A New Machine Learning Approach for parameter regionalization of Flash Flood Modelling in Henan Province, China

Sijia Hao[1,2], *Qiang* Ma[2,*], *Xiaoyan* Zhai[2], *Guomin* Lyu[2], *Suqi* Fan[1,2], *Wenchuan* Wang[1], and *Changjun* Liu[2]

[1]School of Water Resources, North China University of Water Resources and Electric Power, Zhengzhou, China
[2]China Institute of Water Resources and Hydropower Research, Beijing, China

Abstract. China is one of the countries in the world that seriously affected by flash floods disasters. The flash flood caused by extreme rainfall occurred at mountainous small-sized watersheds in China often leads to serious economic damages and obstructs the social development. Setting up an efficient forecasting system for flash flood has been widely accepted as one of the key non-structural measures to improve the control and prevention capability of China. However, due to the data limitation, establishing forecast models in those flash flood areas is challenged by the lack of parameter references. This paper proposed a new machine learning approach based on the Random Forest (RF) algorithm for model parameter regionalization. Integrated with distributed deterministic hydrological models of 20 small-sized watersheds in Henan province, the RF algorithm has been applied for defining the watersheds' similarity and further transferring the parameters from sample watersheds to the objective watershed. Validated through leave-one-out approach, the RF model is able to effectively improve the simulation accuracy of flash floods in Henan province. The presented approach showed high-levelled applicability to be extended in other flash flood areas in China for providing effective reference for parameter regionalization.

1 Introduction

Flash floods in small-sized mountainous watersheds cause serious damage to life and property in China. From 1991 to 2015, the number of deaths due to flash flood disasters was around 27,000, with an average annual death rate nearly 1093 per year. The average annual economic loss caused by flash flood disasters exceeded RMB 40 billion [1]. Improving the accuracy of flood forecasting is the key point to solve the problem of flood disaster reduction. However, it is difficult to determine the parameters of the hydrological model for the ungauged watersheds. Therefore, the study the parameter regionalization of flash flood modelling based on machine learning has scientific significance and practical value.

[*]Corresponding author: maqiang@iwhr.com

Some researches claimed that the transferring the known parameters from a gauged watershed to the ungauged watershed which is hydrologically similar watershed could be one of the best opinions to modelling problems in ungauged area. Obviously, the core of that approach is to define the similarity between referenced watershed and the targeted watershed. Under the current condition, with the development of big data and artificial intelligence analysis technology, the machine learning technology has been widely considered as one of the main approaches to implement the regionalized analysis of model parameters. In 2014, Singh et al. used classification and regression tree (CART) analysis to determine the relationship between catchment similarity and performance of transferred parameters. Using physical and climatic catchment characteristics, as well as streamflow response characteristics, similarity is defined for different geographic regions [2]. In 2017, Ragettli et al. used decision tree learning to explore parameter set transferability in the full space of catchment descriptors and proved that decision tree learning can outperform other regionalization approaches because it generates rules that optimally consider spatial proximity and physical similarity [3].

In this paper, a new machine learning approach integrated with a distributed deterministic hydrological model has been implement in 20 flash flood watersheds in Henan province. Through the leave-one-out approach, the rules defined by the RF model for parameter regionalization has been validated. The modelling results of objected watershed with parameters transferred from RF selected watersheds showed with accepted accuracy. The approach presented in this paper expressed higher applicability to be extended to other flash flood watershed with limited data condition.

2 Study area and data

2.1 Study area

Henan Province is located in the central-eastern part of China. Mountainous areas are mainly concentrated in the southwest, accounting for 26.6% of the province's area. The precipitation is mainly concentrated in summer, accounting for about 45-60% of the annual precipitation. This paper selects 20 small watersheds in Henan Province as the research area, and the distribution of each watershed is shown in Figure 1.

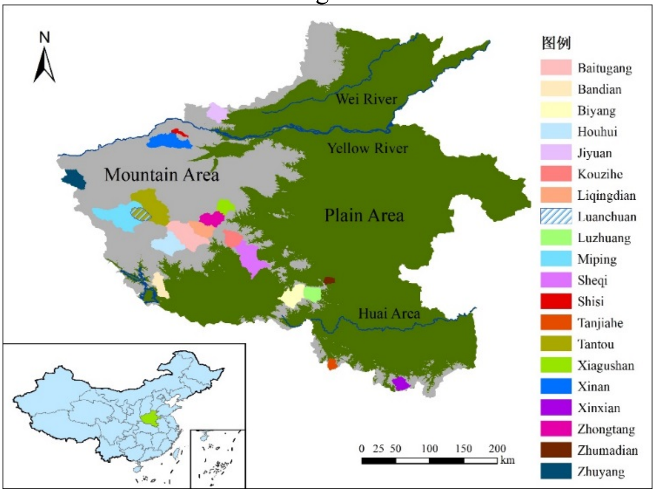

Fig. 1. Study area.

2.2 Data

The establishment of machine learning model needs both feature sets and label sets. In this research, the feature sets are divided into five categories: Topography, drainage, land use, climate attributes and soil texture. The specific contents and source of each attribute category is shown in table 1.

Table 1. Categories of watershed characteristics.

Attribute Category	Characteristics	Source
Topography and landform	Watershed area (km^2); sub-watershed average area (km^2); river length (km); average altitude (m); minimum altitude (m); maximum altitude (m); average slope (°); proportion of flat slope area (%), and gentle slope area (%)	China Centre for Resources Satellite Data and Application
Drainage	River network density (-); first class river percentage (%); second class river percentage (%); third class river percentage (%); and river class (-)	China Centre for Resources Satellite Data and Application
Land use	The proportion of grassland area (%); the proportion of water area (%); the proportion of cultivated land (%); the proportion of construction land (%); the proportion of woodland area (%); the proportion of swamp area (%); and the proportion of other types of land area (%)	National Geomatics Center of China
Climate	Annual average rainfall (mm), one hour maximum rainfall (mm); 24-hour maximum rainfall (mm);	Henan Hydrological Bureau
Soil texture	Proportion of sand and clay area (%); proportion of sandy loam area (%); proportion of soil and clay area (%); proportion of clay area (%); proportion of silt loam area (%); proportion of clay loam area (%); proportion of loam area (%); proportion of sand and clay area (%); and proportion of silt clay loam area (%)	Field survey
Topological Association	Distance between watershed centre points (km)	China Centre for Resources Satellite Data and Application

The Nash coefficient calculated from the model results of the different watershed with parameters values obtained from itself and transferred from other watersheds were defined as label sets. Figure 2 showed the validated modelling results with the deterministic distributed hydrological of 20 watersheds in Henan province. It can be seen that the average Nash coefficient of all watersheds simulated were higher than 0.7. Therefore, the modelling tool we selected in this study is suitable for simulating the flash floods in small-sized watersheds in Henan province.

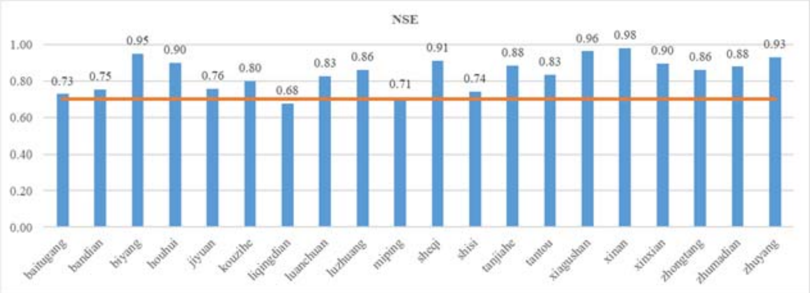

Fig. 2. Evaluation of simulation results.

3 Methodology

3.1 Study design

According to the attribute data of selected watersheds, the correlation coefficients among different properties were firstly calculated. If the value of correlation coefficient is higher than 0.9, those two properties could be considered as strong related and further to keep only of those two in the analysis as the main representative characteristics of the watershed. After determining the main characteristics of the watershed, the difference between the main characteristics of two watersheds were used for detecting the similarity between two tested watersheds through machine learning approach. In order to launch the machine learning algorithm model, the Nash coefficient calculated from the model of targeted watershed with parameters defined with its own property and transferred from other referenced watersheds were used as the label data. In this study, the training samples set is obtained from the data of 20 watersheds consisted with 380 samples (20 target watersheds and 19 donor watersheds). The Random Forest (RF) machine learning model was programed through Python language. The similarity metrics consisted with main characteristic of watersheds is one of the main inputs to launch the model. And the similar watershed discrimination criteria will be automatically calculated by the RF model as the main output. The model is verified by the leave-one-out approach on 20 watersheds in Henan province. According to the model outputs and the feature importance ranking list, the parameter regionalization plan of the small watershed in Henan Province is finally obtained.

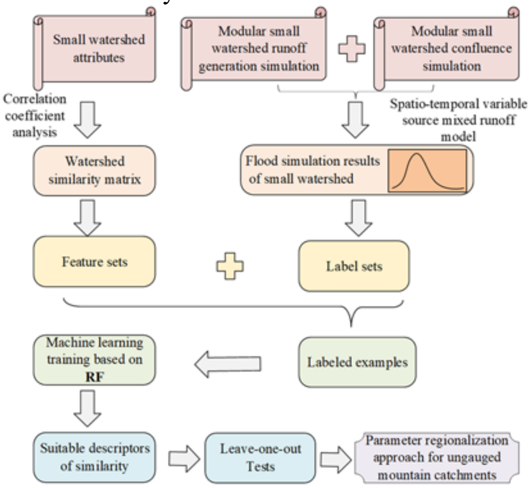
Small watershed attributes
Modular small watershed runoff generation simulation
Modular small watershed confluence simulation
Correlation coefficient analysis
Spatio-temporal variable source mixed runoff model
Watershed similarity matrix
Flood simulation results of small watershed
Feature sets
Label sets
Machine learning training based on RF
Labeled examples
Suitable descriptors of similarity
Leave-one-out Tests
Parameter regionalization approach for ungauged mountain catchments

Fig. 3. Flowchart of parameter regionalization based on RF.

3.2 Random forest

3.2.1 Random forest principle (RF)

The RF regression model is established on the basis of the CART regression tree model. It mainly uses random sampling to extract samples, to build multiple decision trees, to combine the prediction results of all decision trees, and to get the final result by voting [4]. The RF regression principle is shown in the figure 4. The RF method has a good tolerance for outliers and noise and is not prone to overfitting. Compared to the CART tree method, RF adopts the processing characteristics of non-pruning and arbitrary growth of a single tree to obtain a low-bias decision tree, and can ensure the correct rate of classification of new test data. RF has a very good effect on solving multivariate predictions, with strong data mining capabilities and high prediction accuracy.

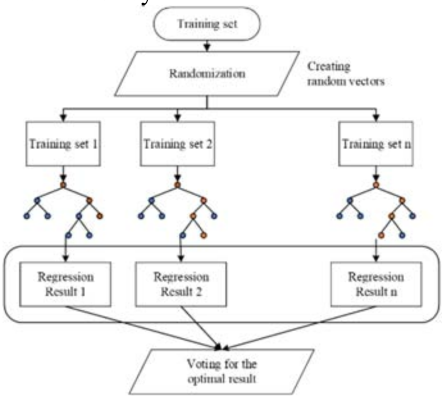

Fig. 4. RF regression principle.

3.2.2 CART regression tree

The CART model generates regression trees based on the square error minimization criterion [5]. Suppose the given training data set is:

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\} \quad (1)$$

In the formula, X and Y are input and output variables respectively.

Assuming that the input space has been divided into M units: R_1, R_2,...,R_M, and each unit R_m has a fixed output value c_m, therefore, the regression tree model can be expressed as:

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m) \quad (2)$$

Using a heuristic method, select the j_{th} variable x^j and its values as the segmentation variable and segmentation point, and thus define two regions:

$$R_1(j, s) = \{x|x^j \leq s\}\, and R_2(j, s) = \{x|x^j > s\} \quad (3)$$

Then by solving:

$$min_{(j,s)}\left[min_{(c_1)} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + min_{(c_2)} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2\right] \quad (4)$$

For the fixed input variable j, the optimal cut points can be found:

$$\hat{c}_1 = \mathrm{ave}\{y_i|x_i \in R_1(j, s)\} \text{ 和 } \hat{c}_2 = \mathrm{ave}\{y_i|x_i \in R_2(j, s)\} \quad (5)$$

Traverse all input variables and find the optimal segmentation variable j to form a pair (j, s). The input space is divided into two regions in turn.

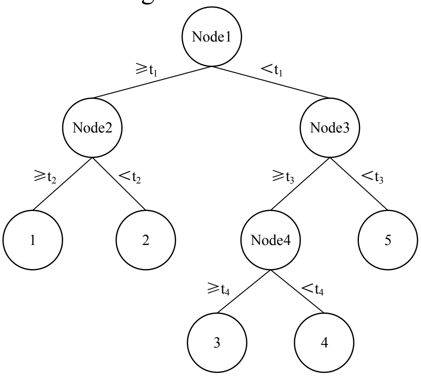

Fig. 5. Growth processes of a single CART regression tree.

3.2.3 Leave-one-out method

The Leave-one-out method was used to verify the results of parameter transplantation of the RF model in ungauged areas. After removing the target watershed in the sample set, the remaining watersheds are used as the input data sets of the training sample to construct the RF model. According to the prediction results, five watersheds with the highest NSE were selected as the similar watersheds and compared with the actual simulation results of the target watershed, which proves the possibility of applying the parameter transplantation standards generated by the RF model.

3.3 Modelling tools

In this study, the spatiotemporal variable source mixed runoff (SVSMR) model proposed by China Institute of Water Resources and Hydropower Research (IWHR) was used to simulate flash floods in 20 small watersheds in Henan Province. The model was built based on the modular modeling method, using different runoff generate mechanism for different hydro-geomorphological response unit. According to the characteristics of the short duration of the flash flood, this model can operate on daily and hourly time scales. Using GIS and remote sensing technology, the model identified the landform features (terrain, land use, vegetation cover, soil type) by hillside scale and established the corresponding relationship between the hydro-geomorphological response unit and the runoff generation mechanism. Using the modular hydrological model construction method, the calculation mode of runoff generation for sub-hydrological unit was constructed based on the dominant runoff characteristics of the hillside geomorphic hydrological response unit. According to the sub-hydrological unit topological series method, each sub-basin production stream was collected in time series to the basin outlet using the slope and river convergence procedure. The slope and river convergence process were calculated using the kinematic wave method [6-7].

In order to evaluate the quality of model simulation with regionalized parameter proposed by RF rules, the Nash coefficient (NSE) was applied in this study:

$$NSE = 1 - \frac{\sum_{i=1}^{N}\left(Q_o^i - Q_s^i\right)^2}{\sum_{i=1}^{N}\left(Q_o^i - \overline{Q_o}\right)^2} \tag{6}$$

where, Q_o is the observed flow, Q_s is the simulated flow, $\overline{Q_o}$ is the observed average flow, and i is the simulation time steps.

4 Results and discussion

4.1 RF result

By using the RF algorithm model the feature importance ranking is defined in Figure 6. The top three are sandy loam, silt loam, and sandy clay, which respectively accounts for 17%, 13% and 7% of the importance among all components. The maximum precipitation in one hour and in 24 hours both accounts for 6% and listed at fourth and fifth in the importance ranking. There are 22 indicators with a cumulative statistical importance of 95%. The results show that the descriptors of similarity for small mountain watersheds in Henan Province are underlying surface conditions and climate.

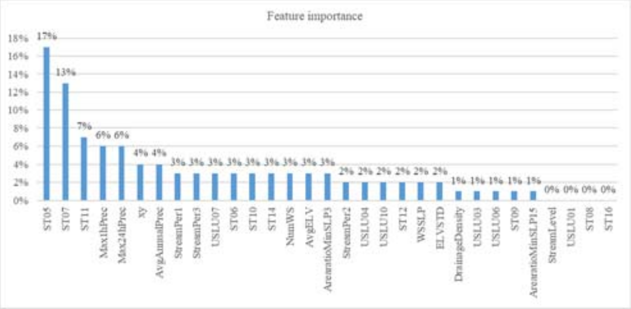

Fig. 6. Index importance degree and contribution of the index.

4.2 Regionalization performance

To verify the RF model performance, the leave-one-out test was implemented for 20 watersheds. Taking Baitugang as an example, the RF regression tree was constructed using training samples formed in 19 watersheds besides Baitugang. The blue line in figure below shows the Nash coefficient values of all potential donor watersheds parameters transplanted to the Baitugang watershed, ranging from -4.09 to 0.28. The orange mark represents the RF model-selected donor watersheds, ranging from -2.50 to 0.27. It can be seen from the results that after RF selected, the donor basins with poor simulation results are removed, and the watersheds with better transplantation results are retained.

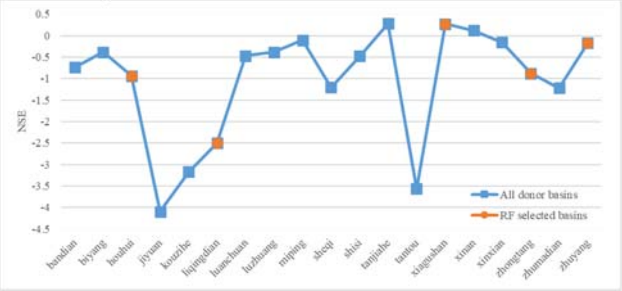

Fig. 7. Parameter transplantation result of Baitugang**.**

Take leave-one-out tests to all 20 watersheds. Figure 6 shows the all-possible transfers and RF selected transfers results. It can be seen that after RF selected, the minimum value of the parameter transplantation results of 15 of the 20 watersheds has been improved, indicating that the RF regression tree model can provide a reference for the selection of similar watersheds.

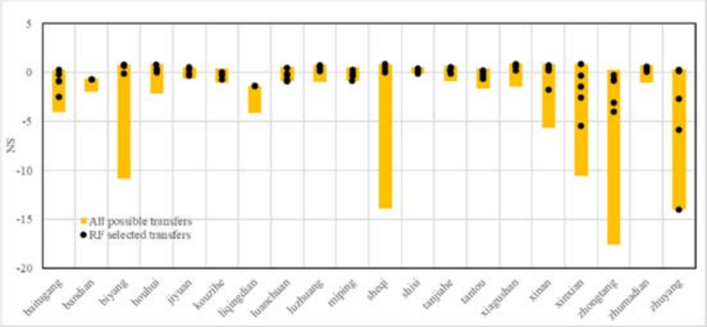

Fig. 8. Parameter transplantation result.

5 Conclusion

In order to improve the flash flood simulation in the ungauged or poor-gauged small-sized watershed in China. The parameter regionalization approach has been widely accepted as one of the main measures. With both measured and simulated data from 20 flash flood watersheds in Henan province of China, a new machine learning approach for parameter regionalization has been proposed with RF algorithm. The results of RF model showed that the percentage of sandy loam, silt loam, sandy clay could be considered as the top three priority components for determining the similarity between objective and sample watershed. Considering the special characteristics of the flash flood in Henan province which often expressed with short concentration time, the underlying properties related to the flow generation conditions are reasonable to have higher importance in the flash flood simulation. Therefore, the further rule produced by RF model for detecting the similar watershed for the objective watershed is able to be used for guiding the parameter regionalization in Henan province. Through the leave-one-out approach, the simulations with RF regionalized parameter showed higher performance than the simulation with random transferred parameters. Thus, the approach presented in this paper could be extended for other flash flood watersheds in China as one of main references to guide the determination of model parameters.

This study was supported by the National Key Research and Development of China (No. 2019YFC1510603), and the National Natural Science Foundation of China (No. 41807171).

References

1. L.Wei,K.H.Hu,Y.H.Huang, Comparisons of present situation and prevention and control of flash floods in China,United States and Japan,YR,**49**,29-33(2018)
2. R.Singh, S.A.Archfield, T.Wagener, Identifying dominant controls on hydrologic parameter transfer from gauged to ungauged catchments–A comparative hydrology approach, JH, **517**, (2014).
3. S.Ragettli, J.Zhou, H.Wang, Modeling flash floods in ungauged mountain catchments of China :A decision tree learning approach for parameter regionalization.,JH, **555**,330-346,(2017)
4. L.Breiman, Random Forests, ML, (2001)
5. H.Li,*Statistical learning methods*,(2012)
6. L.Guo, L.Q.Ding, Y.D.Sun, C.J.Liu, B.S.He, R.H.Liu, Key techniques of flash flood disaster prevention in Chin, JHE, **49**, 101-114,(2018)
7. C.J.Liu, L.Wen, J.Zhou, X.T.Zhao, L.Guo, Y.Q.Wei, Comparative analysis of hydrological and hydrodynamic calculation method for flash flood in small watershed, JCIWRHR, **17**,262-270,(2019)