Research on flood forecasting based on flood hydrograph generalization and random forest in Qiushui River basin, China

Tiantian Tang, Zhongmin Liang, Yiming Hu, Binquan Li and Jun Wang

ABSTRACT

At present, the use of hydrological models is the main technical approach for real-time flood forecasting. However, in semi-arid and arid areas, the use of the hydrological model is restricted by technical and data conditions. With the accumulation of hydrological data deluge, making full use of historical data and mining potential hydrological laws, causal relationships and other valuable information behind them provide new ideas for real-time flood forecasting in the study area. This paper develops a hybrid flood forecasting model that combines the flood hydrograph generalization method and random forest in the Qiushui River basin in the middle reaches of the Yellow River. The performance of this hybrid model is compared to that of the antecedent precipitation index model. For the development of these models, 23 flood events occurring from 1980 to 2010 are selected, of which 18 are used for calibration and 5 are used for validation. The results show that the hybrid model yields accurate predictions. And the comparison shows that the hybrid model performs better than the empirical model in the Qiushui River basin. Thus, this study provides a method for improving the accuracy of flood forecasting.

Key words | flood forecasting, flood hydrograph generalization, Qiushui River basin, random forest

Tiantian Tang
Zhongmin Liang (corresponding author)
Yiming Hu
Binquan Li
Jun Wang
College of Hydrology and Water Resources,
Hohai University,
Nanjing 210098,
China
E-mail: *zmliang@hhu.edu.cn*

Zhongmin Liang
National Cooperative Innovation Center for Water Safety & Hydro-Science,
Nanjing 210024,
China

HIGHLIGHTS

- A hybrid model of flood forecasting is proposed for a semi-arid and arid area.
- The hybrid model combines the random forest model and a flood hydrograph generalization method.
- The hybrid model outperforms the currently used Antecedent Precipitation Index model in the study area.

INTRODUCTION

Flood disasters often cause considerable loss of production and life, resulting in serious consequences. As an important supporting technology for flood control work, real-time flood forecasting plays a critical role in actual flood control. Commonly used real-time flood forecasting methods can be summarized into two types: physical process-driven flood forecasting models and the data-driven flood forecasting models. Both types of models have developed rapidly and have played an important role in production practices.

At present, the use of hydrological models is the main technical approach to real-time flood forecasting. With the vigorous development of computer technology in the mid-1950s, the study of hydrological models also ushered in a new opportunity for development. Conceptual hydrological models, such as the Stanford Basin Model (Crawford & Linsley 1966), the Soil Conservation Service (SCS) Model (McCuen 1982) and the Hec-1 Model (Chem 1947), emerged in this period. In the second half of the 20th century, many multi-parameter and

doi: 10.2166/hydro.2020.147

complex conceptual lumped models have been developed in succession by countries all over the world, such as the TANK model (Sugawara 1961), antecedent precipitation index (API) model (Sittner *et al.* 1969) and Xin'anjiang model (Renjun *et al.* 1980). These conceptual hydrological models have played an important role in studying hydrological laws and solving practical problems in production. Another conceptual model: distributed hydrological models also have made great progress, such as the Soil and Water Assessment Tool (SWAT) model (El-Nasr *et al.* 2005), SHE model (Abbott *et al.* 1986), the Systeme Hydrologique Europeen TRAN (SHETRAN) (Ewen 2000) model and the MIKE Systeme Hydrologique Europeen (MIKESHE) model (Refshaard & Storm 1995), which were developed on the basis of the SHE model. TOPographic Kinematic APproximation and Integration (TOPKAPI) (Ciarapica & Todini 2002) was established by combining the ARNO model with the TOPgraphy based hydrological (TOP) model and fully exploits the potential of the physical mechanisms of distributed models. Other distributed models include the IHDM (Institute of Hydrology Distributed Model) model and Variable Infiltration Capacity (VIC) model (Liang *et al.* 1994; 1996).

However, the existing hydrological models are more suitable for flood forecasting in humid areas. Our study area, Qiushui River basin, consists of an arid and semi-arid region where the spatial composition of flood sources is complex (Li *et al.* 2019), the forecast accuracy of hydrological models is often low, which is difficult to meet the needs of flood control and disaster reduction in this area (Li *et al.* 2018). Hence, there is an urgent need for a flood forecasting method that can not only avoid the direct simulation of physical flood formation processes in arid and semi-arid areas but also meet forecasting accuracy requirements. Another type of the flood forecasting model is the data-driven model. This type of model does not consider the physical mechanism of the hydrological process, regards the hydrological process as a black box and determines the mathematical function according to input and output data. Random forest (RF) is one of a data-driven model which combines the prediction from an ensemble of decision trees (Breiman 2001). RF has become popular in various industries due to its prediction power and the speed of processing (Svetnik *et al.* 2003; Belgiu & Drăguţ 2016; Dai *et al.* 2018; Zahedi *et al.* 2018). In hydrology, Peters *et al.* (2007) used RF as a tool for ecohydrological distribution modeling. Carlisle *et al.* (2010) used natural watershed characteristics to predict the value of each runoff metric using RF. Wang *et al.* (2015) proposed an approach for the flood hazard risk assessment model based on RF. Albers *et al.* (2015) determined the relative importance of contributing upstream discharges to the main stem during significant flood events. Yang *et al.* (2017) used RF to predict reservoir inflows for two headwater reservoirs in USA and China. Based on the flood hydrograph generalization method and RF model, this study intends to use advanced intelligent analysis technology to deeply extract knowledge from the data deluge and establish a new real-time flood forecasting method. Compare the new method with the empirical model: the API model, which is currently used in actual work.

The remainder of this paper is organized as follows: 'Study area and data' introduces the study area and the data used. The flood hydrograph generalization, RF model and rainfall–runoff relation methods are described in detail in the 'Methods' section. The results of flood forecasting and comparison of the hybrid model and empirical model are provided in the 'Results and discussion' section. Finally, the conclusions obtained from the study are outlined in the 'Conclusions' section.

STUDY AREA AND DATA

The Qiushui River basin is located in the left bank of the middle of the Yellow River. It is a tributary of the Yellow River and covers an area of 1,989 km^2. There are more than 20 branch ditches in the basin (the basin area is larger than 10 km^2) that are asymmetric pinnate inlets. Among the main branch, ditches are Taiping ditch, Chengzhuang ditch, Yulin ditch, Chegan ditch, Anye ditch, Dayu ditch and Zhaoxian ditch. The Qiushui River control station, named the Linjiaping Hydrological Station, is located in the upper 13 km of the Yellow River mouth, with a control area of 1,873 km^2. There are nine precipitation stations in the Qiushui River basin, namely, Zhangjiawan, Daipo, Yangposhuiku, Yaotou, Chengjiata, Linxian, Huangcaolin, Chegan and Linjiaping, with a network density of 208 km^2 per station. Since the Zhangjiawan and Daipo precipitation stations are located in the upper reaches of the Yangpo reservoir, only the other seven stations and Zhaojiagou were used to analyze the precipitation data (Figure 1).

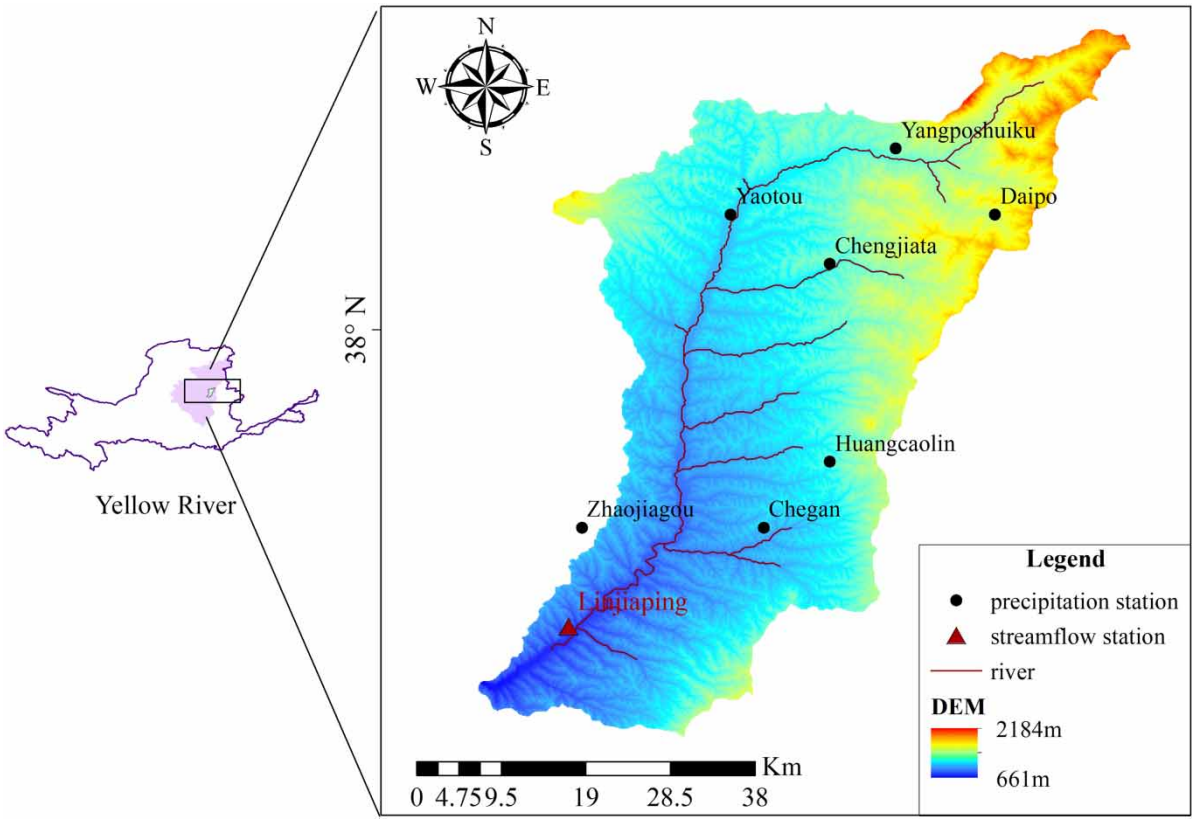

Figure 1 | Map of the Qiushui River basin.

This study uses 23 flood events that occurred during the period 1980–2010, with hourly observations of the discharge of Linjiaping station and the hourly precipitation data of Yangposhuiku, Daipo, Yaotou, Chengjiata, Huangcaolin, Chegan, Zhaojiagou and Linjiaping precipitation station. Eighteen flood events during the period 1980–1996 are used for calibration, while the remaining datasets during the period 1997–2010 are used for validation. According to the characteristics in this basin, the forecast lead time selected for this study is 5 h. Table 1 lists the information of these flood events, including the beginning and ending time.

METHODS

Flood hydrograph generalization method

Flood hydrograph generalization refers to the production of a representative flood hydrograph based on the observed flood hydrograph data of a large number of flood events at a hydrological station. The method flood hydrograph generalization comprises the following steps: first, combine each flood hydrograph into the same drawing, wherein the ordinate represents the ratio of Q_i and Q_m, the abscissa represents the ratio of T_i and T. Q_m is the peak discharge, T is the total duration of the flood process, and Q_i and T_i represent the discharge and time, respectively, at any time. Then, overlap the time of peak discharge in one place, one common hydrograph that summarizes the station flood shape characteristics of an average hydrograph is chosen as the generalized flood hydrograph. In this paper, considering that the flood recession process is long, the flood progress is divided into two parts: the rising and recession processes. Moreover, we assume that the total flood duration is twice as long as the duration of the rising process. The 6-point generalization method is taken as an example to control the hydrograph characteristics, as shown in Figure 2. When the flood hydrograph is calculated by the generalization map, the coordinates of the points in the graph are $(0, \alpha_1 Q_m)$, $(\beta_1 T, \alpha_2 Q_m)$, $(\beta_2 T, Q_m)$, $(\beta_3 T, \alpha_3 Q_m)$, $(0.5T, Q_g)$, $(0, T)$, respectively. Here, Q_m is the peak discharge,

Table 1 | Information of the 23 flood events used for calibration and validation of the models

Flood number	Start time	End time
Calibration		
19800818	10:00, 18 August 1980	8:00, 19 August 1980
19810620	6:00, 20 June 1981	20:00, 21 June 1981
19810703	16:00, 3 July 1981	20:00, 4 July 1981
19810707	14:00, 7 July 1981	4:00, 8 July 1981
19840701	5:00, 1 July 1984	20:00, 2 July 1984
19850805	20:00, 5 August 1985	20:00, 6 August 1985
19880715	2:00, 15 July 1988	0:00, 16 July 1988
19880718	12:00, 18 July 1988	20:00, 19 July 1988
19890716	20:00, 16 July 1989	16:00, 17 July 1989
19890722	5:00, 22 July 1989	20:00, 23 July 1989
19900811	19:00, 11 August 1990	16:00, 12 August 1990
19910610	4:00, 10 June 1991	8:00, 11 June 1991
19910721	16:00, 21 July 1991	8:00, 22 July 1991
19910727	22:00, 21 July 1991	16:00, 28 July 1991
19910915	1:00, 15 September 1991	20:00, 15 September 1991
19920802	16:00, 2 August 1992	20:00, 3 August 1992
19920828	20:00, 28 August 1992	12:00, 29 August 1992
19960809	19:00, 9 August 1996	12:00, 10 August 1996
Validation		
19970718	6:00, 18 July 1997	20:00, 19 July 1997
19970731	9:00, 31 July 1997	8:00, 1 August 1997
19990711	10:00, 11 July 1999	2:00, 12 July 1999
20000708	4:00, 8 July 2000	0:00, 9 July 2000
20100919	6:00, 19 September 2010	20:00, 19 September 2010

Q_g is the maximum discharge of the recession process, and T is the flood duration.

Random forest

RF (Breiman 2001) is a machine learning algorithm combining the Bagging ensemble learning theory (Breiman 1996) and the random subspace method (Ho 1998). An RF is a classifier consisting of a collection of tree-structured classifiers $\{h(x, \Theta_n), n = 1, \ldots\}$, where $\{\Theta_n\}$ are independent identically distributed random vectors, and each tree casts a unit vote for the most popular class at input x (Breiman 2001). RF utilizes bootstrap resampling technology to sample original samples to generate a number of training samples, each of which randomly selects feature attributes through random subspace methods to construct a decision tree. Finally, the optimal result is obtained by the voting or averaging method. Previous studies have found that RF can effectively overcome the problems of noise and overfitting and obtain a high prediction precision (Wang *et al.* 2015). RF has two main technological aspects: the first is bootstrap resampling technology and out-of-band error estimation; the second is decision tree construction and the random subspace theory (Liang *et al.* 2017). The main structure of the model is shown in Figure 3.

Hybrid model

The hybrid model combines the two methods above. First, the generalized flood hydrograph was obtained by the flood hydrograph generalization method. Then, construct correlations between the flood factors and precipitation factors to screen predictors. And the flood factors series and predictors series were as the input to the RF model to forecast flood factors. The RF algorithm is implemented by Matlab. Finally, according to the forecasted flood factors, the forecasted flood hydrograph was obtained through scaling up the generalized flood hydrograph. The main structure of the hybrid model is shown in Figure 4.

API model

Sittner *et al.* (1969) proposed the API model for computing a groundwater flow hydrograph; the API uses the unit hydrograph (UH) method to develop a model to simulate the flow hydrograph.

The API model is based on the physical mechanism of rainfall and runoff generation in basins and takes the main influencing factors as parameters to establish the quantitative correlation between rainfall and runoff. Some common parameters are antecedent precipitation, seasonal characteristics and precipitation duration.

We use the $(P + P_a) - R$ relation graph (Kohler & Linsley 1951), which is a graph between the sum of two values of precipitation (P) and antecedent precipitation (P_a) and runoff (R), as shown in Equation (1):

$$R = f(P, P_a) \tag{1}$$

The main steps of this method are as follows: first, calculate the average daily precipitation in the basin. Second,

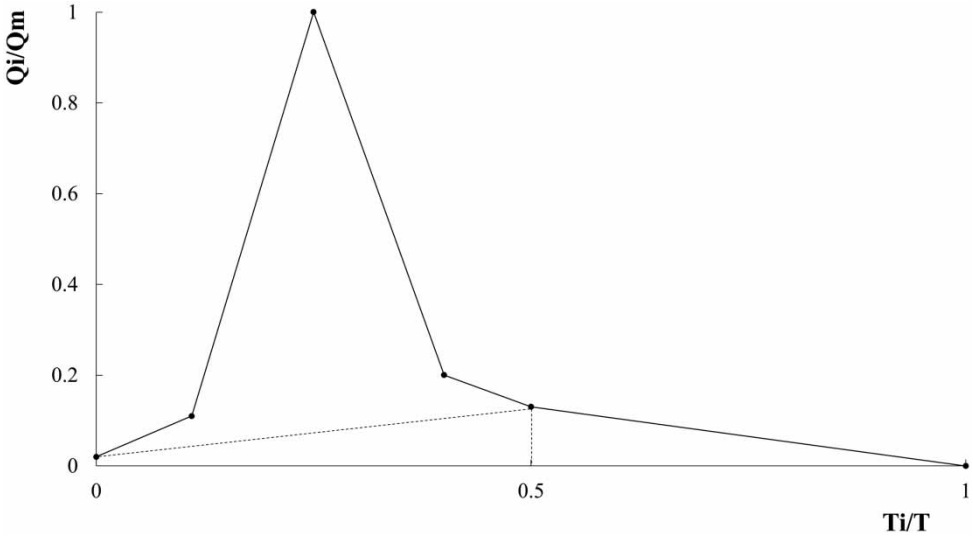

Figure 2 | Sketch map of flood hydrograph generalization.

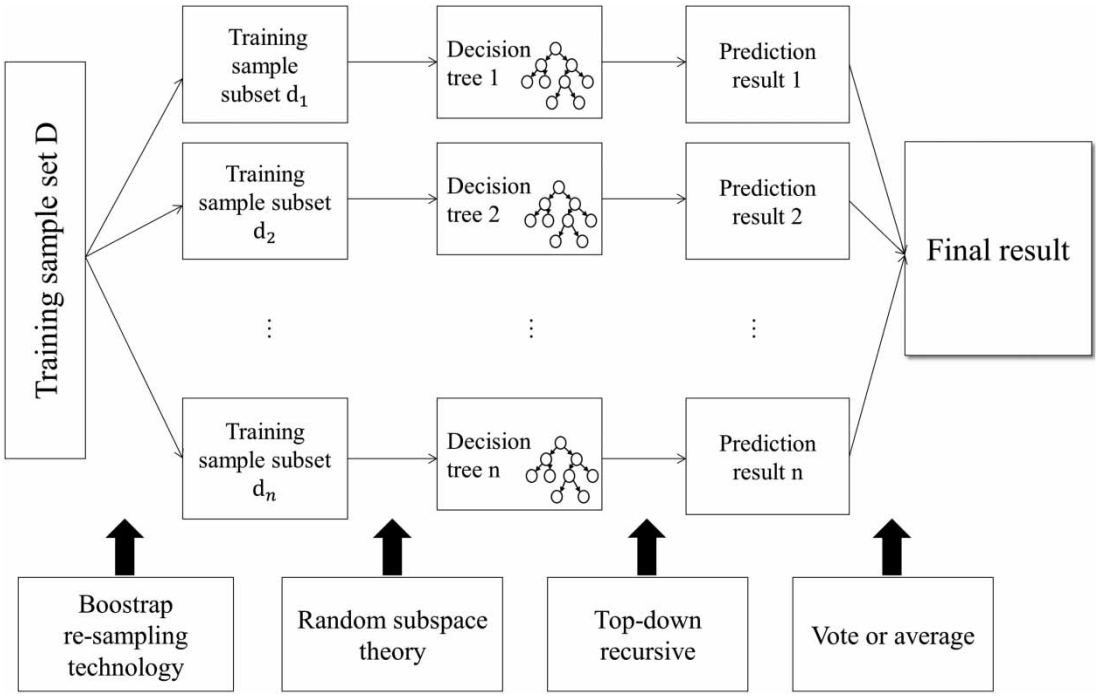

Figure 3 | The main structure of the RF.

compute the values of antecedent precipitation at the early stage of the forecast period as follows:

$$P_{a,t} = kP_{t-1} + k^2P_{t-2} + \cdots + k^nP_{t-n} \quad (2)$$

where n is the number of days that influence the flood event, with a general value of 15 days; k is a constant coefficient, with a general value of 0.85 (Bao 2006). Third, R is calculated from the relation between $(P + P_a) - R$. Finally, taking R as the input quantity, the flood hydrograph

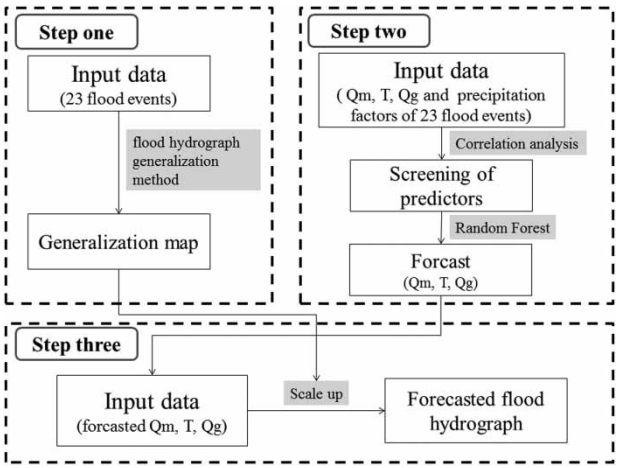

Figure 4 | The main structure of the hybrid model.

is calculated according to the UH of the basin for flood forecasting.

Evaluation measures

To evaluate the forecasting ability of the models, the simulation accuracy of each flood event is summarized. The relative error of peak discharge (DQ_m), relative error of the maximum discharge of the recession process (DQ_g), absolute error of the duration of the rising process (DT), coefficient of correlation (CC) and root mean square error (RMSE) are used as evaluation measures (see Supplementary Material). The CC measures the degree of correlation among the observed and simulated values, ranging between $(-\infty, 1]$, and the model is ideal if the CC value is close to 1. The RMSE evaluates the variance of errors, and the smaller the value of the RMSE, the better is the performance of the model. These variables are defined as follows:

$$DQ_m = \frac{|Q_{cal} - Q_{obv}|}{Q_{obv}} \times 100\% \tag{3}$$

$$DQ_g = \frac{|Q_{cal} - Q_{obv}|}{Q_{obv}} \times 100\% \tag{4}$$

where Q_{obv} and Q_{cal} are observed and simulated discharge, respectively.

$$DT = |T_{cal} - T_{obv}| \tag{5}$$

where T_{obv} and T_{cal} are observed and simulated flood duration, respectively.

$$CC = \frac{\frac{1}{N}\sum_{t=1}^{N}(Q_{obs}(i) - \overline{Q_{obs}})(Q_{cal}(i) - \overline{Q_{cal}})}{\sqrt{\frac{1}{N}\sum_{t=1}^{N}(Q_{obs}(i) - \overline{Q_{obs}})^2}\sqrt{\frac{1}{N}\sum_{t=1}^{N}(Q_{cal}(i) - \overline{Q_{cal}})^2}} \tag{6}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}[Q_{obs}(i) - Q_{cal}(i)]^2}{N}} \tag{7}$$

where $Q_{obs}(i)$ and $Q_{cal}(i)$ are the observed and simulated discharge series, respectively; $\overline{Q_{obs}}$ and $\overline{Q_{cal}}$ are the mean observed and simulated discharge series, respectively; and N is the length of the time series considered.

RESULTS AND DISCUSSION

Hybrid model development

Flood hydrograph generalization

Considering the long process of flood recession, the flood progress is divided into two parts: the rising and recession processes. The generalization method is used to generalize these two processes. The rising and recession processes of 18 flood events of the calibration period were generalized for each flood hydrograph, and finally, the general flood hydrograph was obtained by averaging the individual flood hydrographs (Figures 5–8).

Screening of predictors

A correlation analysis between the precipitation factors (peak hourly precipitation (P_m), accumulated 5 h precipitation (AP_5), accumulated 10 h precipitation (AP_{10}), accumulated 15 h precipitation (AP_{15}), precipitation intensity during rising process (PI), time of peak discharge (T_{Q_m}), time of peak precipitation (T_{P_m}) and peak discharge (Q_m), duration of the rising process (T_s) and the maximum discharge of the recession process (Q_g) were established, respectively, to select the key influential predictors, as shown in Table 2.

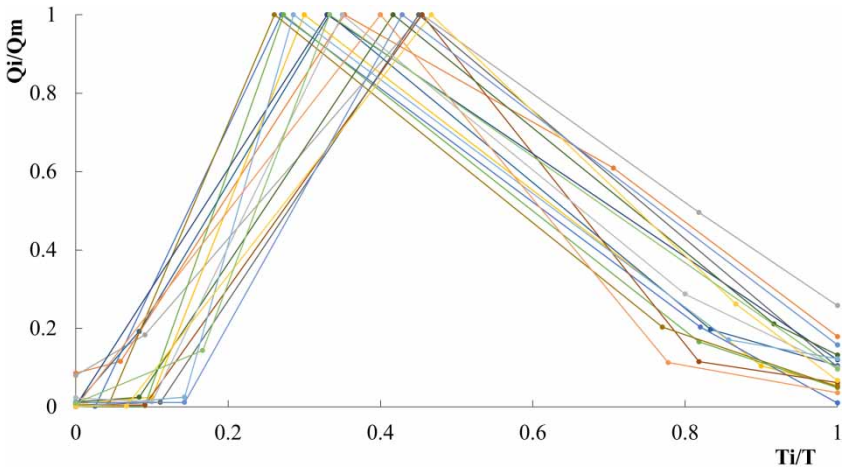

Figure 5 | Sketch map of 23 flood hydrographs generalization of the rising process.

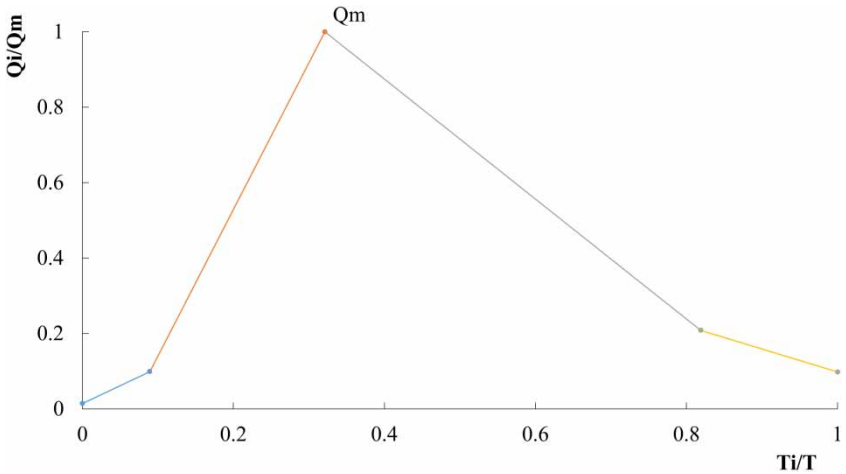

Figure 6 | Sketch map of the general flood hydrograph generalization of the rising process.

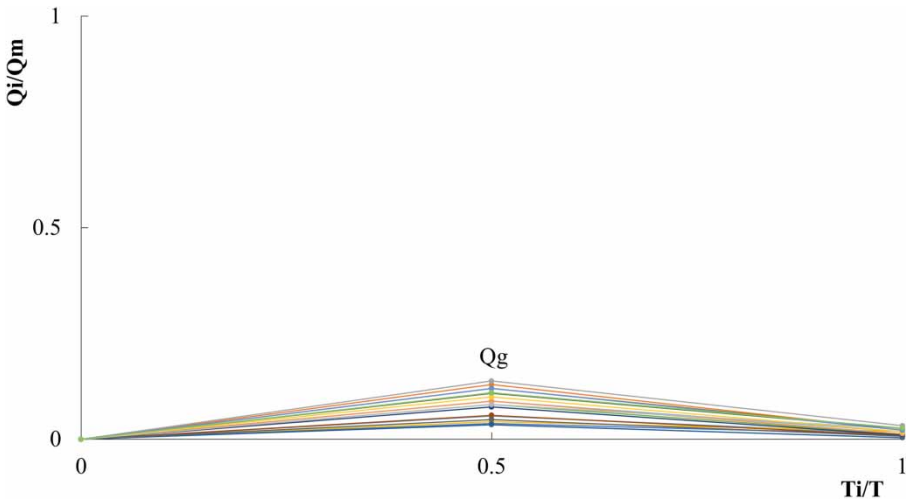

Figure 7 | Sketch map of 23 flood hydrographs generalization of the recession process.

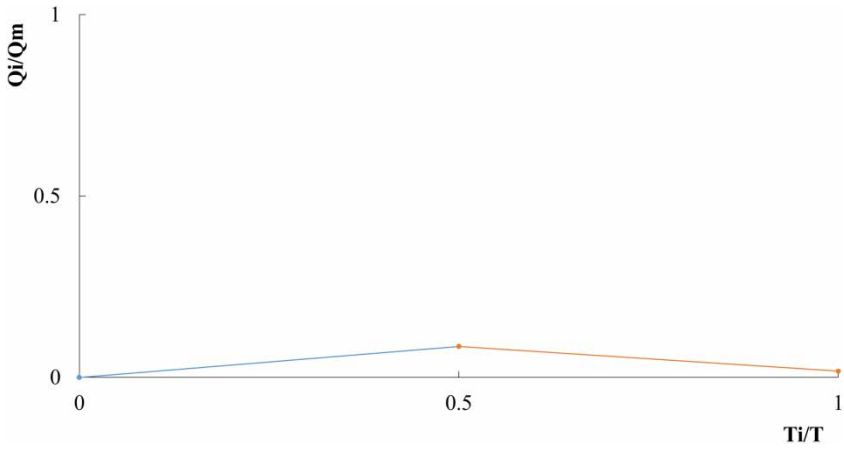

Figure 8 | Sketch map of the general flood hydrograph generalization of the recession process.

Table 2 | Predictands and predictors

Predictands	Predictors
Q_m	P_m, AP_{10}, AP_{15}, PI
T_s	T_{Q_m}, T_{P_m}, AP_{15}, PI
Q_g	P_m, AP_{10}, AP_{15}

RF model development

The model was built in three steps. In the first step, peak discharge was forecasted. The duration of the rising process and maximum discharge of the recession process were forecasted in the second and third steps, respectively, with the total duration set to twice the duration of the rising process. The predicted flood hydrograph was obtained by substituting the predicted time into the generalized flood process. Eighteen floods were used to calibrate the model, and five floods were used to validate the model.

First, the selected M predictors were used to construct the training sample set D together with the predictand series.

$$D = \{(x_i, y_i), x_i \in X, y_i \in Y, i = 1, 2, \ldots, N\} \quad (8)$$

where X is the M-dimensional explanatory variable vector composed of predictors, Y is the target variable of the predictand series, and N is the sample capacity. Second, n training sample subsets were randomly taken from the training sample set D through bootstrap resampling, and the size of the training sample subset was N. Third, n decision trees were constructed for the n training sample subsets. According to the random subspace theory, m indexes (generally, $m = \sqrt{M}$) were randomly selected from the M indexes. Then, the optimal value was selected based on the principle of entropy increasing, and this value was the final node attribute value. n in this research was set at 100. Finally, each decision tree was executed based on top-down recursive growth to obtain a predicted value. The results of the n decision trees were then voted or averaged to obtain the ultimate classification or regression results, namely, the final values of predictands.

When forecasting the predictands, the training sample of the RF model was calculated as follows:

$$D = \{(x_i, y_i), y_i \in Q_m, x_i \in |P_m, AP_{10}, AP_{15}, PI|, i = 1, 2, \ldots, N\} \quad (9)$$

$$D = \{(x_i, y_i), y_i \in T_s, x_i \in |T_{p_m}, T_{Q_m}, AP_{15}, PI|, i = 1, 2, \ldots, N\} \quad (10)$$

$$D = \{(x_i, y_i), y_i \in Q_g, x_i \in |P_m, AP_{10}, AP_{15}|, i = 1, 2, \ldots, N\} \quad (11)$$

The forecasting statistics of the RF model in the calibration period are shown in Table 3. Moreover, according to the accuracy requirement of the flood forecasting in the standard for hydrological information and hydrological forecasting in China (GB/T22482-2008), a 20% variation between the observed peak discharge and forecasted peak

Table 3 | Error statistics of the RF model in the calibration period

	Q_m (m³/s)			T_s (h)			Q_g (m³/s)		
Flood number	Q_{obv}	Q_{cal}	DQ_m	T_{obv}	T_{cal}	DT	Q_{obv}	Q_{cal}	DQ_g
19800818	373	447	19.8%	11	9	2	21	25	19.0%
19810620	312	371	18.9%	18	15	3	40	40	0.0%
19810703	238	327	37.4%	12	13	1	33	45	36.4%
19810707	1480	1221	17.5%	10	10	0	61	37	39.3%
19840701	284	299	5.3%	14	15	1	24	21	12.5%
19850805	838	617	26.4%	12	10	2	31	45	45.2%
19880715	582	689	18.4%	13	12	1	48	43	10.4%
19880718	975	832	14.7%	12	12	0	34	28	17.6%
19890716	698	780	11.7%	10	13	3	40	43	7.5%
19890722	1520	1280	15.8%	23	15	8	78	61	21.8%
19900811	221	238	7.7%	7	8	1	7	8	14.3%
19910610	283	356	25.8%	13	13	0	31	30	3.2%
19910721	207	261	26.1%	8	9	1	16	11	31.3%
19910727	797	648	18.7%	10	8	2	21	27	28.6%
19910915	256	283	10.5%	11	10	1	21	27	28.6%
19920802	385	477	23.9%	16	16	0	19	32	68.4%
19920828	328	392	19.5%	8	10	2	32	30	6.3%
19960809	523	627	19.9%	7	9	2	40	35	12.5%

discharge is taken as the permissible error, and the qualified rate (QR) is calculated.

In the calibration period of the model, the values of the average relative error of the peak discharge of 13 events were less than 20%; thus, QR of forecasting of Q_m is 72%. In addition, the average value of DQ_m in the calibration period was 18.8%. As for the flood duration, there were four flood events with DT values at 0, which is a considerable result. However, the DT value of No. 19890722 was relatively large, mainly because the observed value is relatively large. Ten of the 18 flood events had average relative error of Q_g less than 20%; therefore, QR of forecasting of Q_g is 55.6%. The average value of the DQ_g of the calibration period was 22.4%. Broadly speaking, however, the accuracy was satisfactory. Table 4 shows the forecasting statistics of the RF model in the validation period.

During the validation period, QR of forecasting of Q_m and Q_g is 80 and 60%, respectively. The average value of

Table 4 | Error statistics of the RF model in the validation period

	Q_m (m³/s)			T (h)			Q_g (m³/s)		
Flood number	Q_{obv}	Q_{cal}	DQ_m	T_{obv}	T_{cal}	DT	Q_{obv}	Q_{cal}	DQ_g
19970718	201	235	16.9%	14	16	2	39	52	33.3%
19970731	800	960	20.0%	10	12	2	70	66	5.7%
19990711	345	380	10.1%	8	10	2	39	36	7.7%
20000708	1100	1030	6.4%	13	13	0	75	75	0.0%
20100919	2280	1170	48.7%	8	10	2	274	71	74.1%

the DQ_m and DQ_g in the validation period was 20.4 and 24.2%, respectively. It has to be noted, for the event No. 20100919, the accuracies of forecasting of the peak discharge and maximum discharge of the recession process are both very low. It is mainly because of the peculiarity of the RF model. RF cannot make prediction beyond the range of training set data despite it being a powerful model. Namely, when the maximum and minimum of the peak discharge in the training set is 1,520 and 207 m^3/s, the forecasted discharge cannot be greater than 1,520 m^3/s or smaller than 207 m^3/s. However, the observation of the Q_m of No. 20100919 is 2,280 m^3/s, which is beyond the maximum value of the training set. Therefore, the effect will be relatively poor in any case. The same is true for Q_g. In general, the model provided acceptable accuracy in both the calibration and validation periods.

Empirical model development

The $(P + P_a) - R$ relation graph is shown in Figure 9. The UH was derived in a conventional manner using the selected events, as shown in the figure.

Generally, if the precipitation intensity is large, the peak discharge of UH is higher and the peak time is earlier. However, the peak discharge is lower and the peak time lags behind. When the precipitation center is in the upstream, due to the long confluence path, the peak discharge of UH is lower and the peak time lags behind. However, the peak discharge is higher and the peak time is earlier. Therefore, the UH is classified and compiled according to the location of the precipitation center and the magnitude of the net precipitation. We summarized the flood events into four types of unit hydrographs, as shown in Figure 10. The classification of the flood events is shown in Table 5.

Counting the values of $P + P_a$, according to the relation graph, the values of R will be calculated, then, according to the UH and Equation (12), the flood process is obtained.

$$Q_{d,t} = \sum_{j=k_1}^{k_2} r_{d,j} q_{t-j+1} \tag{12}$$

Comparison of the hybrid model and empirical model

The CC and the RMSE of the hybrid model and empirical model in the calibration period are summarized and shown in Table 6.

It is evident from Table 5 that the hybrid model performs better than the empirical model in the calibration

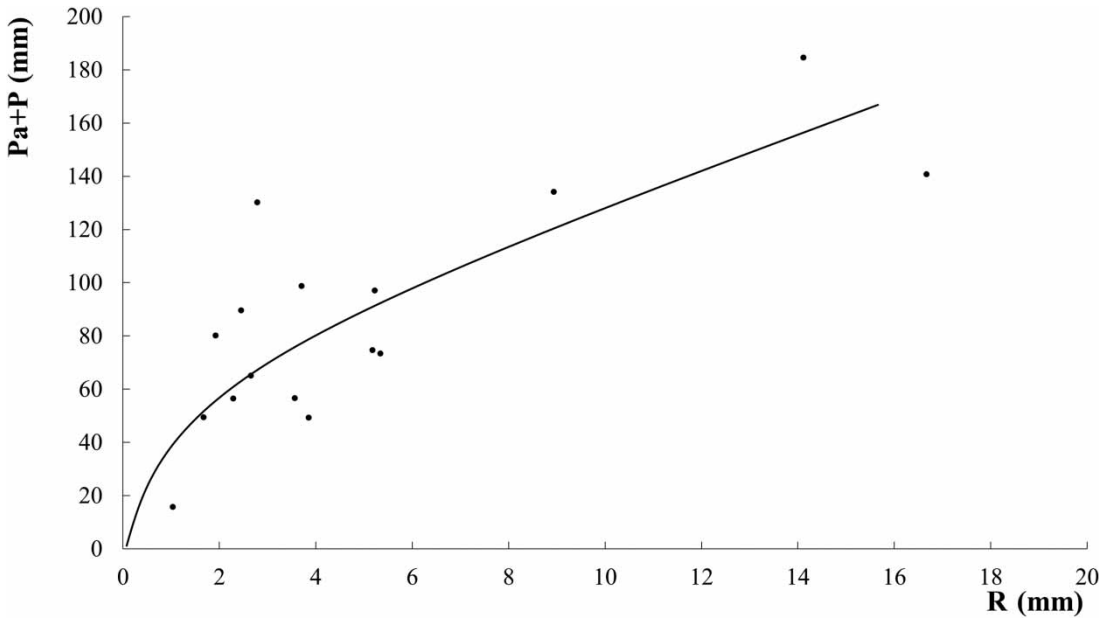

Figure 9 | $(P + P_a) - R$ relation graph.

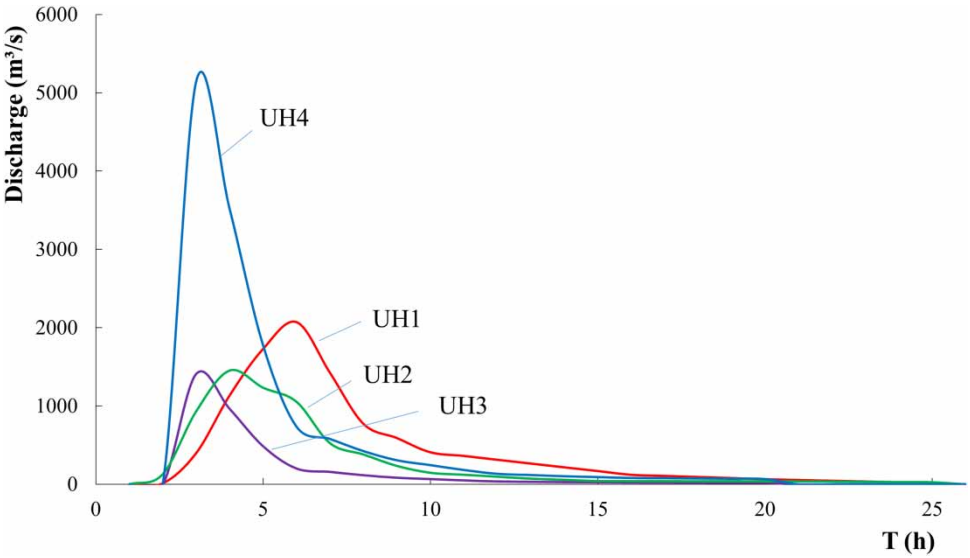

Figure 10 | Four types of UH.

Table 5 | Classification of the flood events

UH	Flood number
UH1	19880715 19890716 19960809
UH2	19800818 19810620 19810703 19840701 19910610 19920802
UH3	19900811 19910721 19910915 19920828
UH4	19810707 19850805 19880718 19890722 19910727

Table 6 | Performance comparison of the hybrid and empirical model in the calibration period

CC		RMSE (m^3/s)	
Hybrid model	**Empirical model**	**Hybrid model**	**Empirical model**
0.80	0.67	155	226

period. Yet, there are two events which have different results. The results of No. 19840701 and No. 19910721 indicate that the empirical model has better values of CC, which is 0.02 and 0.06 higher than the hybrid model. This might be explained by the variance of the antecedent precipitation. When the antecedent precipitation is well distributed in the temporal scale, the empirical model performs well, and when the antecedent precipitation is more concentrated in a short time, the hybrid model is better. However, the real problem of Qiushui River basin is that the spatial and temporal distribution of precipitation is often uneven and expressed in peaks with rising and dropping steeply. Thus, the hybrid model can be more suitable than the traditional model in the

Table 7 | Performance comparison of the hybrid and empirical model in the validation period

Flood Number	CC		RMSE (m^3/s)	
	Hybrid model	**Empirical model**	**Hybrid model**	**Empirical model**
19970718	0.88	0.43	40	194
19970731	0.88	0.86	228	253
19990711	0.87	0.84	65	223
20000708	0.88	0.63	179	374
20100919	0.73	0.35	538	692

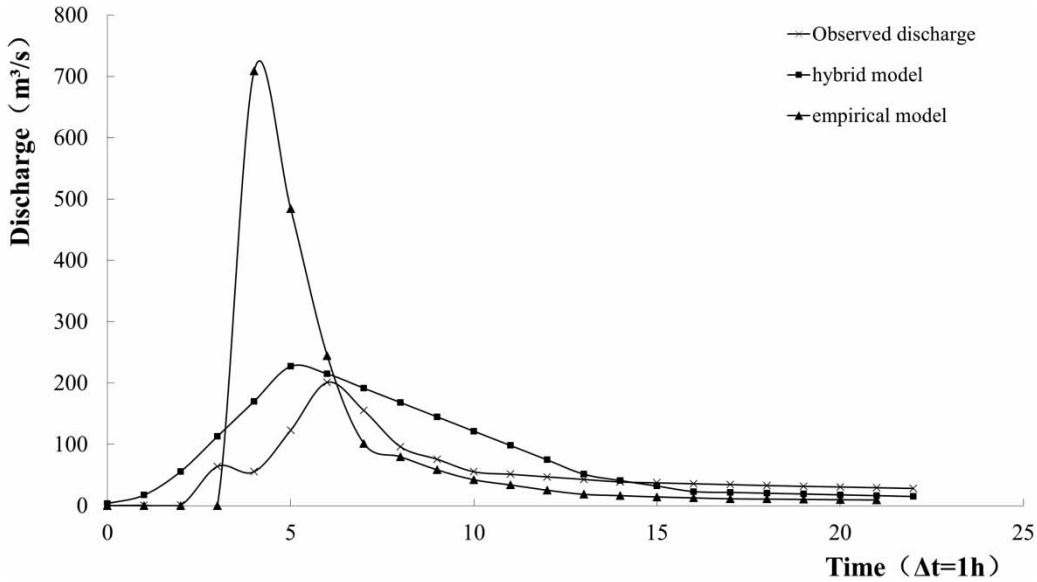

Figure 11 | Observed and forecasted flood hydrograph of event No. 19970718.

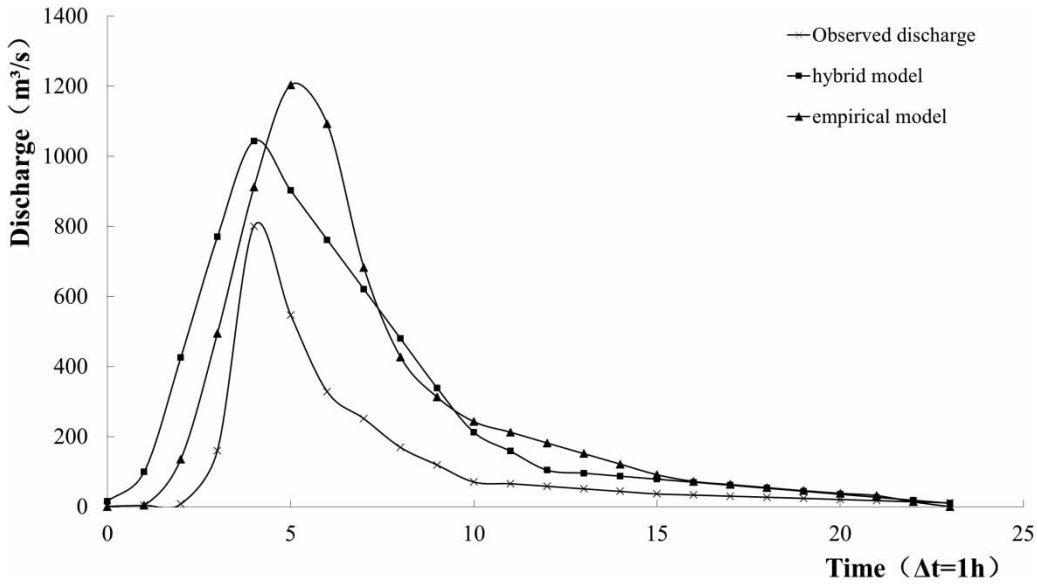

Figure 12 | Observed and forecasted flood hydrograph of event No. 19970731.

study area. Moreover, Table 7 lists the CC values and RMSE values of the hybrid model and empirical model in the validation period. See Figures 11–15 for comparison of observed and forecasted flood hydrographs of the five flood events in the validation period.

During the validation period, the CC values of the hybrid model were all above 0.7 and there were four flood events with CC values above 0.85. However, the CC values of only two events were above 0.8 for the empirical model. In addition, for event No. 19990711, although the CC value of the empirical model is 0.84, Figure 12 shows that the hydrograph forecasted by the empirical model only follows the trend of the observed hydrograph and the difference between forecasted peak discharge and observed peak discharge is relatively large. On average, the values of CC of the hybrid model and the empirical model are 0.85

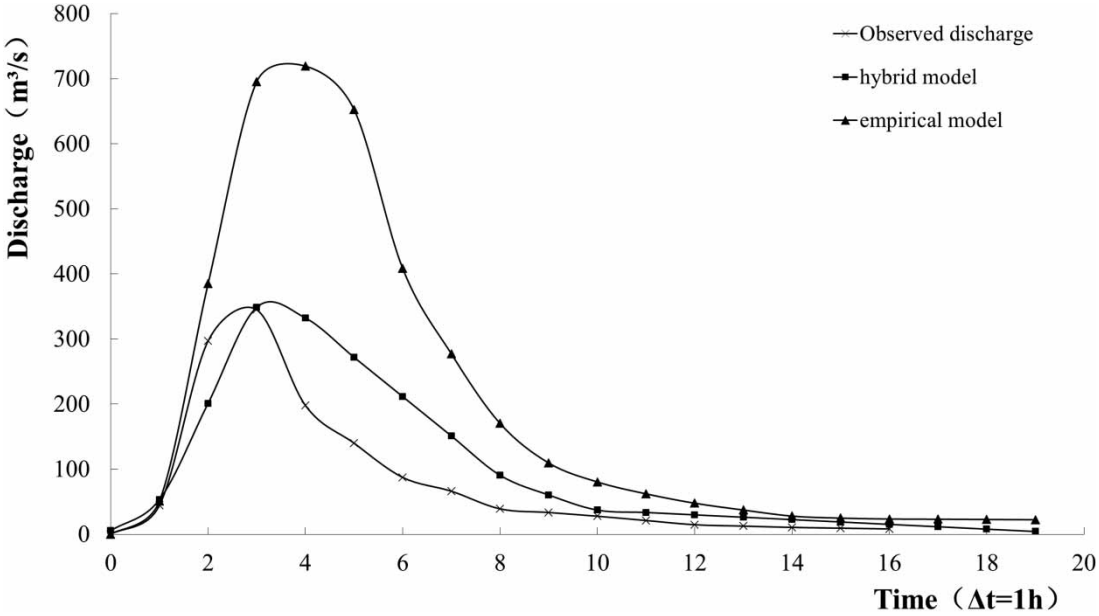

Figure 13 | Observed and forecasted flood hydrograph of event No. 19990711.

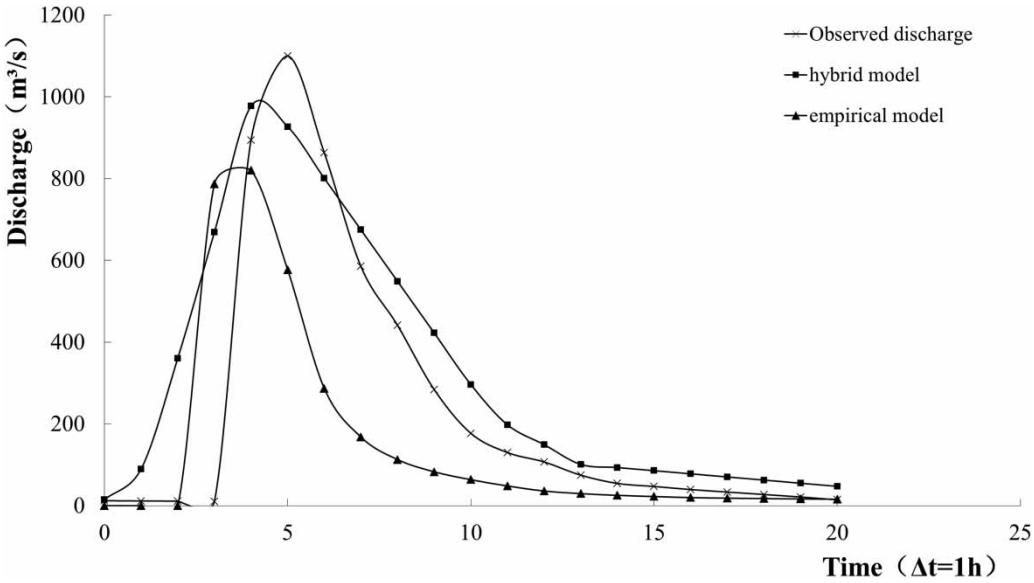

Figure 14 | Observed and forecasted flood hydrograph of event No. 20000708.

and 0.62, respectively, and the values of RMSE of the hybrid model and the empirical model are 210 and 347 m^3/s, respectively. The comparison indicates that the hybrid model provided a better flood forecasting fit based on observations compared to the empirical model both in the calibration and validation period.

CONCLUSIONS

In order to solve the problem of the low forecasting accuracy of hydrological models in arid and semi-arid areas, this paper develops a flood forecast method that combines the flood hydrograph generalization method and RF in the

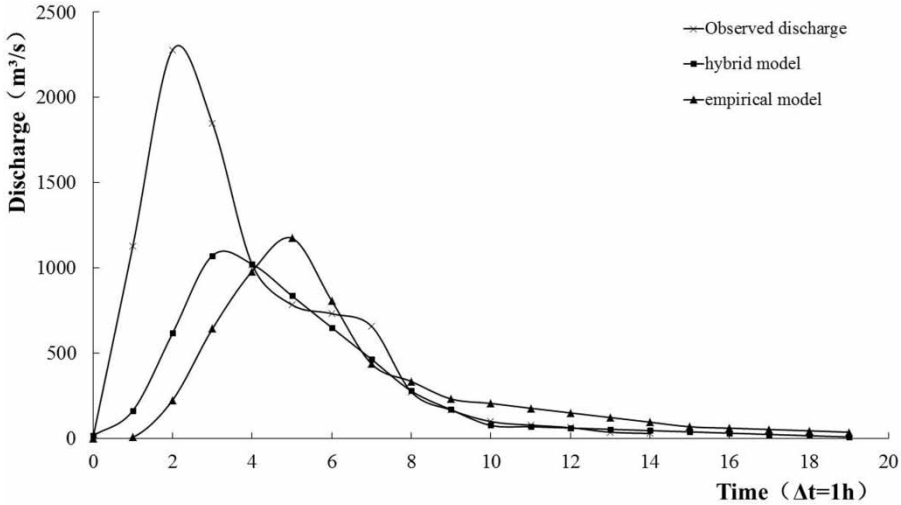

Figure 15 | Observed and forecasted flood hydrograph of event No. 20100919.

Qiushui River basin. First, selected flood events from 1980 to 2010 were generalized using the flood hydrograph generalization method. Then, the peak discharge and flood duration were forecasted using the RF method, and the flood processes were deduced. The specific findings of this study are as follows:

1. RF cannot make prediction beyond the range of training set data, which may lead to the poor prediction effect when we do the extreme value prediction. The solution for that problem could not be proposed in this study and must be left for future work.
2. Our study found that when the antecedent precipitation is well distributed in the temporal scale, the empirical model performs well, and when the antecedent precipitation is more concentrated in short time, the hybrid model is better. Thus, the accuracy of the current hydrological model is often lower in arid and semi-arid areas with more complex hydrological processes. The Qiushui River basin is an arid and semi-arid area where the spatial and temporal distribution of precipitation is often uneven and expressed in peaks with rising and dropping steeply. In this study, for the hybrid model, the average CC of the calibration period and validation period of the forecasted and observed flood progress were 0.80 and 0.85, respectively. For the empirical model, the average correlation coefficient of the calibration period and validation period of the forecasted and observed flood progress were 0.67 and 0.62, respectively. The results indicate that the hybrid model provides a better flood forecasting performance than the empirical model. Consequently, this study provides a new method for improving the accuracy of flood forecasting.

ACKNOWLEDGEMENTS

This work is supported in part by the National Key Research and Development Program of China (2016YFC0402709, 2016YFC0402706), National Natural Science Foundation of China (41730750), and National Natural Science Foundation of China (41877147). In addition, the authors are indebted to the editors/reviewers for their valuable comments and suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this paper is available online at https://dx.doi.org/10.2166/hydro.2020.147.

REFERENCES

Abbott, M. B., Bathurst, J. C., Cunge, J. A., O'Connell, P. E. & Rasmussen, J. 1986 An introduction to the European hydrological system – Systeme Hydrologique Europeen, 'SHE', 1: history and philosophy of a physically-based, distributed modelling system. *Journal of Hydrology* **87** (1), 61–77. doi:10.1016/0022-1694(86)90114-9.

Albers, S. J., Déry, S. J. & Petticrew, E. L. 2015 Flooding in the Nechako River Basin of Canada: a random forest modeling approach to flood analysis in a regulated reservoir system. *Canadian Water Resources Journal/Revue canadienne des ressources hydriques* 1–11. doi:10.1080/07011784.2015.1109480

Bao, W. M. 2006 *Hydrological Forecasting*, Vol. 3. China Water & Power Press, Beijing, p. 29

Belgiu, M. & Drăguţ, L. 2016 Random forest in remote sensing: a review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing* **114**, 24–31.

Breiman, L. 1996 Bagging predictors. *Machine Learning* **24** (2), 123–140. doi:10.1007/bf00058655.

Breiman, L. 2001 Random forests. *Machine Learning* **45** (1), 5–32. doi:10.1023/A:1010933404324.

Carlisle, D. M., Falcone, J., Wolock, D. M., Meador, M. R. & Norris, R. H. 2010 Predicting the natural flow regime: models for assessing hydrological alteration in streams. *River Research and Applications.* **26** (2), 118–136.

Chem, A. 1947 Corps of engineers. *Analytical Chemistry* **19** (3), 16A. doi:10.1021/ac60003a715.

Ciarapica, L. & Todini, E. 2002 TOPKAPI: a model for the representation of the rainfall-runoff process at different scales. *Hydrological Processes* **16** (16), 207–229. doi:10.1002/hyp.342.

Crawford, N. H. & Linsley, R. K. 1966 Digital Simulation in Hydrology: Stanford Watershed Model IV. Department of Civil Engineering, Stanford University, California. Technical Rep. 39.

Dai, B., Gu, C., Zhao, E. & Qin, X. 2018 Statistical model optimized random forest regression model for concrete dam deformation monitoring. *Structural Control Health Monitoring* **25** (6), 1–15.

El-Nasr, A. A., Arnold, J. G., Feyen, J. & Berlamont, J. 2005 Modelling the hydrology of a catchment using a distributed and a semi-distributed model. *Hydrological Processes* **19** (3). doi:0.1002/hyp.5610.

Ewen, J. 2000 SHETRAN: distributed river basin flow and transport modeling system. *Journal of Hydrologic Engineering* **5** (3), 250–258.

Ho, T. K. 1998 The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20** (8), 832–844. doi:10.1109/34.709601.

Kohler, M. A. & Linsley, R. K. 1951 Predicting the runoff from storm rainfall. *Lasers in Surgery and Medicine* **19** (4), 407–412. doi:10.1002/(SICI)1096-9101(1996)19:4 < 407:: AID-LSM4 > 3.0.CO;2-W.

Li, B., Liang, Z., Zhang, J., Wang, G., Zhao, W., Zhang, H., Wang, J. & Hu, Y. 2018 Attribution analysis of runoff decline in a semiarid region of the Loess Plateau, China. *Theoretical and Applied Climatology* **131** (1–2), 845–855. doi:10.1007/s00704-016-2016-2.

Li, B., Liang, Z., Bao, Z., Wang, J. & Hu, Y. 2019 Changes in streamflow and sediment for a planned large reservoir in the middle Yellow River. *Land Degradation & Development* **30** (7), 878–893. doi:10.1002/ldr.3274.

Liang, X., Lettenmaier, D. P., Wood, E. F. & Burges, S. J. 1994 A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research Atmospheres* **99** (D7), 14415–14428. doi:10.1029/96JD01448.

Liang, X., Lettenmaier, D. P. & Wood, E. F. 1996 One-dimensional statistical dynamic representation of subgrid spatial variability of precipitation in the two-layer variable infiltration capacity model. *Journal of Geophysical Research Atmospheres* **101** (D16), 21403–21422.

Liang, Z., Tang, T., Li, B., Liu, T., Wang, J. & Hu, Y. 2017 Long-term streamflow forecasting using SWAT through the integration of the random forests precipitation generator: case study of Danjiangkou Reservoir. *Hydrology Research* **49** (5), 1513–1527. doi:10.2166/nh.2017.085.

McCuen, R. H. 1982 *A Guide to Hydrologic Analysis Using SCS Methods*. doi:10.1007/BF00820732.

Peters, J., Baets, B. D., Verhoest, N. E. C., Samson, R., Degroeve, S., Becker, P. D. & Huybrechts, W. 2007 Random forests as a tool for ecohydrological distribution modelling. *Ecological Modelling* **207** (2–4), 304–318.

Refshaard, J. C. & Storm, B. 1995 *MIKE SHE*. in *Computer Models of Watershed Hydrology*.

Renjun, Z., Yilin, Z., Lerun, F., Xinren, L. I. U. & Quan, Z. 1980 The Xinanjiang model. *Proceedings of the Oxford Symposium on Hydrological Forecasting IAHS Publ* **135** (1), 371–381.

Sittner, W. T., Schauss, C. E. & Monro, J. C. 1969 Continuous hydrograph synthesis with an API-type hydrologic model. *Water Resources Research* **5** (5), 1007–1022. doi:10.1029/wr005i005p01007.

Sugawara, M. 1961 On the analysis of runoff structure about several Japanese rivers. *Japanese Journal of Geophysics* **2** (4), 1–76.

Svetnik, V., Liaw, A., Tong, C., Christopher Culberson, J., Sheridan, R. P. & Feuston, B. P. 2003 Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences* **43** (6), 1947–1958. doi:10.1021/ci034160 g.

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S. & Bai, X. 2015 Flood hazard risk assessment model based on random forest. *Journal of Hydrology* **527**, 1130–1141. doi:10.1016/j.jhydrol.2015.06.008.

Yang, T., Asanjan, A. A., Welles, E., Gao, X., Sorooshian, S. & Liu, X. 2017 Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. *Water Resources Research* **53** (4), 2786–2812.

Zahedi, P., Parvandeh, S., Asgharpour, A., McLaury, B. S., Shirazi, S. A. & McKinney, B. A. 2018 Random forest regression prediction of solid particle erosion in elbows. *Powder Technology* **338**, 983–992.

First received 11 August 2019; accepted in revised form 21 May 2020. Available online 18 August 2020