Real-Time Flood Classification Forecasting Based on k-Means++ Clustering and Neural Network

Hu Caihong

Zhengzhou University

Zhang Xueli

Zhengzhou University

Li Changqing

Shandong University

Liu Chengshuai

Zhengzhou University

Wang Jinxing

Information center of ministry of water resources

Jian Shengqi (✉ jiansq@zzu.edu.cn)

Zhengzhou University

Research Article

Keywords: flood classification forecasting, real-time classification, k-means++, BP neural network.

Posted Date: July 20th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-611718/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Water Resources Management on November 13th, 2021. See the published version at <https://doi.org/10.1007/s11269-021-03014-y>.

Real-time Flood Classification Forecasting Based on k-means++ Clustering and Neural Network

Author names and affiliations:

Hu Caihong

College of Water Conservancy Science & Engineering, Zhengzhou University,
Zhengzhou 450001, China;

Yellow River Institute for Ecological Protection & Regional Coordinated
Development, Zhengzhou 450001, China;

Zhang Xueli

College of Water Conservancy Science & Engineering, Zhengzhou University,
Zhengzhou 450001, China;

Yellow River Institute for Ecological Protection & Regional Coordinated
Development, Zhengzhou 450001, China;

Li Changqing

Shangdong survey and design institute of water conservancy, Jinan 250013, China;

Liu Chengshuai

College of Water Conservancy Science & Engineering, Zhengzhou University,
Zhengzhou 450001, China;

Yellow River Institute for Ecological Protection & Regional Coordinated
Development, Zhengzhou 450001, China;

Wang Jinxing

Information Center of Ministry of Water Resources, Beijing 100032, China;

Jian Shengqi

State Key Laboratory of Soil Erosion and Dryland Farming on the Loess Plateau ,
Institute of Soil and Water Conservation, CAS & MWR, Yangling 712100 , China;

College of Water Conservancy Science & Engineering, Zhengzhou University,
Zhengzhou 450001, China;

Corresponding author:

Name: Jian Shengqi

E-mail: jiansq@zzu.edu.cn

1 **Abstract:** Floods are one of the most dangerous disasters that affect human beings.
2 Timely and accurate flood forecasting can effectively reduce losses to human life and
3 property as well as improve the utilization rate of flood resources. In this study, an F-
4 K++ real-time flood classification forecasting framework based on *k-means++*,
5 principal component analysis, and backpropagation (BP) neural network was
6 constructed using the Jingle sub-basin, a tributary of the Yellow River Basin, as the
7 study area. The model parameters of different flood types were obtained, and a flood
8 classification and predictions were made based on the mature *M-EIES* model applied
9 in the study area. The relationship between the hydrological model parameters and
10 hydrological characteristics of the different flood types was analyzed. The results
11 show that the flood real-time classification framework established in this study can be
12 used for flood classifications and predictions in real time. The parameters of the flood
13 classification and prediction model were consistent with the characteristics of the
14 flood events. Compared with unclassified predictions, the accuracy of the
15 classification and prediction models significantly improved. The Nash coefficient
16 increased by 5%–11.62%, the relative error of the average flood peak was reduced by
17 6.08%–12.7%, the relative error of average flood volume was reduced by 5.74%–
18 8.07%, and the time difference of the average peak was reduced by 43%–66%. The
19 methodology proposed in this study can identify extreme flood events and provide
20 scientific support for flood classifications and predictions, flood control and disaster
21 reduction in river basins, and the efficient utilization of water resources.

Key words: flood classification forecasting; real-time classification; *k-means++*; BP neural network.

1. INTRODUCTION

Floods have always been one of the most dangerous disasters affecting human health. In addition, with the increase in global temperatures and the intensification of human activities, the frequency of flood disasters is gradually increasing, which poses a considerable threat to human life, property, and ecosystems (Moore et al., 2005). However, flood forecasting has always been a challenging topic in hydrology (UNPD, 2005). The hydrological model is a generalization of the law of the water cycle (Gupta et al., 1998), and it cannot be completely consistent with the actual water cycle. There are many factors that affect forecast results (Arkesteijn and Pande, 2013; Gourley and Vieux, 2006; Wagener and Gupta, 2005), among which the model parameters are an important part of the hydrological model of a watershed and have certain physical significance. The rationality of the parameter settings directly affects the final forecast results (Benke et al., 2008). Optimization methods are usually used to calibrate the model parameters based on historical hydrological data (Duan et al., 2004; Moradkhani et al., 2005). Owing to the inevitable uncertainty of the model parameters, it is difficult to obtain a set of optimal parameters regardless of which optimization method is adopted. If a set of hydrological forecast model parameters is used to forecast a flood in an entire basin, certain errors will be generated.

42 To address these problems, a series of studies have been performed both nationally
43 and globally in the recent years (Onema et al., 2012; Minglei et al., 2010; Yin et al.,
44 2007). In the same area of the flood have similar flood characteristics (Gan and
45 Burges, 2006; Sikorska et al., 2015), these similar characteristics would be divided
46 into several categories (Stein et al., 2020; Tarasova et al., 2019), flood, and for
47 different categories with different parameters of the model or the flood forecast. Thus,
48 improving the prediction accuracy of clustering technology is an effective method,
49 and this approach has been successfully applied in flood forecasting, achieving
50 important results (Jain, 2010; Kaufman and Rousseeuw, 2005). The clustering
51 algorithms of Minglei et al. (2010) and Wei and Peng (2015) were used to classify and
52 forecast floods and thus improve the accuracy of classification flood forecasting.
53 Additional clustering approaches utilize principal component analysis to determine
54 similar features (Kaufman and Rousseeuw, 2005; Sivakumar et al., 2013) and data-
55 driven artificial neural network methods (Hong et al., 2003; Hu et al., 2018; Kumar et
56 al., 2013).

57 The products of this study are as follows: (1) a real-time flood classification
58 framework based on *k-means++*, factor analysis, and a BP neural network was
59 established to classify long time series flood events, and hydrological model
60 parameters of corresponding flood categories were obtained through an intelligent
61 optimization algorithm. (2) Based on the mature M-EIES model applied in the study
62 area, flood classifications and predictions are produced. (3) The relationship between

the hydrological model parameters and hydrological characteristics of different types of floods is analyzed.

2. MATERIAL AND METHODS

2.1. Study area

The Fenhe River, located in the middle reaches of the Yellow River, is the second largest tributary of the Yellow River and is located in the middle and southwest of Shanxi Province. This area in the middle latitudes has a semi-arid and semi-humid temperate continental monsoon climate, with a monthly average temperature of 4–13 °C and a maximum water surface evaporation of 1120 mm. The average annual precipitation in this region is 503 mm, 60% of which occurs from July to September, and the average annual runoff of the basin is $2.28 \times 10^9 \text{ m}^3$. In the middle and upper reaches of the Fenhe River Basin, there are sand shrub mountains on both sides of the main channel, metamorphic rock shrub mountains in the southeast of the basin, and loess hilly terraces on both sides of the main channel in the lower reaches of the basin.

The main stream of the Fenhe Jingle sub-basin is 83.9 km in length and 2799 km² in area, with an average slope of 6.7% and a basin shape coefficient of 0.398. Loess coverage in the control basin is relatively concentrated, with a thickness of approximately 30–50 m. The lithology of the loess is loose, and cracks and caves are relatively developed. Loess can easily absorb and conduct water, but its water-holding capacity is relatively poor. The average annual precipitation of the basin controlled by

Jingle Station is 497.85 mm, an average annual maximum 24 h rainstorm produces between 50–55 mm, the single station maximum 24 h rainstorm value is 109.6 mm, the average annual maximum peak flood discharge is 594 m³/s, and the measured maximum peak flood discharge is 2230 m³/s. In addition, there are few coal mines, minimal water diversions, and no water conservancy projects in the Jingle sub-basin. In recent years, urbanization has accelerated in the Jingle sub-basin has accelerated, and urban land has expanded annually. However, the main land use types in this region are still woodlands and arable land (Fig. 1).

2.2. Data collection

In this study, the runoff and flood extract data of annual runoff days from 1971 to 2018 were collected from the Jingle Hydrological Station in the Jingle sub-watershed. The daily rainfall data from 17 rain measurement stations (Xinmiao, Xinmin Gushan, and Gaoshiya stations) in the sub-basin, from 1971 to 2018, were unified into 1 h time step area average rainfall and evaporation data; these data were obtained via the Tyson polygon method. Based on the flood event identification methods of Kundzewicz and Robson (2004) and Sikorska et al. (2015), a total of 88 flood events were selected.

2.3. Flood classification methods

Based on factor analysis and the *k-means++* clustering algorithm, this study combined the SEE function and BP neural network model to construct a F-K++ real-time flood classification and prediction framework (Fig. 2). The process of the F-K++

103 real-time flood classification method mainly includes three modules: sample
104 processing, cluster analysis, and flood classification.

105 (a) Data Disposal Module

106 First, the flood characteristic indexes that are closely related to floods are extracted
107 according to the floods' information. After considering and analyzing the factors
108 affecting floods in the middle reaches of the Yellow River, flood characteristic indices
109 that were easily obtained in the early rainfall period and were significantly related to
110 the flood formation were selected, as follows: duration of precipitation T , maximum 1
111 h precipitation P_{1h} , maximum 3 h rainfall P_{3h} , mean precipitation intensity i , total
112 precipitation P , preceding affected precipitation P_p (Heggen and Richard, 2001),
113 rising flow Q_0 , and precipitation center D , denoted as $X_1, X_2, \dots, X_8$, respectively.

114 Second, the factor analysis method was used to reduce the dimension of the flood
115 characteristic indices extracted from the flood process of the basin, all the extracted
116 characteristic indexes were converted into several unrelated common factors, and the
117 common factors of each flood were calculated (Meredith, 1993).

118 (b) Cluster analysis module

119 The *k-means++* algorithm was used to cluster the extracted flood common factors
120 (Arthur and Vassilvitskii, 2007; Likas et al., 2002; Nielsen and Nock, 2013). This
121 process is as follows: (1) Take one center c_1 , chosen uniformly at random from X . (2)

122 Take a new center c_i , choosing $x \in X$ with probability $\frac{D(x)^2}{\sum_{x \in X} D(x)^2}$. (3) Repeat Step (1)
 123 until k centers are taken together. (4) For each $i \in \{1, \dots, k\}$ (Merz. and Blöschl,
 124 2008) set the cluster C_i to be the set of points in X that are closer to c_i than they are to
 125 c_j for $j \neq i$. (5) For each $i \in \{1, \dots, k\}$, c_i is the center of mass of all points in C_i :
 126 $c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$ (6) Repeat Steps (4) and (5) until C no longer changes.

127 The reasonable value of the flood classification number k is the key to ensuring a
 128 reasonable flood classification. In this study, the sum of error squares (SSE) was used
 129 to determine the optimal value of the flood classification number k :

$$130 \quad SSE = \sum_{i=1}^k \sum_{x \in S_i} (d(c_i, x))^2 \quad (1)$$

131 where d is the Euclidean distance of two sample points in the Euclidean space, k is the
 132 number of categories, x is the sample points, c_i is the center point of each cluster, and
 133 S_i represents the set of sample points in class i .

134 (c) Real-time classification module

135 A back propagation (BP) neural network is a multi-layer feed-forward neural network
 136 that consists of an input layer, hidden layer, and output layer, and is trained according
 137 to the error back propagation algorithm. It is usually used as a classifier with tutors
 138 (Benediktsson and Swain, 1990; Heermann and Khazenie, 1992). In this study, eight
 139 characteristic indexes ($T, P_{1h}, P_{3h}, P, i, P, P_p, Q_0, D$) of field floods were used as the
 140 training sample input of the BP neural network. A part of each category of

classification results obtained by *k-means++* clustering was randomly selected as the training sample output to train the BP neural network, and the remaining part was used to verify the real-time classification effect of the BP neural network.

2.4. M-EIES model

In utilizing the Xin–Zhejiang River model (Zhao, 1992), the *M-EIES* model of full reservoir-superpermeability was improved by our team (Hu et al., 2005; Wen et al., 2020). The improved comprehensively considered the complex mechanism of the combined action of superpermeability and runoff generation, and simulated the flood process by analyzing the main runoff generation modes of different rainfall processes. The model demonstrated a strong applicability for flood simulations under complex underlying surface conditions.

(a) Model structure

The *M-EIES* model consists of five parts: evapotranspiration, watershed unit runoff, watershed unit confluence, underground runoff confluence, and watershed confluence. The *M-EIES* model organically combines the area distribution curve of the basin storage capacity and the area distribution curve of the basin infiltration capacity, and comprehensively considers the infiltration capacity, soil moisture content, and distribution. The distribution curve of infiltration capacity in a given period adopts an *m*-degree parabola:

$$\beta = 1 - \left(1 - \frac{F'_{\Delta t}}{F'_{m\Delta t}} \right)^m \quad (2)$$

where β is the relative area, $F'_{\Delta t}$ is the lower body capacity (mm) of a certain point in the basin, m is the experience index, $F'_{m\Delta t}$ is the maximum infiltration capacity at the point (mm) within the time period t , and Δt represents the time period t .

According to the area distribution curve of the infiltration capacity, the average amount of infiltration water in the basin was $F_{m\Delta t}$.

$$F_{m\Delta t} = \int_0^t F'_{\Delta t} d\beta = \int_0^{F'_{m\Delta t}} (1 - \beta) dF'_{\Delta t} = \frac{F'_{m\Delta t}}{m+1} \quad (3)$$

The water storage capacity distribution curve of the basin adopts the n -th empirical parabola:

$$\alpha = 1 - \left(1 - \frac{W'}{W'_m} \right)^n \quad (4)$$

where W' is the water storage capacity value at a certain point (mm), W'_m is the maximum water storage capacity of the basin (mm), n is the experience index, and α is the relative area.

The average storage capacity of the basin (W_m) was calculated according to the area distribution curve of the basin storage capacity:

$$W_m = \int_0^{W'_m} (1 - \alpha) dW' = \frac{W'_m}{n+1} \quad (5)$$

The basin area is divided into two parts, $1-\beta$ and β , according to the area distribution curve $F'_{\Delta t} \square \beta$ of the basin infiltration capacity. In terms of the runoff producing area of the relative area $1-\beta$, the soil infiltration capacity of the basin is greater than the rainfall intensity. Currently, the basin does not produce runoff. The basin area rainfall on the watershed distribution curve $W' \square \alpha$ is divided into parts: α and $1-\alpha$ on the relative area of alpha, including the entirety of the basin rainfall included in the runoff, and the relative area of $1-\alpha$ alpha, including basin rainfall first used to supplement the soil water deficit, which can result in runoff. The rainfall runoff yield model at this time is the accumulation of runoff.

(b) Model parameters and calibration method

The model has 17 main parameters, among which evapotranspiration calculation parameters include four runoff generation parameters, five sub-water source parameters, four confluence parameters and four water source parameter (Table 1). In this study, the manual trial and error method and SCE-UA algorithm (Duan et al., 1994) were used to optimize and calibrate the model parameters.

(c) Model evaluation method

The Nash coefficient, relative error, and time difference of the flood peak were used to evaluate the optimal value of periodic parameters of parameter rate and the effect of the classified flood prediction. NSE was used to verify the credibility of the model (Nash and Sutcliffe, 1970), expressed as:

$$NSE = 1 - \frac{\sum_{m=1}^n (Q_m^{sim} - Q_m^{obs})^2}{\sum_{m=1}^n (Q_m^{obs} - \bar{Q}_m^{obs})^2} \quad (5)$$

196

197 where Q_m^{obs} is the measured value of runoff, Q_m^{sim} is the simulated value of runoff, $\bar{Q}_m^{obs}$
 198 is the average value of the measured runoff, and n is the sequence length. The value of
 199 NSE is between 1 and ∞ ; the closer it is to 1, the better the performance and the
 200 higher the reliability of the model. A value close to 0 indicates that the simulation
 201 results are close to the average value of observation, and the overall results are
 202 credible, but the process simulation error is large; a value less than 0 indicates that the
 203 result is not credible (Moriasi, 2007).

204 The relative error is expressed as:

$$RE = \frac{\sum_{m=1}^n (Q_m^{obs} - Q_m^{sim})}{\sum_{m=1}^n Q_m^{obs}} \quad (6)$$

205

206 The optimal RE value is 0. A positive value means that the simulated value is less
 207 than the observed value, whereas a negative value means that the simulated value is
 208 greater than the observed value. The greater the absolute value of RE, the greater the
 209 simulation error of the model (Gupta et al., 1999). In this study, the flood peak
 210 discharge relative error and flood volume relative error evaluation models were used
 211 to predict the performance.

212 There is a time lag between the peak of the flood:

$$E = t_m^{obs} - t_m^{sim} \quad (7)$$

where t_m^{obs} represents the observed value of the occurrence time of the flood peak, t_m^{sim} represents the simulated value of this time, and E is the error between the two (Khu and Madsen, 2005).

3. RESULTS

3.1. Flood event analysis

All the flood events occurred from May to October (Fig. 3a), and 78.41% of the flood events occurred in July and August in summer. The precipitation of each flood event ranged from 0.98–154 mm, and the rainfall intensity of 85.23% of the flood events ranged from 5–29.9 mm/h (Fig. 3b).

3.2. Real-time flood classification

(a) Clustering classification

Based on the eight flood characteristic indicators (T , P_{1h} , P_{3h} , P , i , P_p , Q_0 , D) (Fig. 5(a)) and the new flood factors obtained via factor analysis, the *k-means++* clustering method was applied to cluster flood events, and five types of floods were classified (Fig. 4 and Fig. 5(b)).

According to the statistics of the characteristics of various flood events (Fig. 6):

The first type of flood (C1) had two flood events, with an average rainfall duration of 55.5 h and a large total rainfall, which were second only to those of the third type

232 (C3). The average rainfall of the earlier stage was 99.56 mm, and the average flood
233 volume was 12357.06 m³, which were second only to those of C3. The peak flood
234 discharge was 289.5 m³/s, denoting a long duration rainfall event.

235 The second type of flood (C2) had 32 flood events, with a median rainfall duration of
236 16.5 h, a small total rainfall duration of 11.05 mm, a median rainfall total of 109.03
237 mm, a median flood volume of 2765.59 m³ and a median peak flow of 252.5 m³/s,
238 which is a short duration rainfall event.

239 The third type of flood (C3) had only one flood event, but it had the longest rainfall
240 duration of 102 h, a maximum total rainfall total of 153.83 mm, and a preceding
241 rainfall total of 123.06 mm. Its flood volume and peak discharge were the largest, at
242 63448.4 m³/s and 948 m³/s, respectively. C3 represents an extreme flood event caused
243 by heavy rainfall over a long duration.

244 The fourth type of flood (C4) had 31 flood events, with a median rainfall duration of
245 16 h, a small total rainfall of 10.27 mm, a median rainfall of 59.63 mm, a median
246 flood volume of 958.39 m³, and a median peak flow of 132 m³/s. C4 represents a
247 flood with a short rainfall duration, similar to C2, although the latter affected more
248 rainfall in the early stage.

249 The fifth type of flood event (C5) had 22 flood events, with a median rainfall duration
250 of 17.5 h, a large total rainfall of 30.3 mm, a median rainfall of 73.29 mm, a median

flood volume of 2618.02 m³, and a median peak flow of 300.93 m³/s. C5 represents a short duration, heavy rainfall flood.

(b) Real-time classification

The BP neural network flood classification model used in this study contained a network structure of three layers. After a trial calculation, the number of nodes in the hidden layer was eight, and the rainfall centers of the first 80 floods in the historical flood data of Jingle Station were selected. The data for the eight variables, such as rainfall in the early stage, were input as training samples for the BP neural network, and the number of categories obtained via clustering was output as training samples to train the BP neural network. The remaining flood characteristic data were input into the trained BP neural network for verification (Table 2).

The real-time classification results of BP network were mostly consistent with the classification results of clustering analysis, and the classification accuracy reached 100%. The maximum error of the detection results was 0.083, and the relative error was 2.1%. Therefore, the BP neural network is fully capable of correctly classifying floods using the K-F++ real-time flood classification method, which presents the flood events as a mixed mechanism and realizes the real-time classification of floods in the basin according to the rainfall data and other information.

3.3. Classified flood forecasting

(a) Parameter calibration

First, the scope of parameter optimization was preliminarily set according to the physical significance of the parameters in the model. Then, combined with three automatic optimization algorithms, the parameters in each reservoir-superpermeability model were optimized and adjusted individually using historical flood data and classified historical flood data, and the optimal value of each parameter was obtained (Table 3).

The parameter W_m represents the degree of drought in the basin. The drier the earlier stage, the greater the W_m value is; the W_m values were the largest in C4 and the smallest in C3. Regarding precipitation prior to the flood, C4 was the smallest and C3 was the largest. Parameter fc is the most sensitive parameter, and the increase in parameter value will reduce the total production flow; the fc values were the largest in C4 and the smallest in C3. The parameter k represents the soil permeability coefficient, which is a sensitive parameter. The increase in the k value will reduce the total yield flow. The k values were the largest in C4 and the smallest in C3. The parametric characteristics were consistent with those of the flood events. The parameter specificity increased compared to the flood classification, and it was more consistent with the flood event characteristics of the category.

(b) Classified flood forecasting

Among the five categories of flood events controlled by Jingle Station, C1 had two flood events: one for event calibration parameters and simulation, and the other for

291 verification. C3 had a flood event that only included parameter calibration and
292 simulation. Ten flood events were randomly selected for the other three types of flood
293 events, of which seven were used for model parameter calibration and three were used
294 to verify the simulation of the model. The simulation results (Figs. 7 and 8) show that,
295 compared with the traditional flood prediction framework with a single parameter set,
296 the real-time flood classification prediction framework using a multi-parameter set
297 can more accurately predict the peak arrival time of total runoff and the simulated
298 flood process.

299 Compared with the simulation accuracy before classification, the average Nash
300 efficiency coefficient after classification of C1 increased by 8.8% to 0.74, the average
301 flood peak relative error was reduced by 7.68%, the average flood volume relative
302 error was reduced by 8.07%, and the average peak occurrence time difference was
303 reduced by 66%.

304 Compared with the simulation accuracy before the classification, the average Nash
305 efficiency coefficient after the classification of C2 increased by 5.6% to 0.75, the
306 average flood peak relative error was reduced by 9.15%, the average flood volume
307 relative error was reduced by 5.74%, and the average peak occurrence time difference
308 was reduced by 60%.

309 Compared with the simulation accuracy before classification, the average Nash
310 efficiency coefficient after classification of C3 increased by 11.6% to 0.77, the

average flood peak relative error was reduced by 12.7%, the average flood volume relative error was reduced by 6.82%, and the average peak occurrence time difference was reduced by 50%.

Compared with the simulation accuracy before classification, the average Nash efficiency coefficient after classification of the C4 increased by 10% to 0.72, the average flood peak relative error was reduced by 8.94%, the average flood volume relative error was reduced by 6.68%, and the average peak occurrence time difference was reduced by 58%.

Compared with the simulation accuracy before classification, the average Nash efficiency coefficient after classification of C5 increased by 5% to 0.73, the average flood peak relative error was reduced by 6.08%, the average flood volume relative error was reduced by 6.58%, and the average peak occurrence time difference was reduced by 43%.

4. DISCUSSION

4.1. Real-time flood classification

According to the clustering classification results, there are five flood types in the study basin: long rainfall, short rainfall, flash, high preceding precipitation/long rainfall, and high preceding precipitation/long rainfall floods. This finding is consistent with the event statistics and is even somewhat consistent with the flood types proposed by Merz and Blöschl (2003) (long rain, short rain, flash, rain-on-snow,

and snowmelt floods). However, the precipitation in the Jingle watershed mainly occurs in summer, with less rainfall in winter and no snow cover, and there are no rain-snow and snowmelt floods. In addition, the cluster classification considers precipitation that occurred prior to the flood events and separates events with large rainfall in the early stage into different rain types. However, because of the complex relationship between various indicators, similar rainfall in the earlier period is also divided into different types, and the specific reasons for this division must be discussed in further studies.

The calculation process of clustering classification is cumbersome, and the prediction efficiency of real-time data after a clustering classification is low. Using the clustering and classification results of historical data to train a BP neural network can solve this problem and ensure real-time data classification. However, the main limitations of the real-time flood classification simulation are as follows: first, a classification of river basin floods requires historical flood event data with a sufficiently long time series to reflect all the characteristics of river basin floods. Real-time rainfall and other information should also be obtained. For areas with incomplete data or no data, it is difficult to classify floods directly using a real-time classification method, and thus the accuracy of the flood simulation and prediction cannot be improved. Second, the selection of factors is limited by the existing cognition of the causes of floods, and the static attribute influencing factors of the basin are not considered. Thus, the factors affecting the flood process cannot be fully reflected, and the genetic mechanism of

flooding cannot be fully reflected, thereby affecting the accuracy of flood simulations.

Third, the calculation process of flood classification is cumbersome as it requires a large amount of data, and flood classifications are time consuming. For a large basin area with many flood events, the amount of data is overly large, and the calculation time is excessive. Therefore, classification software can be constructed by coupling all modules of real-time classification.

4.2. Classified flood forecasting

The causes of flooding are complex, and the interaction mechanisms of individual factors vary under different climatic conditions. The underlying surface conditions in the basin are also changing owing to the influence of human activities (Nied et al., 2014; Merz and Blöschl, 2008). The existing models only have a set of parameters and cannot reflect different flood generation mechanisms. The K-F++ real-time flood classification method presents flood events as a mixed mechanism and ensures a real-time flood classification of the watershed based on rainfall data and other information.

The existing model only has a set of parameters that cannot reflect the real-time flood classification method of K-F++ with different flood generation mechanisms. Flood events are represented as a mixed mechanism, and a real-time flood classification is created based on rainfall data and other information. A classification simulation can select matching parameters according to different types of flood events and obtain

model parameters reflecting the characteristics of different flood event types to improve the prediction accuracy.

The accuracy of the C3 flood type improved the most. C3 represents an extreme flood event caused by a heavy rainfall intensity during and before the event and a long duration. The traditional forecast method cannot reflect the characteristics of extreme floods, and the prediction error is large. However, in the classification and prediction model, the W_m index, fc , and k , are significantly different from the traditional prediction parameters; these parameters are mostly in accordance with the mechanism of extreme flood generation to ensure a higher prediction accuracy.

Therefore, the real-time flood classification and prediction approach is superior in extreme flood predictions and simulations. Concurrently, the main characteristics of extreme floods can be determined through analyzing the characteristics of different flood types. Moreover, characteristics that are similar to those of historical extreme floods can be quickly captured in the initial rainfall data to preliminarily classify and determine the flood types, thereby providing an important early reference for flood management and disaster prevention and reduction. Therefore, the real-time classified flood prediction framework provides an efficient and robust tool for real-time flood prediction in watersheds, especially for extreme flood events.

5. CONCLUSIONS

In this study, a real-time flood classification framework based on *k-means++*, principal component analysis, and a BP neural network was established, and the model parameters of different types of floods were obtained. The rationality and effectiveness of this method were verified using real river basins. The real-time flood classification method can classify watershed flood events in real time, and the real-time flood classification and prediction framework can classify floods in real time using specific indicators. Selecting the flood model parameters that can reflect different generation mechanisms and match with the type to forecast the flood may improve the prediction accuracy, especially for large-scale disaster floods. However, the selection of the flood characteristic index is the key step of this method, and needs to be further studied and discussed in future research.

Institutional Review Board Statement: No applicable.

Informed Consent Statement: No applicable.

Data Availability Statement: The data presented in this study is available on request from the corresponding author.

Comments and suggestions. We also thank my other colleagues for their valuable comments and suggestions that have helped improve the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Acknowledgements. This project was supported by National key research priorities program of China (2019YFC1510703), National Natural Science Foundation of China (51979250), Key projects of National Natural Science Foundation of China (51739009), Natural Science Foundation of Henan Province (212300410413), Young Talents Promotion Project of Henan Province(2021HYTP030), State Key Laboratory of Soil Erosion and Dryland Farming on the Loess Plateau Foundation (A314021402-2021011).

REFERENCES

- 416 Arkesteijn, L., Pande, S., 2013. On hydrological model complexity, its geometrical
417 interpretations and prediction uncertainty. *Water Resources Research* 49, 7048-
418 7063.
- 419 Arthur, D., Vassilvitskii, S., 2007. k-means++: the advantages of careful seeding. In
420 Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete
421 Algorithms, New Orleans, SIAM, pp. 1027-1035.
- 422 Benediktsson, J.A., Swain, P.H., 1990. Neural Network Approaches Versus Statistical
423 Methods In Classification Of Multisource Remote Sensing Data. *IEEE*
424 *Trans.geosci. & Remote Sensing* 28, 489-492.
- 425 Benke, K.K., Lowell, K.E., Hamilton, A.J., 2008. Parameter uncertainty, sensitivity
426 analysis and prediction error in a water-balance hydrological model.
427 *Mathematical and Computer Modelling* 47, 1134-1149.
- 428 Duan, Q., Gupta, H.V., Sorooshian, S., Rousseau, A.N., Turcotte, R., 2004.
429 Calibration of watershed models.
- 430 Duan, Q., Sorooshian, S., Gupta, V.K., 1994. Optimal use of the SCE-UA global
431 optimization method for calibrating watershed models. *Journal of Hydrology*
432 158, 265-284.
- 433 Gan, T.Y., Burges, S.J., 2006. Assessment of soil-based and calibrated parameters of
434 the Sacramento model and parameter transferability. *Journal of Hydrology* 320,
435 117-131.
- 436 Gourley, J.J., Vieux, B.E., 2006. A method for identifying sources of model
437 uncertainty in rainfall-runoff simulations. *Journal of Hydrology* 327, 68-80.
- 438 Gupta, H.V., Sorooshian, S., Yapo, P.O., 1998. Toward improved calibration of
439 hydrologic models: Multiple and noncommensurable measures of information.
440 *Water Resources Research* 34, 751– 763.
- 441 Gupta, H.V., Sorooshian, S., Yapo, P.O., 1999. Status of Automatic Calibration for

442 Hydrologic Models: Comparison With Multilevel Expert Calibration. Journal of
443 Hydrologic Engineering 4, 135-143.

444 Heermann, P.D., Khazenie, N., 1992. Classification of Multispectral Remote Sensing
445 Data Using a Backpropagation Neural Network. IEEE Transactions on
446 Geoscience and Remote Sensing 30, 81-88.

447 Heggen, Richard, J., 2001. Normalized Antecedent Precipitation Index. Journal of
448 Hydrologic Engineering 6, 377-381.

449 Hong, Y., Hsu, K., Sorooshian, S., 2003. Precipitation Estimation from Remotely
450 Sensed Information using ANN-Cloud Classification System. journal of applied
451 meteorology.

452 Hu, C., Guo, S., Xiong, L., Peng, D., 2005. A modified Xinanjiang model and its
453 application in northern China. Hydrology Research 36, 175-192.

454 Hu, C., Wu, Q., Li, H., Jian, S., Li, N., Lou, Z., 2018. Deep Learning with a Long
455 Short-Term Memory Networks Approach for Rainfall-Runoff Simulation. Water
456 10.

457 Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recognition
458 Letters 31, 651-666.

459 Kaufman, L., Rousseeuw, P.J., 2005. Finding Groups in Data: An Introduction to
460 Cluster Analysis. Wiley.

461 Khu, S.T., Madsen, H., 2005. Multiobjective calibration with Pareto preference
462 ordering: An application to rainfall-runoff model calibration. Water resources
463 research 41, p.W03004.03001-W03004.03014.

464 Kileshye Onema, J.M., Taigbenu, A.E., Ndiritu, J., 2012. Classification and flow
465 prediction in a data-scarce watershed of the Equatorial Nile region. Hydrology
466 and Earth System Sciences 16, 1435–1443.

467 Kundzewicz, Z.W., Robson, A.J., 2004. Change detection in hydrological records — a
 468 review of the methodology. *Hydrological sciences journal* 49, 7-19.

469 Likas, A., Vlassis, N., Verbeek, J.J., 2002. The Global K-Means Clustering Algorithm.
 470 *Pattern Recognition* 36, 451-461.

471 Manuela Nied, Tobias Pardowitz, Katrin Nissen, Uwe Ulbrich, Yesheatesfa
 472 Hundecha, Bruno Merz, 2014. On the relationship between hydro-meteorological
 473 patterns and flood types. *Journal of Hydrology* 519, 3249-3262.

474 Meredith, W., 1993. Measurement Invariance, Factor Analysis and Factorial
 475 Invariance. *Psychometrika* 58, 525-543.

476 Merz, R., Blöschl, G., 2003. A process typology of regional floods. *Water Resources*
 477 *Research* 39.

478 Merz., R., Blöschl, G., 2008. Flood frequency hydrology: 1. Temporal, spatial, and
 479 causal expansion of information. *Water Resources Research* 44, 8432.

480 Minglei, REN, and, Bende, WANG, and, Qiuhua, LIANG, and, Guangtao, 2010.
 481 Classified real-time flood forecasting by coupling fuzzy clustering and neural
 482 network. *International Journal of Sediment Research* 25, 134-148.

483 Moore, R.J., Bell, V.A., Jones, D.A., 2005. Forecasting for flood warning. *Comptes*
 484 *rendus - Géoscience* 337, 203-217.

485 Moradkhani, H., Sorooshian, S., Gupta, H.V., Houser, P.R., 2005. Dual state-
 486 parameter estimation of hydrological models using ensemble Kalman filter.
 487 *Advances in Water Resources* 28, 135-147.

488 Moriasi, D., 2007. Model evaluation guidelines for systematic quantification of
 489 accuracy in watershed simulations. *Transactions of the Asabe* 50, 885–900.

490 Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models
 491 part I — A discussion of principles - ScienceDirect. *Journal of Hydrology* 10,

492 282-290.

493 Nielsen, F., Nock, R., 2013. Total Jensen divergences: Definition, Properties and k-
494 Means++ Clustering. Computer Science.

495 Senthil kumar A.R., Manish Kumar Goyal, Ojha C.S.P., Singh R.D., Swamee P. K. ,
496 K., N.R., 2013. Application of ANN, Fuzzy Logic and Decision Tree Algorithms
497 for the Development of Reservoir Operating Rules. Water Resources
498 Management 27, 911-925.

499 Sikorska, A.E., Viviroli, D., Seibert, J., 2015. Flood-type classification in
500 mountainous catchments using crisp and fuzzy decision trees. Water Resources
501 Research 51, 7959-7976.

502 Sivakumar, B., Singh, V.P., Berndtsson, R., Khan, S.K., 2013. Catchment
503 Classification Framework in Hydrology: Challenges and Directions. Journal of
504 Hydrologic Engineering 20, A4014002.

505 Stein, L., Pianosi, F., Woods, R., 2020. Event-based classification for global study of
506 river flood generating processes. Hydrological Processes 34, 1514– 1529.

507 Tarasova, L., Merz, R., Kiss, A., Basso, S., Blschl, G., Merz, B., Viglione, A., Pltner,
508 S., Guse, B., Schumann, A., 2019. Causative classification of river flood events.
509 Wiley Interdisciplinary Reviews: Water 6.

510 UNPD, 2005. Reducing Disaster Risk: A Challenge for Development. United Nations
511 Development Programme, Bureau for Crisis Prevention and Recovery.

512 Wagener, T., Gupta, H.V., 2005. Model identification for hydrological forecasting
513 under uncertainty. Stochastic Environmental Research & Risk Assessment 19,
514 378-387.

515 Wei, X., Peng, Y., 2015. Research on classified real-time flood forecasting framework
516 based on K-means cluster and rough set. Water Science & Technology A Journal
517 of the International Association on Water Pollution Research 71, 1507-1515.

518 Wen, Y., Hu, C., Zhang, G., Jian, S., 2020. Response of the parameters of excess
 519 infiltration and excess storage model to land use cover change. Journal of
 520 Hydrology and Hydromechanics 68, 99-110.

521 Yin, X.R., Zhang, X., Xia, J., 2007. Classification-based flood forecasting model
 522 using artificial neural networks. Journal of Sichuan University(Engineering
 523 Science Edition).

524 Zhao, R.J., 1992. The Xinanjiang Model Applied in China. Journal of Hydrology 135,
 525 371--381.

526

528

529

530

531

Fig. 1. Topographic map of the Jingle control catchment. (a) Description of Yellow River in China; (b) Description of Jingle control catchment of Yellow River; (c) The control catchment of Jingle station in Fen River and distribution of rainfall gauge stations.

Fig. 2. K-F++ real-time classification framework

Fig. 3 Distribution of flood event features.(a) Seasonal characteristics of flood events;(b) Rainfall intensity characteristics of flood events.

Fig. 4 Flood clustering results

540

541

542 Fig. 5 Flood event characteristic indexes. (a) Values of flood event characteristic factors;(b) Values
543 of flood event characteristic factors after classification.

Fig. 6 Classified flood event features map.

Fig. 7 Diagram of model evaluation results.

550

551 Fig. 8 Classification Simulation of Five Types of Flood Events, Conventional and observed flood
552 process lines.

Table 1. The meanings of the *M-EIES* model parameters.

Parameters	Parameters meaning	Parameters	Parameters meaning
Wm	Average catchment storage capacity	SM	Free water storage capacity
WUm	Water storage capacity of upper watershed	EX	Index of free water storage capacity curve
WLm	Water storage capacity of the lower watershed	CI	Outflow coefficient of soil
n	Index of the water storage capacity distribution curve	CG	Underground runoff coefficient
m	Index of infiltration capacity distribution curve	CKI	Soil flow regression coefficient
C	Deep evaporation coefficient	CKG	Subsurface runoff regression coefficient
CKE	Conversion coefficient of evaporation	CN	The number of river sections divided by channel calculation
fc	Stable infiltration rate	X	Dimensionless flow specific gravity factor and storage constant
k	Soil permeability coefficient		

555

Table 2. BP Neural network detection results table

Event No.	Cluster No.	Output value	Event No.	Cluster No.	Output value
81	2	2.045	85	4	3.943
82	4	3.917	86	2	2.009
83	4	4.019	87	2	1.961
84	4	4.053	88	2	2.034

556

557

Table 3. parameters set for classified flood forecasting and traditional flood forecasting

558

in the *M-EIES* model.

Parameters	Unclassified	Classified				
		1	2	3	4	5
<i>Wm</i>	138.248	136.352	123.178	103.437	163.623	159.971
<i>WUm</i>	0.347	0.271	0.399	0.302	0.271	0.399
<i>WLm</i>	0.288	0.234	0.254	0.371	0.199	0.273
<i>n</i>	0.176	0.421	0.087	0.455	0.087	0.101
<i>m</i>	0.313	0.102	0.373	0.089	0.477	0.273
<i>CKE</i>	0.763	0.885	0.882	0.734	0.545	0.718
<i>C</i>	0.689	0.86	0.109	0.455	0.854	0.249
<i>f_c</i>	8.034	8.089	11.412	6.551	13.551	7.319
<i>k</i>	0.441	0.565	0.433	0.341	0.655	0.354
<i>CKG</i>	0.981	0.921	0.998	0.976	0.965	0.997
<i>CN</i>	7.792	6.455	8.416	6.541	7.732	9.416
<i>X</i>	5.211	4.777	6.968	5.888	7.21	6.528
<i>EX</i>	2.677	1.117	2.233	3.525	6.23	1.987
<i>SM</i>	13.229	8.718	11.812	15.324	16.555	11.331
<i>CKI</i>	0.996	0.985	0.998	0.999	0.995	0.997
<i>CI</i>	0.431	0.434	0.171	0.688	0.431	0.211
<i>CG</i>	0.222	0.199	0.363	0.498	0.112	0.337

559