertainty and limitation of the results

The uncertainty of this study came from three aspects. First, the selected catchment descriptors imposed limitations on the interpretation of some ungauged catchments. This also illustrated a fundamental challenge to the parameter regionalization, that is, the number or quality of selected catchment descriptors was insufficient to represent the catchment heterogeneity [58]. Second, the nonlinear relationship between the model parameters and the catchment descriptors is difficult to express perfectly with statistical models. Third, the SWAT model has an excellent performance in the runoff simulation, but uncertainty still exists [59,60], which due to the following reasons: ① parameter uncertainty, in which the inconsistency of the model inputs and parameters in space and time leads to error in model parameter values; ② data uncertainty, in which the variability of natural conditions, limitation of measurement conditions, and the uncertainty of measurement methods all affect the accuracy of the model input data; and ③ the model uncertainty, in which the hydrological models generalize hydrological processes and cannot accurately represent the actual physical process of a watershed. Additionally, most of the parameters in the SWAT model adopt the default values, which deviate from the actual values, which also affects the accuracy of the model simulation.

6. Conclusions

The regionalization approach is a crucial method for solving the problem of runoff modeling in ungauged catchments. Different regionalization methods were used to estimate SWAT model parameters in this study, and runoff simulation was studied in 38 catchments located in the YHHRB. Due to the weakness of the LR-based method in coping with nonlinear relationships, five machine learning algorithms (SVR, RF, *k*NN, DT, and RBF) were used to describe the quantitative relationships between the model parameters and the catchment descriptors to improve the parameter regionalization performance. We found that the SVR-based regression scheme had the highest simulation accuracy in ungauged catchments, indicating that its performance was better than traditional LR-based and similarity-based approaches. The performance of different regionalization methods was similar in humid regions due to the relatively simple hydrometeorological processes and easy runoff simulation. However, the runoff simulation results in arid areas were more sensitive to the model parameters, and the advantages of the machine learning techniques were outstanding in these regions. The regionalization performance of the SVR, RBF, and RF based methods was better than that of the traditional LR techniques in arid regions. When the study area contained nested catchments, the best parameter regionalization performance was derived through similarity-based methods because of the high basin density and similarity among catchments. The study results enrich the method of parameter regionalization and provide a reference for future water resources planning and management in ungauged catchments.

Acknowledgments

This research is funded by the National Key Research and Development Program of China (2017YFA0605002, 2017YFA0605004, and 2016YFA0601501); the National Natural Science Foundation of China (41961124007, 51779145, and 41830863); and “Six top talents” in Jiangsu Province (RJFW-031).

Compliance with ethics guidelines

Houfa Wu, Jianyun Zhang, Zhenxin Bao, Guoqing Wang, Wensheng Wang, Yanqing Yang, and Jie Wang declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Guo Y, Zhang Y, Zhang L, Wang Z. Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: a comprehensive review. Wiley Interdiscip Rev Water 2021;8(1):e1487.

[2] Yang Q, Almendinger JE, Zhang X, Huang M, Chen X, Leng G, et al. Enhancing SWAT simulation of forest ecosystems for water resource assessment: a case study in the St. Croix River basin. Ecol Eng 2018;120:422–31.

[3] Beven KJ, Kirkby MJ, Freer JE, Lamb R. A history of TOPMODEL. Hydrol Earth Syst Sci 2021;25(2):527–49.

[4] Gong J, Yao C, Li Z, Chen Y, Huang Y, Tong B. Improving the flood forecasting capability of the Xinanjiang model for small- and medium-sized ungauged catchments in South China. Nat Hazards 2021;106(3):2077–109.

[5] Woo SY, Kim SJ, Lee JW, Kim SH, Kim YW. Evaluating the impact of interbasin water transfer on water quality in the recipient river basin with SWAT. Sci Total Environ 2021;776:145984.

[6] Clark GE, Ahn KH, Palmer RN. Assessing a regression-based regionalization approach to ungauged sites with various hydrologic models in a forested catchment in the northeastern United States. J Hydrol Eng 2017;22(12):05017027.

[7] Wang GQ, Zhang JY, Jin JL, Liu YL, He RM, Bao ZX, et al. Regional calibration of a water balance model for estimating stream flow in ungauged areas of the Yellow River Basin. Quat Int 2014;336:65–72.

[8] Golian S, Murphy C, Meresa H. Regionalization of hydrological models for flow estimation in ungauged catchments in Ireland. J Hydrol Reg Stud 2021;36:100859.

[9] Samuel J, Coulibaly P, Metcalfe RA. Estimation of continuous streamflow in Ontario ungauged basins: comparison of regionalization methods. J Hydrol Eng 2011;16(5):447–59.

[10] Knight RR, Gain WS, Wolfe WJ. Modelling ecological flow regime: an example from the Tennessee and Cumberland River basins. Ecohydrology 2012;5(5):613–27.

[11] Yang X, Magnusson J, Xu CY. Transferability of regionalization methods under changing climate. J Hydrol 2019;568:67–81.

[12] Beck HE, van Dijk AIJM, de Roo A, Miralles DG, McVicar TR, Schellekens J, et al. Global-scale regionalization of hydrologic model parameters. Water Resour Res 2016;52(5):3599–622.

[13] Boughton W, Chiew F. Estimating runoff in ungauged catchments from rainfall, PET and the AWBM model. Environ Model Softw 2007;22(4):476–87.

[14] Jafarzadegan K, Merwade V, Moradkhani H. Combining clustering and classification for the regionalization of environmental model parameters: application to floodplain mapping in data-scarce regions. Environ Modell Softw 2020;125:104613.

[15] Oudin L, Kay A, Andreassian V, Perrin C. Are seemingly physically similar catchments truly hydrologically similar? Water Resour Res 2010;46(11): W11558.

[16] Guiamel IA, Lee HS. Watershed modelling of the Mindanao River Basin in the Philippines using the SWAT for water resource management. Civ Eng J 2020;6(4):626–48.

[17] Reichl JPC, Western AW, McIntyre NR, Chiew FHS. Optimization of a similarity measure for estimating ungauged streamflow. Water Resour Res 2009;45(10): W10423.

[18] Sellami H, La Jeunesse I, Benabdallah S, Baghdadi N, Vanclooster M. Uncertainty analysis in model parameters regionalization: a case study involving the SWAT model in Mediterranean catchments (Southern France). Hydrol Earth Syst Sci 2014;18(6):2393–413.

[19] Ly S, Charles C, Degre A. Different methods for spatial interpolation of rainfall data for operational hydrology and hydrological modeling at watershed scale: a review. Biotechnol Agron Soc 2013;17(2):392–406.

[20] Heng S, Suetsugi T. Comparison of regionalization approaches in parameterizing sediment rating curve in ungauged catchments for subsequent instantaneous sediment yield prediction. J Hydrol 2014;512:240–53.

[21] Kittel CMM, Arildsen AL, Dybkjær S, Hansen ER, Linde I, Slott E, et al. Informing hydrological models of poorly gauged river catchments—a parameter regionalization and calibration approach. J Hydrol 2020;587:124999.

[22] Zhang YQ, Chiew FHS. Relative merits of different methods for runoff predictions in ungauged catchments. Water Resour Res 2009;45(7):W07412.

[23] Saadi M, Oudin L, Ribstein P. Random forest ability in regionalizing hourly hydrological model parameters. Water 2019;11(8):1540.

[24] Soni P, Tripathi S, Srivastava R. A comparison of regionalization methods in monsoon dominated tropical river basins. J Water Clim Chang 2021;12(5):1975–96.

[25] Lary DJ, Alavi AH, Gandomi AH, Walker AL. Machine learning in geosciences and remote sensing. Geosci Front 2016;7(1):3–10.

[26] Hao S, Ma Q, Zhai X, Lyu G, Fan S, Wang W. A new machine learning approach for parameter regionalization of flash flood modelling in Henan Province, China. In: Stanciu S, Kassmi K, Shmavonyan G, editors. Proceedings of the 2021 2nd International Conference on Energy, Power and Environmental System Engineering; 2021 Jul 4–5; Shanghai, China. Les Ulis: EDP Science; 2021. p. 02010.

[27] Ragettli S, Zhou J, Wang H, Liu C, Guo L. Modeling flash floods in ungauged mountain catchments of China: a decision tree learning approach for parameter regionalization. J Hydrol 2017;555:330–46.

[28] Ministry of Water Resources People's Republic of China. China water resources bulletin 2019. Beijing: China Water & Power Press; 2020.

[29] Wu J, Gao XJ. A gridded daily observation dataset over China region and comparison with the other datasets. Chin J Geophys 2013;56(4):1102–11. Chinese.

[30] Samal DR, Gedam S. Assessing the impacts of land use and land cover change on water resources in the Upper Bhima River Basin, India. Environ Chall 2021;5:100251.

[31] Tan ML, Gassman PW, Yang X, Haywood J. A review of SWAT applications, performance and future needs for simulation of hydro-climatic extremes. Adv Water Resour 2020;143:103662.

[32] Arnold JG, Moriasi DN, Gassman PW, Abbaspour KC, White MJ, Srinivasan R, et al. SWAT: model use, calibration, and validation. Trans ASABE 2012;55(4):1491–508.

[33] Li C, Fang H. Assessment of climate change impacts on the streamflow for the Mun River in the Mekong Basin, Southeast Asia: using SWAT model. Catena 2021;201:105199.

[34] Mohammadi B, Mehdizadeh S. Modeling daily reference evapotranspiration via a novel approach based on support vector regression coupled with whale optimization algorithm. Agric Water Manage 2020;237:106145.

[35] Park SY, Park M, Lee WY, Lee CY, Kim JH, Lee S, et al. Machine learning-based prediction of Sasang constitution types using comprehensive clinical information and identification of key features for diagnosis. Integr Med Res 2021;10(3):100668.

[36] Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D. Machine learning in agriculture: a review. Sensors 2018;18(8):2674.

[37] Feng K, González A, Casero M. A *k*NN algorithm for locating and quantifying stiffness loss in a bridge from the forced vibration due to a truck crossing at low speed. Mech Syst Signal Proc 2021;154:107599.

[38] Mary AH, Miry AH, Kara T, Miry MH. Nonlinear state feedback controller combined with RBF for nonlinear underactuated overhead crane system. J Eng Res 2021;9(3A):197–208.

[39] Hu Z, Chen X, Zhou Q, Chen D, Li J. DISO: a rethink of Taylor diagram. Int J Climatol 2019;39(5):2825–32.

[40] Bao Z, Zhang J, Liu J, Fu G, Wang G, He R, et al. Comparison of regionalization approaches based on regression and similarity for predictions in ungauged catchments under multiple hydro-climatic conditions. J Hydrol 2012;466-467:37–46.

[41] Ligaray M, Kim H, Sthiannopkao S, Lee S, Cho KH, Kim JH. Assessment on hydrologic response by climate change in the Chao Phraya River Basin, Thailand. Water 2015;7(12):6892–909.

[42] Yu D, Xie P, Dong X, Hu X, Liu J, Li Y, et al. Improvement of the SWAT model for event-based flood simulation on a sub-daily timescale. Hydrol Earth Syst Sci 2018;22(9):5001–19.

[43] Samimi M, Mirchi A, Moriasi D, Ahn S, Alian S, Taghvaeian S, et al. Modeling arid/semi-arid irrigated agricultural watersheds with SWAT: applications, challenges, and solution strategies. J Hydrol 2020;590:125418.

[44] Oudin L, Andreassian V, Perrin C, Michel C, Le Moine N. Spatial proximity, physical similarity, regression and ungaged catchments: a comparison of regionalization approaches based on 913 French catchments. Water Resour Res 2008;44:W03413.

[45] Booker DJ, Snelder TH. Comparing methods for estimating flow duration curves at ungauged sites. J Hydrol 2012;434–435:78–94.

[46] Elith J, Graham CH, Anderson RP, Dudík M, Ferrier S, Guisan A, et al. Novel methods improve prediction of species' distributions from occurrence data. Ecography 2006;29(2):129–51.

[47] Penas FJ, Barquin J, Alvarez C. A comparison of modeling techniques to predict hydrological indices in ungauged rivers. Limnetica 2018;37(1):145–58.

[48] Patel SS, Ramachandran P. A comparison of machine learning techniques for modeling river flow time series: the case of Upper Cauvery River Basin. Water Resour Manage 2015;29(2):589–602.

[49] Swain JB, Patra KC. Streamflow estimation in ungauged catchments using regionalization techniques. J Hydrol 2017;554:420–33.

[50] Boscarello L, Ravazzani G, Cislaghi A, Mancini M. Regionalization of flow-duration curves through catchment classification with streamflow signatures and physiographic-climate indices. J Hydrol Eng 2016;21(3):05015027.

[51] Merz R, Blöschl G. Regionalisation of catchment model parameters. J Hydrol 2004;287(1–4):95–123.

[52] Mwakalila S. Estimation of stream flows of ungauged catchments for river basin management. Phys Chem Earth 2003;28(20–27):935–42.

[53] Razavi T, Coulibaly P. Streamflow prediction in ungauged basins. Review of regionalization methods. J Hydrol Eng 2013;18(8):958–75.

[54] Parajka J, Viglione A, Rogger M, Salinas JL, Sivapalan M, Blöschl G. Comparative assessment of predictions in ungauged basins-part 1: runoff-hydrograph studies. Hydrol Earth Syst Sci 2013;17(5):1783–95.

[55] Yang X, Magnusson J, Huang S, Beldring S, Xu CY. Dependence of regionalization methods on the complexity of hydrological models in multiple climatic regions. J Hydrol 2020;582:124357.

[56] Pool S, Vis M, Seibert J. Regionalization for ungauged catchments—lessons learned from a comparative large-sample study. Water Resour Res 2021;57(10):WR030437.

[57] Abdulelah Al-Sudani Z, Salih SQ, Sharafati A, Yaseen ZM. Development of multivariate adaptive regression spline integrated with differential evolution model for streamflow simulation. J Hydrol 2019;573:1–12.

[58] Choubin B, Solaimani K, Rezanezhad F, Roshan MH, Malekian A, Shamshirband S. Streamflow regionalization using a similarity approach in ungauged basins: application of the geo-environmental signatures in the Karkheh River Basin. Catena 2019;182:104128.

[59] Abbas T, Hussain F, Nabi G, Boota MW, Wu RS. Uncertainty evaluation of SWAT model for snowmelt runoff in a Himalayan watershed. Terr Atmos Ocean Sci 2019;30(2):265–79.

[60] Wang Y, Jiang R, Xie J, Zhao Y, Yan D, Yang S. Soil and water assessment tool (SWAT) model: a systemic review. J Coast Res 2019;93(SI):22–30.