Research on classified real-time flood forecasting framework based on K-means cluster and rough set

Wei Xu and Yong Peng

ABSTRACT

This research presents a new classified real-time flood forecasting framework. In this framework, historical floods are classified by a K-means cluster according to the spatial and temporal distribution of precipitation, the time variance of precipitation intensity and other hydrological factors. Based on the classified results, a rough set is used to extract the identification rules for real-time flood forecasting. Then, the parameters of different categories within the conceptual hydrological model are calibrated using a genetic algorithm. In real-time forecasting, the corresponding category of parameters is selected for flood forecasting according to the obtained flood information. This research tests the new classified framework on Guanyinge Reservoir and compares the framework with the traditional flood forecasting method. It finds that the performance of the new classified framework is significantly better in terms of accuracy. Furthermore, the framework can be considered in a catchment with fewer historical floods.

Key words | conceptual hydrological model, flood classification, K-means cluster, real-time flood forecasting, rough set

Wei Xu
College of River and Ocean Engineering,
Chongqing Jiaotong University,
Chongqing 400074,
China;
National Engineering Research Center for Inland
    Waterway Regulation,
Chongqing Jiaotong University,
Chongqing 400074,
China;
and
State Key Laboratory of Hydraulics and Mountain
    River Engineering,
Sichuan University,
Sichuan 610065,
China

Yong Peng (corresponding author)
School of Hydraulic Engineering,
Dalian University of Technology,
Dalian 116024,
China
E-mail: pengyong@dlut.edu.cn

INTRODUCTION

Floods are one of the most common natural hazards in China and throughout the world. Floods are influenced by many factors, including physical geography, hydrology, meteorology and human activity. Thus, floods are difficult to simulate due to spatial and temporal variability of precipitation and geomorphology (Xu *et al.* 2013); great complexity and uncertainty are the characteristics of flood simulation and forecasting.

Traditionally, the unit hydrograph is the most basic and important tool in flood simulation (Sherman 1932; Labat *et al.* 2000). Moreover, the unit hydrograph is typically treated as having the same shape for the entire simulation (Ren & Wang 2010; Xu *et al.* 2013). In the conceptual flood forecasting model, the parameters are calibrated using an optimization algorithm with historical hydrological data. During calibration, the simulation performances of historical floods are evaluated and taken to assess the objective function of the optimization algorithm (Cheng *et al.* 2006). In this way, the characteristics of floods during calibration are represented by a single set of calibrated parameters. In addition, the single set of parameters is assumed to be associated with a catchment and to be applicable to different types of flood (Cheng *et al.* 2006). In the traditional method, the

influential factors of geomorphology and precipitation are not fully considered during flood simulation and forecasting.

In fact, the characteristics of floods, e.g. peak flow and volume, direction and velocity, are affected by geomorphology and precipitation (Rodriguez & Valded 1979; Rinaldo *et al.* 1991; Zhang *et al.* 2007). Rodriguez & Valded (1979) proposed a modified unit hydrograph to account for the varying geomorphology and precipitation in different drainage basins. Following the development of this modified unit hydrograph, classified flood forecasting methods were established (Gupta *et al.* 1980; Rinaldo *et al.* 1991; Saghafian & Julien 2002; Ren & Wang 2010). In the classified flood forecasting method, historical floods are clustered into several categories by using statistical methods or cluster models (Ma *et al.* 1997; Lu & Hou 2007). Many algorithms have been applied to classify historical floods in recent years, including the projection pursuit model, artificial neural network, fuzzy method, C-means clustering model and rough set approach (Wang *et al.* 2002; Jain & Srinivasulu 2006; Yin *et al.* 2007; Ren & Wang 2010; Xu *et al.* 2013).

These earlier studies first classified historical floods into several categories (Ren & Wang 2010) and then calibrated the flood forecasting model parameters for each category.

Using these calibrated parameters improves the accuracy of forecasted floods (Xu *et al.* 2013). However, the main limitation of this method is that only a few historical floods are included in some categories due to the limited amount of information about historical floods. Thus, these calibrated parameters display high uncertainty and lack credibility. In fact, the characteristics of precipitation (e.g. intensity, volume, and center) vary with time in a flood event, and other influential factors (e.g. soil storage capacity) also vary with time. Therefore, historical floods can be disaggregated according to time period and used in related research.

The main purpose of this paper is to investigate the performance of flood clusters with time variation and real-time flood forecasting. This paper has three goals: firstly, the K-means cluster model is employed to classify historical floods in each time period according to the influential factors of each time period; secondly, a genetic algorithm (GA) is applied to calibrate forecasting model parameters; thirdly, the rough set algorithm is established to extract the identification rules for parameter selection. This study takes the Guanyinge Reservoir basin in northeast China as a case example. In this study, historical floods in each time period are taken as study cases. The cases are defined by both condition attributes and a decision attribute. The condition attributes are weather systems (WSs), the underlying surface of the basin, the precipitation center (PC) and intensity during a given time period, and the decision attribute is the flood category of a given time period. Finally, the newly developed methodology is compared to the performance of the traditional flood forecasting model.

CLASSIFIED REAL-TIME FLOOD FORECASTING MODEL

The classified real-time flood forecasting model has three basic components: the K-means clustering model, optimization algorithm and rough set. This model is illustrated in Figure 1.

In this framework, the K-means clustering model is used to classify historical floods of each time period into i categories in terms of influential factors. As shown in Figure 1, N categories are investigated. The historical floods in each time period are marked as research cases. The cases are defined by condition attributes and decision attribute. In each case, the flood category of a given time period is taken as the decision attribute and the condition attributes include initial surface storage capacity (SSC), initial lower storage capacity (LSC), WSs, the PC of the given time
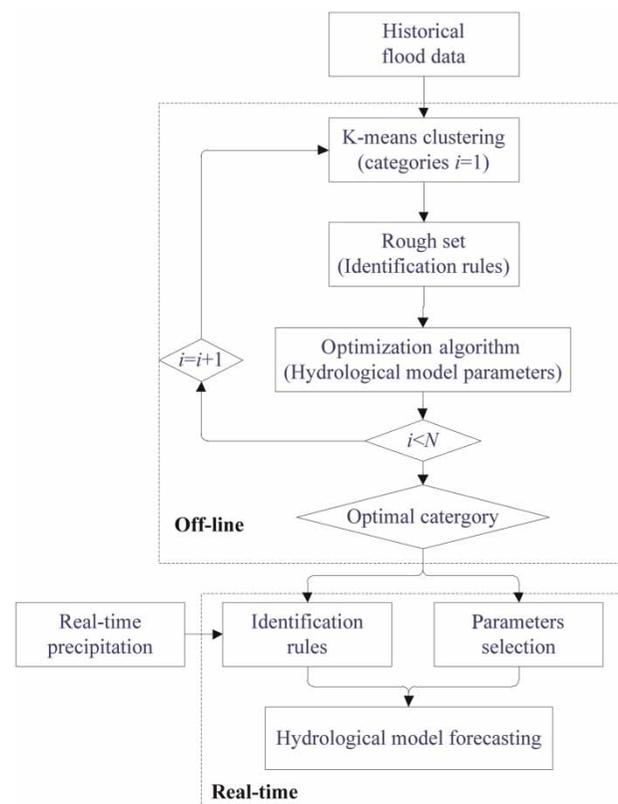

Figure 1 | The real-time flood classification forecasting model structure.

period and precipitation intensity (PI) of the given time period. According to the condition attributes and the decision attribute, the identification rules are extracted by rough set for i categories. Then, historical floods can be simulated using the identification rules to select the forecasting model parameters for each time period. Next, the conceptual hydrological model parameters for i categories can be calibrated using the optimization algorithm. At last, the identification rules, hydrological model parameters and best simulation performance of N categories are chosen to forecast floods in real-time. In this framework, historical flood clustering and parameter calibration are performed off-line because they are computationally expensive.

Cluster analysis for historical floods

In the K-means algorithm, the initial category centers are selected randomly during a set of study cases (Lai *et al.* 2008). The Euclidean distance equation is used to classify the samples. Then, the category centers are recalculated and the cases are reclassified. This process is repeated until the category centers are stable. The Euclidean distance

(J) equation is defined by Lai *et al.* (2008) as

$$J = \min \sum_{i=1}^{L} \sum_{x_i \in S_j} |x_i - c_j|^2 \tag{1}$$

where x_i represents the classified category of the data sample, $i = 1, 2, \cdots, L$. L represents the number of samples. c_j and S_j represent the category and its center, respectively, $j = 1, 2, \cdots, k$.

In the classified real-time flood forecasting framework, historical floods in each time period are classified by a K-means clustering model according to influential factors. The classified category x_i is the decision variable in the identification rules extraction. The classified category c_j is k categories. The conceptual hydrological model parameters of k categories must be calibrated. Owing to the number of the categories, (k) is unknown, so different values of k are investigated according to influential factors.

Rough set for identification rules extraction

The rough set theory, introduced by Pawlak (1982, 2007), is a mathematical approach to deal with a specific type of uncertainty in data related to granulation of the information. This type of uncertainty is very different from the uncertainty considered in the fuzzy set (Zadeh 1965). This type of uncertainty is concerned with a type of arising imprecision that occurs when the boundaries of a class of objects are not sharply defined.

In the field of water management, the rough set methodology is applied in many study cases (Shrestha *et al.* 1996; Fontane *et al.* 1997; Labadie 2004; Salvatore *et al.* 2006), and some studies try to define decision rules. The main difference between the rough set methodology and fuzzy set theory is that, in general, fuzzy set theory does not allow an information reduction process based on the relevance of particular subsets of attributes (reduces and core) (Salvatore *et al.* 2006). However, the rough set approach has specific advantages compared to standard statistical analysis, and the 'if-then-else' decision rule of the rough set approach is expressed in a natural and easily understandable language.

More precisely, the rough set approach allows information to be described in a data table in which the row refers to research cases and the column refers to condition attributes and decision attribute. In rough set theory, the knowledge expression system is written by $S = <U, A, V, f>$. S represents the data table. U represents the set of samples. If the set of attributes (A) in the data table is divided into *condition* attributes (set $C \neq 0$) and *decision* attributes (set $D \neq 0$), such a table is called a *decision* table ($A = C \cup D$) (Salvatore *et al.* 2006). $V = \bigcup_{a \subset A} V_a$ represents the range of attributes value, where V_a represents the range of attribute a ($a \in A$). f represents the information mapping from $U \times A$ to V (Pawlak 1982; Salvatore *et al.* 2006). According to the important degree formula proposed by Salvatore *et al.* (2006), the important degree of attribute B, subset of attribute C, can be calculated by

$$\sigma_{(C,D)}(B) = \frac{r_C(D) - r_{C-B}(D)}{r_C(D)} = 1 - \frac{r_{C-B}(D)}{r_C(D)} \tag{2}$$

where $r_C(D)$ represents the dependence degree of D for attribute C. According to Equation (2), the attribute of B, when the important degree of B ($\sigma_{(C,D)}(B)$) equals zero, can be removed from attribute C, but otherwise cannot.

Hydrological model parameter calibration

The conceptual hydrological model of Dahuofang, which has been widely used in the northeast of China, is then constructed (Ren & Wang 2010). This model framework is constituted using a super infiltration runoff model and a variable intensity and velocity of flood confluence model (Ren & Wang 2010). Its parameter can be automatic calibrated by a GA (Gupta & Sorooshian 1985), which has been widely used in conceptual hydrological model parameter calibration (Wang 1991; Seibert 2000). This study uses a GA as an available algorithm; the information is described by Cheng *et al.* (2006). The GA methodology consists of two components: a GA algorithm and a technique for order performance by similarity to ideal solution (TOPSIS). TOPSIS was first developed by Hwang & Yoon (1981) for solving a multiple criteria decision-making problem. In the GA methodology, TOPSIS was adopted to evaluate and select the most-fit chromosomes to mate and reproduce.

In the process of parameter calibration, flood peak flow, volume and peak time are taken as evaluation indicators (Seibert 2000; Khu & Madsen 2005). Thus, the objective function consists of four components for deviation by comparing forecasted and observed floods (You *et al.* 2003; Cheng *et al.* 2006).

(1) Deviation of flood volume

$$F_1(\theta) = \left| \frac{\sum_{i=1}^{n} Q_{\text{obs},t} - \sum_{i=1}^{n} Q_{\text{sim},i}(\theta)}{\sum_{i=1}^{n} Q_{\text{obs},i}} \right| \times 100\% \tag{3}$$

(2) Deviation of flood peak flow

$$F_2(\theta) = \left| \frac{\max Q_{\text{obs}} - \max Q_{\text{sim}}(\theta)}{\max Q_{\text{obs}}} \right| \times 100\% \qquad (4)$$

(3) Deviation of flood process

$$F_3(\theta) = 1 - \frac{\sum\limits_{i=1}^{n} [Q_{\text{sim},i}(\theta) - Q_{\text{obs},i}]^2}{\sum\limits_{i=1}^{n} [Q_{\text{obs},i} - \hat{Q}_{\text{obs}}]^2} \qquad (5)$$

(4) Deviation of flood peak time

$$F_4(\theta) = |\tau_{\text{obs}} - \tau_{\text{sim}}(\theta)| \qquad (6)$$

where n is the number of floods for hydrological model parameter calibration; θ represents a set of hydrological model parameters; $Q_{\text{obs},i}$ is the observed flood flow, $i = 1, \cdots, n$; $Q_{\text{sim},i}(\theta)$ is the simulated flood flow; $\hat{Q}_{\text{obs}}$ is the average value of observed flood flow; $\max Q_{\text{obs}}$ is the value of observed flood peak flow and $\max Q_{\text{sim}}(\theta)$ is the value of simulated flood flow; τ_{obs} and $\tau_{\text{sim}}(\theta)$ represent the observed and simulated of flood peak time, respectively.

In the GA algorithm, the fitness function includes two components, i.e. deviation of flood process and penalty for deviation in the limited system range (You et al. 2003).

$$f(\theta) = F_3(\theta) + \sum_{j=1,2,4} \alpha_j M_j \qquad (7)$$

$$M_j = \begin{cases} \left| F_j(\theta) - \beta_j \right| & F_j > \beta_j \\ 0 & F_j \leq \beta_j \end{cases} \qquad (8)$$

where $f(\cdot)$ represents fitness function in the GA algorithm; β_j is the deviation for system allowed, $j = 1, 2, 4$; α is penalty factor.

CASE STUDY

Description of study reservoir system

The Guanyinge Reservoir is used to test the newly developed methodology in this study. In the catchment, the average annual precipitation is about 826.3 mm and about 70–80% of the precipitation occurs during wet seasons (from May to September). Meanwhile, WSs, such as the north China cyclone, the low pressure front, the Yangtze–Huaihe cyclone, the upper trough and typhoons are the main influential factors of precipitation.

In this study, the traditional flood forecasting method is used and a single set of parameters is applied to forecast all flood events, which reveals some problems over recent years: (a) the predicted flood peak flow is much lower for large floods and is much higher for small floods than what has been observed; (b) the time of flood peak flow for small floods is inaccurately predicted. In this work, the traditional flood forecasting method and the new classified real-time flood forecasting framework are applied to the study catchment for flood forecasting.

The Dahuofang hydrological forecasting model (DHFM) has been applied to forecast flood processes (Ren & Wang 2010). DHFM has two main components: an infiltration model and a transportation model. In the infiltration model, rainfall loss is calculated using the double-layer infiltration curves that involve an under-layer infiltration curve and a deep-layer infiltration curve (Ren & Wang 2010).

Parameter calibration is an important step for establishing a conceptual hydrological model. The accuracy and reliability of the model are affected significantly by the calibrated parameter. In DHFM, two types of parameters are calibrated, i.e. measured parameters and others. The measured parameters, such as the catchment area and stream slope, are determined for the study catchment. Other parameters, e.g. maximum water-storage capacity and evapotranspiration coefficient, are obtained by calibrating with historical floods.

Datasets

The security of Guanyinge Reservoir and downstream cities needs to be safeguarded. Flood forecasting in the catchment has been studied for many years and detailed information about historical floods is available. This study uses data from 45 historical floods in the catchment upstream of Guanyinge Reservoir from 1960 to 2010. In this study, seven historical floods from 1960 to 1969, including 76 time periods, are chosen for classification, calibration and identification rules extraction. The other 38 historical floods are used to verify the performance of the classified real-time flood forecasting model.

RESULTS AND DISCUSSION

Cluster analysis of historical floods in each time period

During the process of flood confluence, the influential factors have different importance. In this study, initial SSC,

Table 1 │ The combinations of influence factors

| Combination	SSC	LSC	PC	WS	PI
1	o	o	o	o	o
2	x	o	o	o	o
3	o	x	o	o	o
4	o	o	o	x	o
5	x	x	o	o	o
6	x	o	o	x	o
7	o	x	o	x	o
8	x	x	o	x	o

initial LSC, WS, the PC of the given time period and PI of the given time period, are used to cluster the flood categories for each time period. The PC is divided into five areas, i.e. downstream, middle east, middle west, upstream and average. Meanwhile, the weather system is mainly classified as the north China cyclone, the upper trough, the Yangtze–Huaihe cyclone, and typhoons.

In real-time operations, effective influential factors for flood classification are hard to select. Therefore, different factor combinations (FCs) are used as condition attributes to classify historical floods in this work, as shown in Table 1. According to the different FCs, historical floods are classified and simulated. The performances are compared, and the best FC is selected. In Table 1, the 'o' represents the influential factor included in the particular combination, and the 'x' represents the factor excluded in the particular combination.

During the clustering process, the historical floods of each time period are firstly classified according to FC. The classified category is defined in Equation (1) as x_i. Then, the parameters of the conceptual hydrological model are calibrated for each category, which is also defined in Equation (1) as classified category c_i. In calibration, the sets of the parameters for classified category c_i are initialized by the GA and the initialized parameters are selected according to the category of each time period. In this way, the floods for calibration are simulated and the fitness values are evaluated to select optimal parameters.

During GA, the fitness value is used to evaluate the simulation performance by comparing simulated to observed floods. The fitness value curves, for FC from 1 to
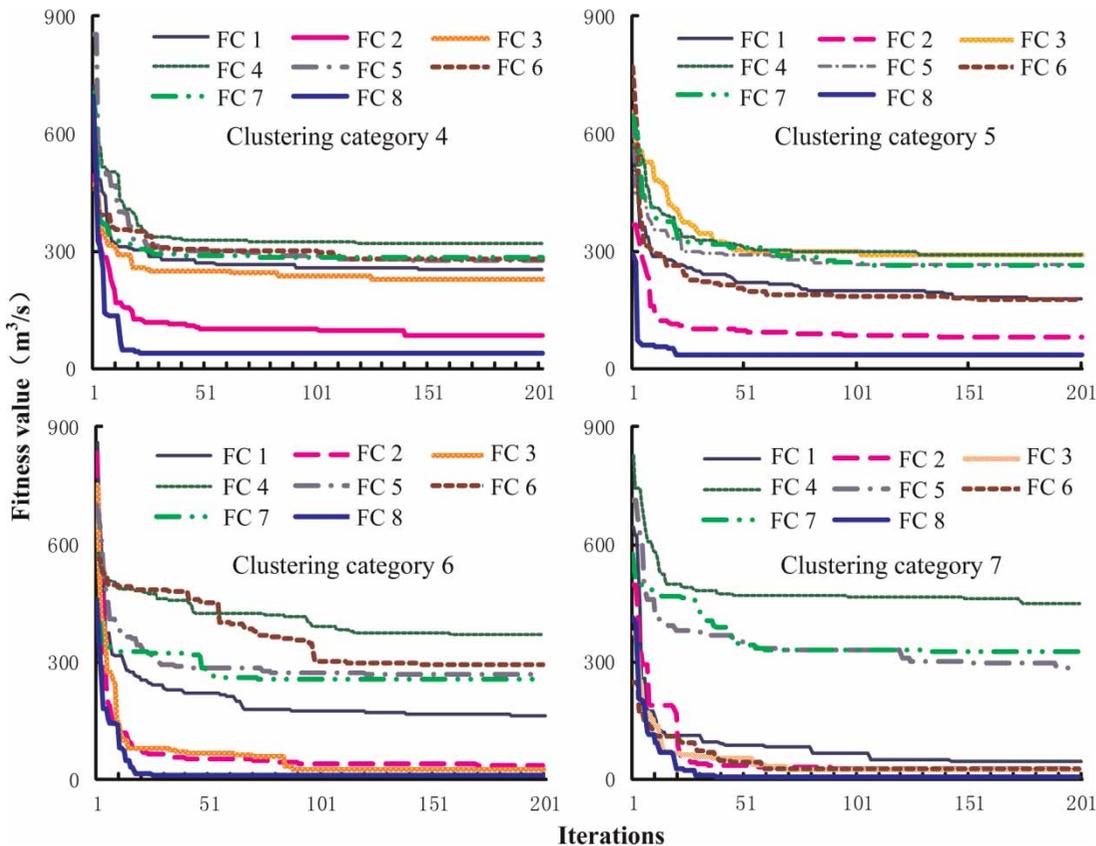
Figure 2 │ The fitness values varying curve with different FC and clustering category.

8 and classification category from 4 to 7, are shown in Figure 2. Generally, rapid convergence rate and minimum value represent the optimal solution. In Figure 2, the different FC and classification categories have different convergence rates and values. FC 8 has a rapid convergence rate, and the fitness value of FC 8 is lower than that of the others. This demonstrates that the influential factors of PI and PC are the most important factors for classification and identification.

The optimal fitness values of the FC 8, from the classification categories 1 to 10, are shown in Table 2. The fitness value of category 1 in Table 2 represents the performance of the traditional method by using a single set of parameters to forecast different types of flood. The fitness values in Table 2 are decreased when the classification category increases. Regarding the variation value, the fitness values decrease quickly from category 2 to category 6, while the fitness values from category 7 decreases slowly. This demonstrates that the flood simulations perform better as the classification category increases. However, there is a threshold value for the classification category. When the classification category exceeds the threshold value, the simulation performance is not improved significantly. In this work, category 6 is chosen as the threshold value.

Extraction results of identification rules

At present, data mining algorithms are widely used in many fields, including knowledge discovery in databases, data harvesting, data archaeology and data pattern analysis (Bessler et al. 2003). Moreover, there are a large number of definitions for this group of algorithms, of which the rough set is the widespread algorithm. In this section, the identification rules are constituted by the condition attributes and the decision attribute. The condition attributes are constituted by PC and the intensity of the given time period. And the flood confluence category of the given time period is taken as the decision attribute.

In this study, seven historical floods are divided into 76 time periods, and the 76 time periods are taken as study cases. The identification rules, as shown in Table 3, are extracted based on the 76 cases by a rough set algorithm for real-time flood forecasting. The identification rules are evaluated according to three indicators, i.e. support, coverage and confidence. The first identification rule in Table 3 is taken as an example to illustrate the indicators. In this example, the identification rule is extracted from 22 cases. These 22 cases represent the support, and the percentage of support in all 76 cases is represented by coverage. The categories of the 22 cases are all equal to 1; thus, the confidence is 100%.

In real-time forecasting, information about a given time period is obtained and used to identify the flood confluence category. In this process, several identification rules may be appropriate for the information. Thus, coverage and confidence are used to identify the category. The principle in this study is that coverage is more important than confidence. In this way, the flood confluence category for the given time period is identified. Following this, the parameters of the corresponding category are chosen for flood forecasting.

Simulation results of classified real-time flood forecasting model

During verification, 38 historical floods are used to evaluate the performance of the classified real-time flood forecasting model. The peak flood flow, flood volume and flood peak time are used to evaluate and compare the performances

Table 2 │ The fitness values of FC 8 from clustering category 1–10

Category	Fitness value	Variation value	Category	Fitness value	Variation value
1	82.09	—	6	11.66	21.95
2	73.47	8.62	7	9.25	2.41
3	52.85	20.62	8	9.01	0.24
4	38.23	14.62	9	8.52	0.49
5	33.61	4.62	10	8.47	0.05

Table 3 │ The identification rules for real-time flood forecasting

| Number	PC	PI	Category	Support	Coverage	Confidence (%)
1	3	2	1	22	0.293	100
2	2	*	4	8	0.107	100
3	*	6	5	5	0.067	100
4	2	3	3	4	0.053	100
5	3	*	3	6	0.081	100
6	2	2	2	2	0.027	100
7	1	*	6	7	0.095	100
8	2	7	5	2	0.027	100
9	3	5	5	8	0.107	100
10	*	1	1	5	0.067	100
11	*	8	6	2	0.027	100
12	*	9	6	2	0.027	100
13	1	4	5	3	0.04	100

Table 4 | Performances of classified vs. traditional flood forecasting during calibration

	Peak flood flow		Flood volume		Flood peak time
	Relative error (%)	Percentage of pass (%)	Relative error (%)	Percentage of pass (%)	Percentage of pass (%)
Traditional	13.74	82.85	13.23	71.42	80
Classified	7.19	94.28	5.78	97.14	91.42

Table 5 | Performances of classified vs. traditional flood forecasting during verification

	Peak flood flow		Flood volume		Flood peak time
	Relative error (%)	Percentage of pass (%)	Relative error (%)	Percentage of pass (%)	Percentage of pass (%)
Traditional	9.75	80	16.52	70	80
Classified	5.07	95	6.01	97	94

of the traditional flood forecasting model and the classified real-time flood forecasting model. Relative error and percentage of pass are chosen as metrics to evaluate the accuracy of the models (Ren & Wang 2010).

The performance results during calibration and verification are shown in Tables 4 and 5, respectively. It can be seen from Tables 4 and 5 that the performances of the classified real-time flood forecasting model are better than those of the traditional flood forecasting model. The traditional

flood forecasting model, using a single set of parameters, is used to forecast all the flood events. In this way, when flood peak flow exceeds 7,000 m^3/s, the flood peak flow and flood volume are usually forecasted below those of historic floods. For example, flood number 19600804, generated by a typhoon with a flood peak flow of approximately 10,000 m^3/s is shown in Figure 3(a). When the flood peak flow is under 1,000 m^3/s, the flood peak flow and flood volume are forecasted higher. The flood peak
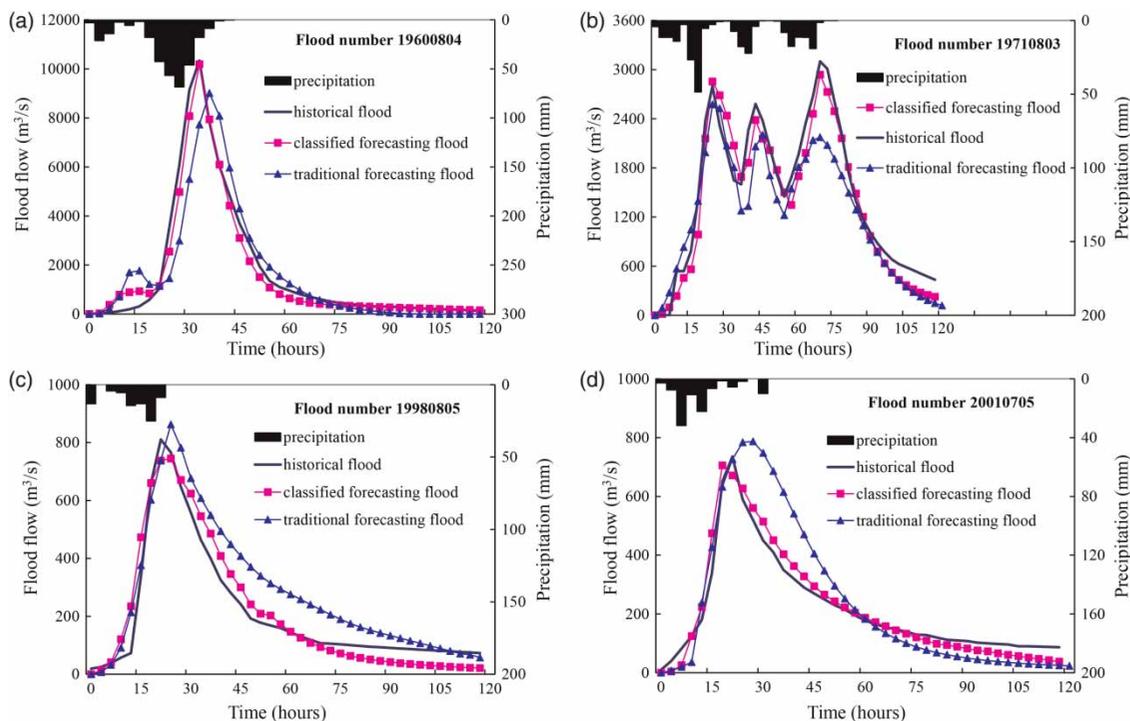
Figure 3 | Flood flow hydrographs of traditional forecasting flood, classified forecasting flood and observed flood process. (a) Flood number 19600804; (b) flood number 19710803; (c) flood number 19980805; and (d) flood number 20010705 (time-interval is 3 hours).

flows and volumes of flood number 19980805 and 20010705, shown in Figure 3(c) and 3(d) respectively, are forecasted higher than those of historical floods. Moreover, the traditional flood forecasting model has poor performance for multi-peak flood forecasting. For example, flood number 19710803 as shown in Figure 3(b), which is a typical multi-peak flood case, is forecasted lower gradually.

SUMMARY AND CONCLUSIONS

It is difficult to forecast floods accurately because of the randomness and complexity in flood events. When the traditional flood forecasting model, which uses a single set of parameters, is used to forecast floods, it performs poorly for large and small flood events. For reservoir operations and decision-making during flood season, this study investigates the framework of the classified real-time flood forecasting model to improve flood forecasting accuracy.

In this framework, historical floods are classified through a K-means cluster model. Subsequently, six sets of hydrological parameters are calibrated by a GA. Then, the flood category for a given time period is identified by identification rules, which have been extracted by the rough set model. Finally, the floods are simulated by the calibrated parameters.

This research finds that the results of the classified real-time flood forecasting framework are more reliable than those of a traditional flood forecasting model with a single parameter set. These findings confirm that it is necessary to consider the variability of precipitation and geomorphology during different periods. This framework is a better tool for improving the accuracy and reliability of flood forecasting and reservoir operations.

In this study, seven floods are divided into 76 periods and used to calibrate the forecasting model parameters. In this way, this framework can be considered in the catchment with fewer historical floods.

ACKNOWLEDGEMENTS

REFERENCES

Bessler, F., Savic, D. & Walters, G. 2003 Water reservoir control with data mining. *Journal of Water Resources Planning and Management* **129**, 26–34.

Cheng, C. T., Zhao, M. Y., Chau, K. W. & Wu, X. Y. 2006 Using genetic algorithm and TOPSIS for Xinanjiang model calibration with a single procedure. *Journal of Hydrology* **316**, 129–140.

Fontane, D., Gates, T. & Moncada, E. 1997 Planning reservoir operation with imprecise objectives. *Journal of Water Resources Planning and Management* **123** (3), 154–162.

Gupta, V. K. & Sorooshian, S. 1985 The automatic calibration of conceptual catchment models using derivative-based optimization algorithms. *Water Resources Research* **21** (4), 473–485.

Gupta, V. K., Waymine, E. C. & Wang, C. T. 1980 A representation of an instantaneous unit hydrograph from geomorphology. *Water Resources Research* **16** (5), 855–862.

Hwang, C. L. & Yoon, K. 1981 *Multiple Attributes Decision Making Methods and Applications*. Springer, Berlin.

Jain, A. & Srinivasulu, S. 2006 Integrated approach to model decomposed flow hydrograph using artificial neural network and conceptual techniques. *Journal of Hydrology* **317** (3–4), 291–306.

Khu, S. T. & Madsen, H. 2005 Multi-objective calibration with Pareto preference ordering: an application to rainfall-runoff model calibration. *Water Resources Research* **41**, art. W03004.

Labadie, J. 2004 Optimal operation of multireservoir systems: state of the art review. *Journal of Water Resources Planning and Management* **130** (2), 93–111.

Labat, D., Ababou, R. & Mangin, A. 2000 Rainfall-runoff relations for karstic springs. Part I: convolution and spectral analyses. *Journal of Hydrology* **238** (3–4), 123–148.

Lai, Y., Liu, J. & Yang, G. 2008 K-means clustering analysis based on genetic algorithm (in Chinese). *Computer Engineering* **34** (20), 200–202.

Lu, Z. & Hou, Z. 2007 Cluster validity analysis for flood clustering (in Chinese). *South-to-North Water Transfers and Water Science & Technology* **5** (2), 87–90.

Ma, Y., Zhou, X., Shang, J. & Zhang, Y. 1997 A new method for flood types prediction and optimal control (in Chinese). *Journal of Hydraulic Engineering* **4**, 1–8.

Pawlak, Z. 1982 Rough sets. *International Journal of Computer and Information Science* **11** (5), 341–356.

Pawlak, Z. 2007 Rough sets and Boolean reasoning. *Inform Sciences* **177**, 41–73.

Ren, M. & Wang, B. 2010 Classified real-time flood forecasting by coupling fuzzy clustering and neural network. *International Journal of Sediment Research* **25** (2), 134–148.

Rinaldo, A., Marania, A. & Rigon, R. 1991 Geomorphological dispersion. *Water Resources Research* **27** (4), 513–525.

Rodriguez, I. & Valded, J. B. 1979 The geomorphologic structure of hydrologic response. *Water Resources Research* **15** (6), 1409–1420.

Saghafian, B. & Julien, P. Y. 2002 Runoff hydrograph simulation based on time variable isochrone technique. *Journal of Hydrology* **261**, 193–203.

Salvatore, B., Simona, C., Nello, P., Salvatore, G. & Santo, M. Z. 2006 Discovering reservoir operating rules by a rough set approach. *Water Resources Management* **20** (1), 19–36.

Seibert, J. 2000 Multi-criteria calibration of a conceptual runoff model using a genetic algorithm. *Hydrology and Earth System Sciences* **4** (2), 215–224.

Sherman, L. 1932 Stream-flow from rainfall by the unit-graph method. *Engineering News-Record* **108**, 501–505.

Shrestha, B., Duckstein, L. & Stakhiv, E. 1996 Fuzzy rule-based modelling of reservoir operation. *Journal of Water Resources Planning and Management* **122** (4), 262–269.

Wang, Q. J. 1991 The genetic algorithm and its application to calibrating conceptual rainfall-runoff models. *Water Resources Research* **27**, 2467–2471.

Wang, S., Zhang, X., Hou, Y. & Ding, J. 2002 Projection pursuit model for evaluating of flood events (in Chinese). *Journal of China Hydrology* **22** (4), 1–4.

Xu, W., Liang, G. & Wang, B. 2013 Study and application of classified real-time flood forecasting based on two step cluster and rough set (in Chinese). *Journal of Hydroelectric Engineering* **32** (2), 60–67.

Yin, X., Zhang, X. & Xia, J. 2007 Classification-based flood forecasting model using artificial neural networks (in Chinese). *Journal of Sichuan University* **39** (3), 34–39.

You, J., Ji, C. & Fu, X. 2003 New method for solving multi-objective problem based on genetic algorithm (in Chinese). *Journal of Hydraulic Engineering* **7**, 64–69.

Zadeh, L. 1965 Fuzzy sets. *Information Control* **8**, 338–353.

Zhang, W., Xia, J., Zhang, X. & Zhang, L. 2007 A unit hydrograph model considering spatio-temporal variation of rainfall (in Chinese). *Journal of China Hydrology* **27** (5), 1–6.