Application of similarity analysis to flood forecasting

Xiao Zhangling[1], Liang Zhongmin[1,2], Li Binquan[1,2], Zhou Yan[1], Wang Yanlan[1]

[1] College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China
[2] State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing 210098, China

Abstract: [Objective] There are plenty of useful information in hydrological observations. Predicting future flood on the basis of similarity information in historical records is an effective and promising approach. [Method] In this paper, a multi-measure similarity analysis method of rainstorms is developed based on "quantity", "type" similarity indicators, the earth mover's distance (EMD) and the rainstorm distribution similarity indicator. Search the similar rainstorm and its corresponding typical flood in historical library and then scale the typical flood process according to the ratio of rainfall amounts to achieve flood forecasting. [Result] The method is applied to a case study in Xinmiao station of Kuye River. The results show that with the accelerating information of rainstorm and flood process, the forecasted flood process is updated continuously, and the prediction accuracy is gradually increasing. [Conclusion] The proposed similarity analysis method is effective and applicable to flood forecasting.

1 Introduction

Flash floods are one of the most serious natural disasters that cause considerable material, social and environmental losses, even lead to loss of human lives (Casagrande et al., 2017; Wang et al., 2017). Real-time flood forecasting is an important non-engineering measure for flood mitigation. Reliable flood forecasting plays a vital role in reducing potential flood risks, generating higher benefits of hydropower assets and increasing the utilization efficiency of water resources (Wang et al., 2015; Li et al., 2016; Barbetta et al., 2017).

Traditionally, conceptual model is the main approach in flood forecasting, that is, establishing the hydrological models that can reflect the runoff production and confluence mechanism to achieve the prediction of hydrological process (Solomatine and Dual, 2003; Vaché and Mcdonnell, 2006). However, the current conceptual models are more suitable for runoff simulation in wet areas while the prediction accuracy is often low in the arid and semi-arid areas. Therefore, there is an urgent need to develop a forecasting method that can not only avoid the direct flood simulation in arid and semi-arid areas, but also meet certain prediction accuracy. Present-day instrumentation networks has already provide immense quantities of data, therefore, hydrology can be a data-intensive science (Spate et al., 2003). With the fast development of contemporary computational techniques, time series data mining can extract meaningful knowledge from massive historical observations to predict the future runoff. Flood forecasting technology based on time series data mining techniques can provide a solution for the problem.

Time series data mining techniques are developed to fully explore the information contained in the data and transform the information into useful knowledge for improving decision making. There are plenty of research in time series data mining, for example, searching similar time series (Agrawal et al., 1993; Berndt and Clifford, 1996; Das et al., 1997), pattern discovery (Aljawarneh et al., 2016), dimensionality reduction (Keogh and Pazzani, 2000; Maciej and Grażyna, 2014) and segmentation (Kehagias and Petridis, 1997; Jamali et al., 2015). Faced with such massive hydrological data, time series data mining technology has played an increasingly important role in dealing with related issues. Among them, similarity analysis is one of the core tasks. Namely, we can find all kinds of similar sequence pairs in the hydrological sequence library, which contains knowledge including the evolution process of climate and underlying surface. The work of similarity search will be helpful in flood forecasting, the analysis of environmental evolution and so on.

Similarity analysis have an important role in flood forecasting. For instance, Veitzer and Gupta (2001) employed random self-similar model to search the for the width function maxima with implications for floods. Wan et al. (2010) developed a similarity based forecasting model and applied it for reservoir operation of water and sediment. Ouyang et al. (2010) used dynamic time warping (DTW) to search similar hydrological discharge processes and discover hydrological discharge process patterns. Ben Daoud et al. (2011) presents the application of a precipitation forecasting technique based on analogs to forecast discharges at the Seine river basin. However, most of exist studies start from the whole flood process to search similarity. In fact, the formation of rainstorm and flood event is a gradual evolution process. With time

going, the more flood information would be obtained, the more favorable it would be to predict the future flood.

Lead time is a key indicator for flood forecasting (Li et al., 2017). Flood forecasts are efficient with long enough lead time to initiate flood warnings or flood emergency measures, such as reservoir operations and transmission of the warning to local decision makers (Klatt and Schultz, 1983). In this work, we forecast the future rainfall through a multi-measure similarity analysis method of rainstorms, searched the similar rainstorm process and its corresponding typical flood process in historical records and then scaled the typical flood process according to the ratio of rainfall amounts to achieve the flood forecasting. The multi-measure similarity method offered an overall rainfall similarity measure on the basis of similarity indicators on "quantity" and "type", the earth mover's distance (EMD) and rainfall distribution similarity indicator.

The rest of this paper is organized as follows: Section 2 describes the multi-measure similarity analysis method of rainstorms. On this basis, a flood forecasting method based on similarity analysis is constructed. Section 3 presents a case study of the proposed method. Finally, the conclusions are summarized in Section 4.

2 Models

2.1 The Multi-measure Similarity Analysis Method

Based on the idea that multi-measure can improve the accuracy of similarity analysis, multi-measure similarity analysis method is put forward to evaluate the similar degree of two rainstorms. Firstly, four indicators including "quantity" similarity, "type" similarity, earth mover's distance (EMD) indicator and the rainstorm distribution similarity, are built to describe the temporal and spatial similar degree of two rainstorms. Then, the overall similarity is computed based on these indicators. In this study, the arithmetic mean method is adopted to calculate areal rainfall.

(1) The similarity indicator on "quantity"

Suppose there are N rainfall stations over the studied basin. Set the observed precipitation of each gauging station and its areal precipitation at time t in one rainstorm process as $X_{t,k}$ and X_t respectively, and those in the other rainstorm process as $Y_{t,k}$ and Y_t respectively, where k=1,2,…,N and t=1,2,…,T. We define the difference between accumulated precipitations in two rainfall processes as "quantity" similarity.

$$X_t = \frac{1}{N}\sum_{k=1}^{N} X_{t,k} \quad , \quad Y_t = \frac{1}{N}\sum_{k=1}^{N} Y_{t,k}$$

$$quantity(X,Y) = \sum_{t=1}^{T} X_t - \sum_{t=1}^{T} Y_t \tag{1}$$

Where X denotes the current rainfall process and Y denotes the historical rainfall process; T represents the comparison time. The "quantity" similarity indicator describes the similar degree of two rainstorms in terms of accumulated precipitations. The lower value of $quantity(X,Y)$ indicates the more similar of two rainstorms in terms of accumulated precipitation.

(2) The similarity indicator on "type"

The temporal similar degree between two rainstorms is defined as "type" similarity. Let $con(t)=(X_t-X_{t+1})/(Y_t-Y_{t+1})$. The "type" similarity of two rainstorm processes at time t can be expressed by a unit step function.

$$Score(t) = \begin{cases} 0, con(t) \le 0 \\ 1, con(t) > 0 \end{cases} \tag{2}$$

Where $con(t)$ represents the consistency of two rainstorms at time t. If the two rainstorm processes have the same trend at time t (both increasing or decreasing), the value of $con(t)$ would be positive and $Score(t)$ would be assigned to 1. On the contrary, the value of $con(t)$ would be negative and $Score(t)$ would be assigned to 0 it they are opposite. Moreover, the accumulated unit step function $\sum_t Score(t)$ is employed to estimate the "type" similarity in two rainstorms. The higher value of this indicator means the more similarity over time, and vice versa.

(3) The earth mover's distance (EMD) indicator

To investigate the similarity of two rainstorms (current rainstorm X and historical rainstorm Y) from several perspectives, we adopt the earth mover's distance (EMD) (Rubner et al., 2000) indicator. The earth mover's distance can assess the similarity of two distributions or signatures, which was firstly applied to image retrieval problems. Suppose each signature with some clusters (feature points). To define the EMD of two signatures, the first step is to define the distance of their feature points, such as the values of signatures at different times, also known as ground distance. Given two signatures, one can be regarded as a set of some earth piles and the other as a set of pits in the same space. The EMD indicator id a measure of the amount of work required to fill the pits with these piles. Obviously, the lower value of the indicator means the smaller difference between the two signatures. Therefore, the earth mover's distance indicator can be used to assess the similarity of two rainstorms.

Suppose d_{ij} as the ground distance between X_i and Y_j and f_{ij} as the flow matrix between X and Y, representing the flow between X_i and Y_j (amount of earth pile to be moved). We need to find a flow matrix to minimize the total work function $\sum_{i=1}^{T}\sum_{j=1}^{T} d_{ij} f_{ij}$ subject to the following conditions:

$$f_{ij} \ge 0 \quad for \ 1 \le i \le T, 1 \le j \le T \tag{3}$$

$$\sum_{j=1}^{T} f_{ij} \le X_i \quad for \ 1 \le i \le T \tag{4}$$

$$\sum_{i=1}^{T} f_{ij} \le Y_j \qquad for \;\; 1 \le j \le T \tag{5}$$

$$\sum_{i=1}^{T}\sum_{j=1}^{T} f_{ij} = \min(\sum_{i=1}^{T} X_i, \sum_{j=1}^{T} Y_j) \tag{6}$$

For the similarity measure of rainstorms, d_{ij} is the distance between the current rainstorm X at time i and the historical rainstorm Y at time j, $d_{ij}=abs(i-j)$, that is, if the moving time differs smaller, the ground distance would be lower and vice versa. f_{ij} denotes the moving rainfall amount from the current rainstorm X at time i to the historical rainstorm Y at time j. Once the optimal flow is determined, the EMD is defined as the resulting work normalized by the total flow:

$$EMD(X,Y) = \frac{\sum_{i=1}^{T}\sum_{j=1}^{T} d_{ij} f_{ij}}{\sum_{i=1}^{T}\sum_{j=1}^{T} f_{ij}} \tag{7}$$

The lower value of EMD indicates the higher similar degree of the two rainstorms. And its higher value means large dissimilarity between the two rainstorms.

(4) The rainstorm distribution indicator

The difference between rainfall amounts at each station in two rainstorms are defined as rainstorm distribution similarity, which is measured by Euclidean distance.

$$Euclidean = \sum_{t=1}^{T} (\sum_{k=1}^{N} X_{t,k} - Y_{t,k})^{\frac{1}{2}} \tag{8}$$

Where $X_{t,k}$ and $Y_{t,k}$ represent the observed precipitation of the kth gauging station at time t in the current rainstorm and the historical rainstorm respectively. This indicator can describe their similarity in spatial distribution of rainstorms. When Euclidean is lower, the difference of the two rainstorms' distribution is smaller and the similar degree is higher.

(5) The overall multi-measure similarity indicator

The above four indicators measure the similarity of rainstorms from different terms and it is necessary to construct an indicator to reflect the overall similarity of two rainstorms. The standard deviation method is employed to convert the indicators' value to [0,1]. Given R historical rainfall processes, the similarity indicator of the rth (r=1,2,…,R) historical rainfall process with the current rainfall process X after standardization is

$$x_s^*(r) = \frac{x_s(r) - x_{s\min}(r)}{x_{s\max}(r) - x_{s\min}(r)} \tag{9}$$

Where $x_s(r)$ represents the sth (s=1,2,3,4) original similarity indicator of the rth historical rainfall process with the current rainstorm X; $x_s^*(r)$ denotes the standardized indicator of $x_s(r)$; x_{smax} and x_{smin} are the maximum and minimum value of the indicator $x_s(r)$ respectively. The overall similarity measure $W(r)$ can be calculated after the average processing of the above four indicators.

$$W(r) = \frac{1}{M}\sum_{s=1}^{M} x_s^*(r) \tag{10}$$

Where $x_s^*(r)$ denotes the standardized similarity indicator; M denotes the number of indicators. In this study, we use four indicators, namely, M=4.

2.2 The Real Time Flood Forecast

Referring to the design flood calculation method, the most similar historical rainfall process searched by the multi-measure rainfall similarity method is regarded as the typical rainfall process, and its corresponding runoff process as the typical flood process. Then we can scale the typical flood process according to the ratio of rainfall amount, and finally realize the real time flood forecast and early warning.

$$Q(t) = \frac{P_{current}}{P_{historical}} Q_{historical}(t) \tag{11}$$

Where $P_{current}$ and $P_{historical}$ are the total rainfall amount of the current rainfall process and the typical rainfall process respectively, and $Q_{historical}(t)$ represents the typical flood process.

The above real time flood forecast method is a simplistic but convenient rainfall-runoff model, which aims to providing an estimation of the future flood event rather than a perfect accurate prediction results. This forecasting method will work efficiently for a successful early warning system.

3 Case Study

In this paper, we realize the real time flood forecast and early warning through the following procedure. Firstly, with the multi-measure similarity analysis method of rainstorms, we can search the most similar rainfall process in the historical rainstorms based on the known information of the current rainstorm. Then, under the assumption of similar flood-causing rainstorm leading to similar flood, we can scale the typical flood process according to the ratio of rainfall amount and obtain the forecasted flood. With the progress of rainstorm and flood, the information of rainstorm and flood is gradually increasing, meanwhile the forecasted flood process is updated constantly. The procedure is applied to the Xinmiao station on Kuye River. There are 85 historical rainfall events and corresponding flood events.

Taking the first 5 time intervals of the number 20040811 rainfall process for research, and comparing with each precipitation event in history, the overall similarity indicator ranking at top 10 are found and displayed in Table 1. Finally, it is found that the number 19960810 rainfall process is the most similar to the number 20040811 rainfall process according to the first 5 time intervals rainfall information. The hydrograph of rainfall process is shown in Figure 1 and the forecasted flood process is presented in Figure 2.

Table 1 Results of rainstorm (Number 20040811) similarity indicators (the first 5 intervals)

Number	"quantity"	"type"	EMD	Rainstorm distribution	Overall indicator
19960810	2.850	2	0.285	3.676	0.705
19820708	2.735	2	0.154	3.930	0.718
19850824	1.608	2	0.226	4.936	0.744
19840613	3.931	2	0.393	3.916	0.765
19910610	2.248	2	0.533	5.050	0.778
20010810	3.415	2	0.634	4.950	0.817
19820709	5.514	2	0.551	4.538	0.872
19840702	5.753	2	0.575	4.540	0.882
19850708	4.051	2	0.413	5.920	0.911
19820815	2.663	1	0.375	11.500	0.914

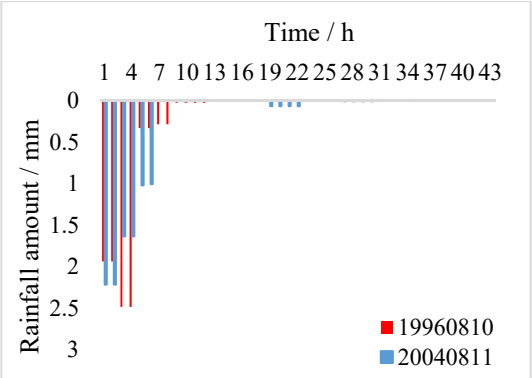

Figure 1 Rainstorms of number 19960810 and 20040811

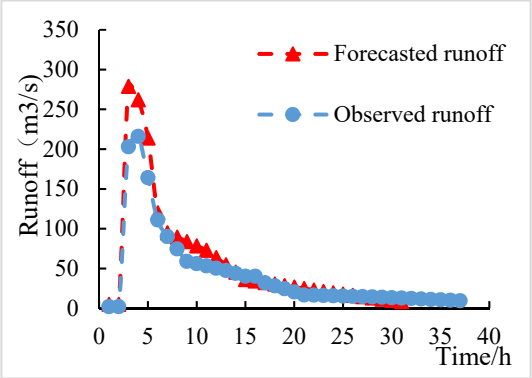

Figure 2 Forecasted flood of number 20040811

Taking the first 10 time intervals of the number 20040811 rainfall process for research, and comparing with each precipitation event in history, the overall similarity indicator ranking at top 10 are found and displayed in Table 2. Finally, it is found that the number 19840704 rainfall process is the most similar to the number 20040811 rainfall process according to the first 10 time intervals rainfall information. The hydrograph of rainfall process is shown in Figure 3 and the forecasted flood process is presented in Figure 4.

Table 2 Results of rainstorm (Number 20040811) similarity indicators (the first 10 intervals)

Number	"quantity"	"type"	EMD	Rainstorm distribution	Overall indicator
19960810	2.850	2	0.285	3.676	0.705
19840704	3.876	5	0.281	3.928	0.682
19840613	4.719	6	0.236	2.398	0.762
19940708	5.321	6	0.266	2.845	0.805
19820709	6.548	6	0.327	2.645	0.818
19840702	6.840	6	0.342	2.710	0.828
19810805	6.208	6	0.326	3.070	0.839
19810729	6.033	6	0.330	3.225	0.846
19830519	7.400	6	0.370	2.840	0.849
19960712	5.177	6	0.386	3.670	0.859

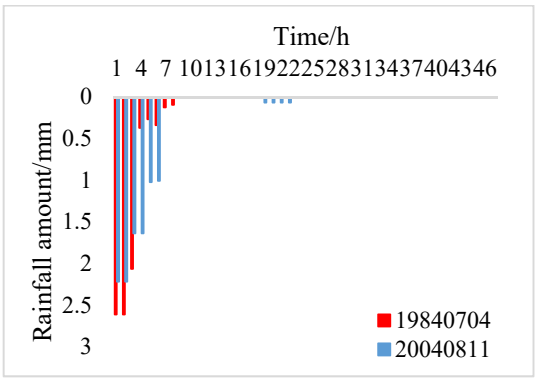

Figure 3 Rainstorms of number 19840704 and 20040811

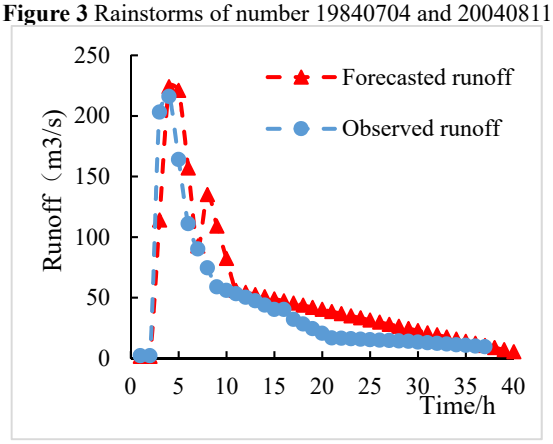

Figure 4 Forecasted flood of number 20040811

When only the first 5 periods of current rainstorm information are inputted, the rainfall of 20040811 is similar to that of 19960810 in the historical rainfall events. But the peak of forecasted flood is higher than the actual flood peak. When the first 10 periods of current rainstorm information are inputted, the rainfall of 20040811 is similar to that of 19840704 in the historical rainfall events. And the peak of forecasted flood is close to the actual flood peak compared to the former search. According to the forecasted results mentioned above, the more information of rainstorm and flood, and the more accurate the forecasted flood. Therefore, the proposed flood forecasting method based on rainfall similarity analysis is effective.

4 Conclusion

In this study, we predict future rainfall by the proposed multi-measure similarity analysis method of rainstorms. Both the observed rainfall and predicted rainfall form the input to a rainfall-runoff model which is employed to achieve the real time forecasts and early flood warning. The used rainfall-runoff model is a simple statistical method, that is, scale the typical flood process according to the ratio of rainfall amount for the sake of realizing real-time flood forecasting. The procedure has been applied in the Xinmiao Station. The main conclusions are as follows:

(1) The evolution of rainstorm flood process usually has the similar law, so the similar rainstorm process can produce similar flood process under certain conditions. Based on the similarity indicators on "quantity", "type", EMD-indicator and rainstorm distribution similarity indicator, an overall measure method of rainstorm similarity is presented in this paper. According to the rainfall ratio, the typical flood process corresponding to the similar rainstorm is scaled to realize the flood forecasting.

(2) Due to the increasing influence of human activities, the underlying surface conditions change greatly, resulting in different floods caused by the same rainfall in different years. In the future, it is necessary to further study the joint similarity between underlying surface conditions and rainstorm processes.

Acknowledge

The study is financially supported by the National Key Research and Development Program of China (2016YFC0402706), National Natural Science Foundation of China (41730750).

Reference

1. Agrawal R, Faloutsos C, Swami A N. Efficient Similarity Search in Sequence Databases [C]. International Conference on Foundations of Data Organization and Algorithms. Springer-Verlag, 1993:69-84.
2. Aljawarneh S, Radhakrishna V, Kumar P V, et al. A similarity measure for temporal pattern discovery in time series data generated by IoT [C]. International Conference on Engineering & Mis. IEEE, 2016.
3. Barbetta S, Coccia G, Moramarco T, et al. The multi temporal/multi-model approach to predictive uncertainty assessment in real-time flood forecasting [J]. Journal of Hydrology, 2017, 551.
4. Ben Daoud, A., Sauquet, E., Lang, M., and Ramos, M.-H. (2011b). Can we extend flood forecastinglead-time by optimising precipitation forecasting based on analogs? Application to the Seine river basin.La Houille Blanche, (1):37–43.
5. Berndt D J, Clifford J. Finding patterns in time series: a dynamic programming approach [M]. Advances in knowledge discovery and data mining. American Association for Artificial Intelligence, 1996:229-248.
6. Casagrande L, Tomasella J, Alvalá R C D S, et al. Early flood warning in the Itajaí-Açu River basin using numerical weather forecasting and hydrological modeling [J]. Natural Hazards, 2017, 88(2):741-757.
7. Das G, Gunopulos D, Mannila H. Finding similar time series [C]. European Symposium on Principles of Data Mining and Knowledge Discovery. Springer, Berlin, Heidelberg, 1997:88-100.
8. Jamali S, Jönsson P, Eklundh L, et al. Detecting changes in vegetation trends using time series segmentation [J]. Remote Sensing of Environment, 2015:182-195.
9. Kehagias A, Petridis V. Time-Series Segmentation Using Predictive Modular Neural Networks [J]. Neural Computation, 1997, 9(8):1691-1709.
10. Keogh E J, Pazzani M J. A Simple Dimensionality Reduction Technique for Fast Similarity Search in Large Time Series Databases [J]. Computers & Education, 2000, 1805(3):122-133.
11. Klatt, P. and Schultz, G. A., 1983. Flood forecasting on the basis of radar rainfall measurement and rainfall forecasting.
12. Li B, Liang Z, Zhang J, et al. Risk Analysis of Reservoir Flood Routing Calculation Based on Inflow Forecast Uncertainty [J]. Water, 2016, 8(11):486.
13. Li J, Chen Y, Wang H, et al. Extending flood forecasting lead time in a large watershed by coupling WRF QPF with a distributed hydrological model[J]. Hydrology & Earth System Sciences Discussions, 2017, 21:1-45.
14. Maciej Krawczak, Grażyna Szkatuła. An approach to dimensionality reduction in time series [J]. Information Sciences, 2014, 260(1):15-36.
15. Ouyang R, Ren L, Cheng W, et al. Similarity search and pattern discovery in hydrological time series data mining[J]. Hydrological Processes, 2010, 24(9):1198-1210.
16. Rubner Y., Tomasi C. and Guibas L.J., 2000. The Earth Mover's Distance as a Metric for Image Retrieval. International Journal of Computer Vision, 40(2):99-121.
17. Spate J M, Crokeb B F W, Jakemanb A J. Data Mining in Hydrology [J]. Hydrological Processes, 2003, 19(7):1511–1515.
18. Solomatine, D., Dulal, K. Model trees as an alternative to neural networks in rainfall-runoff modelling. International Association of Scientific Hydrology Bulletin, 2003, 48(3), 399-411.
19. Vach é K B, Mcdonnell J J. A process - based rejectionist framework for evaluating catchment runoff model structure[J]. Water Resources Research, 2006, 42(2):262-275.
20. Veitzer S A, Gupta V K. Statistical self-similarity of width function maxima with implications to floods [J]. Advances in Water Resources, 2001, 24(9):955-965.
21. Wan X Y, Wang G Q, Peng Y, et al. Similarity-based optimal operation of water and sediment in a sediment-laden reservoir.[J]. Water Resources

Management, 2010, 24(15):4381-4402.

22. Wang J, Shi P, Jiang P, et al. Application of BP Neural Network Algorithm in Traditional Hydrological Model for Flood Forecasting[J]. Water, 2017, 9(1):48.
23. Wang Y, Guo S, Xiong L, et al. Daily Runoff Forecasting Model Based on ANN and Data Preprocessing Techniques[J]. Water, 2015, 2015(7):4144-4160.