Flood Forecasting Method of Small and Medium-sized Watershed Based on Convolutional Neural Network

Jimin Wang¹*, Ying Cao¹, Jiahuan Li², Changzheng Ji¹

¹College of Computer and Information, Hohai University, No.8, Focheng West Road, Nanjing 211100, China
²Information Center of Ministry of Water Resources, No.2, Baiguang Road, Beijing 100053, China

*wangjimin@hhu.edu.cn

Abstract. A data-driven flood forecasting model based on a convolution neural network is proposed for the small and medium-sized watershed. Pearson correlation coefficient analysis was used to determine the time length of the input data. The historical rainfall and discharge were used to create the two-dimensional input data matrix. NAS was used to determine the structure of the model. The experiment results in Tunxi watershed in Anhui Province show that the accuracy is high under the 1~3h forecast period, with the forecast period increased, accuracy would decrease.

Keywords: small and medium-sized watershed; flood forecasting; data-driven model; convolutional neural network.

1. Introduction
Flood forecasting, especially flood forecasting of small and medium-sized watersheds, plays an important role in flood control and rescue, utilization, and protection of water resources, construction, and management of water conservancy projects. Extreme floods in small and medium-sized watersheds seem to occur more frequently in recent years, and cause immense human suffering and enormous economic losses every year worldwide. Therefore, it is necessary to predict the time and size of peak flow accurately before a flood event for the watershed management department.

Over the past decades, many flood forecasting methods have been performed. In terms of methods, they can be roughly divided into the process-driven model and data-driven model. A process-driven model is a generalized simulation of the real hydrological process. In a process-driven model, such as The Stanford model, Sacramento model, TANK model, SMART model, HBV model, and Xin'anjiang model [1-3], a series of parameters with clear physical meaning are used to describe the hydrological mechanism of the watershed. A data-driven model refers to using intelligent algorithms to extract watershed characteristics and other useful information to build the optimal mathematical relationship between forecasting objects (such as the discharge, water level, et al.) and forecasting factors(such as rainfall, soil moisture, et al.) based on historical data, regardless of the physical mechanism of the hy-
Traditional methods used to establish mathematical relations include linear regression, nonlinear regression, support vector machines, and artificial neural networks[4-7].

There are three problems in the current flood forecasting methods: First, the model parameters of the traditional process-driven model are difficult to be determined because of lacking evaporation, soil moisture, and other data for many small and medium-sized watersheds. Second, for some watersheds, the hydrological mechanism has changed because of human activities, so that, no suitable process-driven model could be applied. The last, many existing data-driven models only use shallow networks, which can only look for linear or ordinary non-linear relationships, can’t establish complex non-linear relationships between forecasting objects and factors.

Convolutional Neural Network (CNN) which was proposed in [8] has been widely used in many fields[9]. However, relevant research within the domain of flood forecasting is very limited. In this paper, a flood forecast model based on CNN for small and medium-sized watersheds is proposed. Compared with previous methods, the main contributions and problems solved in this paper are presented as follows.

1. Just the rainfall and discharge were included in the input matrix, and the evaporation, soil moisture, etc. which are missing usually in small and medium-sized watersheds, were not needed in the model.
2. The rainfall of all rainfall stations was included in the input matrix. The influence of rainfall on the outlet discharge at different locations and time, which is considered by a distributed hydrological model, will be carefully considered by CNN, and this could result to improve the forecasting accuracy.
3. Pearson correlation coefficient analysis was used to determine the input time length of rainfall, and reduce the impact of subjective selection.

2. Flood forecast model based on CNN

In this section, we first give an overview of our approach in Section 2.1, which presents the main steps for building a flood forecast model based on CNN. Then, all the steps are detailed in the following subsections.

2.1. Overview

The detailed process is shown in Fig.1.

In the model, the Pearson correlation coefficient is used to analyze the impact time of rainfall on outlet discharge to automatically select the length of input data, NAS[10] is used to select the structure of CNNs, and then train the CNNs with different forecast periods. The deterministic coefficient and flood peak relative error were used to evaluate the model.
2.2. Automatic selection of input data length
Rainfall on different locations in watershed impacts the predicted outlet discharge with different delay times. Assuming that the rainfall of the \(i\)th rainfall station concentrates on the outlet section \(w_i\) time, considering the watershed as a whole, the impact time of rainfall on the outlet discharge is \(w = \max \{w_i\}\), where, \(i = 1, 2, \ldots, m\), \(m\) is the number of rainfall stations in the watershed. \(w\) is the length of input data of our models.

Assuming the outlet discharge time series is \(Q = \{q_0, q_1, q_2, \ldots, q_n\}\), the subsequence of \(Q\) with time delay \(k\) is \(Q[k] = \{q_k, q_{k+1}, q_{k+2}, \ldots, q_n\}\), and the rainfall time series of the \(i\)th rainfall station is \(D_i = \{d_{i0}, d_{i1}, d_{i2}, \ldots, d_{in}\}\). The Pearson correlation coefficient \(P_{ik}\), between \(D_i\) and the subsequence \(Q[k]\), is calculated with time delay \(k\), and \(w\) is the \(k\) of \(\max \{P_{ik}\}\), \(k = 0, 1, 2, \ldots, m\).

2.3. Samples extraction with sliding window
The input data of the CNN model with \(p\) forecast period is a two-dimensional matrix that includes rainfall and discharges with equal interval time, and the output is predicted outlet discharge. A data sample for training and test is shown as follow.

\[
s = \begin{bmatrix}
I = \\
\vdots \\
\vdots \\
O = q_{tk+p}
\end{bmatrix}
\]

Where, \(I\) represent the two-dimensional input matrix of \(w \times (m+1)\). \(w\) represents the row number of the input data matrix, indicating the impact time of rainfall on the outlet discharge, \(m\) is the number of rainfall stations. \(t_k\) presents the closest monitoring time to prediction. \(q_0, q_1, q_2, \ldots, q_n\) represent the outlet discharge sequences, and \(d_1, \ldots, d_m\) are the rainfall sequences of different rainfall stations that affect the outlet discharge in the watershed. Output \(O\) represents the predicted outlet discharge.

The historical floods and rainfall data could be sampled by a slid window with step 1 and prediction period \(p\) to produce a training and test set. In order to eliminate the difference of magnitude and unit in input data, proper normalization methods, such as z-score and max-min normalization, should be used to normalize the data.

2.4. Structure selection and parameter training
This step mainly uses the reinforcement learning method to search the CNN structure to find the optimal network configuration parameters when the number of layers is three. The search model includes a controller and a convolutional neural network that needs to be optimized. The controller is a NAS network structure[10] and an intensive learning method for training and rewarding. The output of the controller is 3-dimensional, which indicates the size of the convolution kernel, the number of convolution kernels, and the pooling layer size in each layer of CNN.

2.5. Evaluation of model
In order to evaluate the performance of the model, the Deterministic Coefficient(DC), flood peak relative error (PRE) are adopted as the evaluation criteria. These evaluation criteria are commonly used in the field of flood forecasting. DC represents the accuracy of the flood process as a whole, and PRE represents the flood peak characteristics. So these evaluation criteria are sufficient to explain the stability and accuracy of the prediction model.

The value of DC ranges from \(-\infty\) to 1, and the closer the value is to 1, the better the accuracy of model prediction is and the higher the credibility of the model is; the closer the value is to 0, the closer the model prediction is to the average level, and the model prediction is generally credible, but there are large errors; the value far less than 0 means the prediction not credible. DC is calculated by the following formula:
\[ DC = 1 - \sum_{i=1}^{n} \left( y^{(i)} - o^{(i)} \right)^2 / \sum_{i=1}^{n} \left( y^{(i)} - \bar{y} \right)^2 \]  

(2)

Where, \( y^{(i)} \) presents the \( i \)th observed discharge of flood process, \( o^{(i)} \) presents the \( i \)th predicted discharge, \( \bar{y} \) presents the average of the observed discharge, and \( n \) presents the number of observed discharge.

PRE is calculated by the following formula:

\[ \text{PRE} = \frac{|y - o|}{y} \]  

(3)

Where, \( y \) presents the observed peak discharge and \( o \) presents the predicted peak discharge.

3. Experimental evaluation

3.1. Data set

Tunxi watershed, in Anhui Province, China, contains 11 rainfall stations and one outlet hydrological station. 37 historical flood data of Tunxi watershed from 1981 to 2003 are used as experimental data, and the time interval of observation is 1 hour. 28 flood processes selected randomly were used to train the forecasting model, and the last 9 were used to test the forecasting model.

3.2. Experimental setup

Six forecasting models with a 1-6h forecast period are established respectively to verify the validation of the method. Flood data were extracted by sliding window with step 1, and discharge was extracted by sliding window with step 1-6 respectively to create samples.

3.3. Parameter analysis

3.3.1. Length of input data

Fig. 2 shows the Pearson correlation coefficients with 0–20 hours delay in the Tunxi watershed. The max\{ \( w^{(i)} \) \} appears at the \( t-11 \) time of the 6th rainfall station, so \( w \), the length of input data, is set to 11 for Tunxi. The input data is a two-dimensional matrix of 11×12, and contains rainfall and discharge in the 11 hours before forecasting.

![Fig. 2 Pearson correlation coefficients between rainfall and outlet discharge](image_url)
3.3.2. Structure of CNN
Through the analysis of section 2.4, we get the following network architecture: the input layer includes 132 nodes, the kernel of the convolution layer is 5x5, and the kernel of the max-pooling layer is 5x5, the full connection layer includes 132 nodes, output layer includes 1 node.

3.4. Experimental results and analysis

3.4.1. DC analysis
The DC of 9 test floods for the 1-6 hours forecast period is shown in Fig.3.

![Fig.3 DC of test floods for 1-6 hours forecast period](image)

It can be seen that, with the increase of the forecast period, the DC all show a downward trend, but the DC of 9 test floods are all higher than 0.90, reaching A-level, except one case in which DC is 0.88 and reaches B-level.

3.4.2. PRE analysis
Fig. 4 shows the PRE of the 1-6h forecast period, in 1h and 2h forecast period, the PRE is less than 5% except for one case (the error is 7.6%); In the 3-6h forecast period, the PRE are all less than 20% except one case. The experimental results meet the requirement "error for rainfall-runoff forecasting should be less than 20% of the observed peak discharge"[11] with 100% in the 1-4h forecast period and with 89% in the 1-6h forecast period.

![Fig.4 PRE of test floods for the 1-6h forecast period](image)
From the analysis of experimental results, we can see the forecasting model based on CNN can meet the forecasting requirement for the Tunxi watershed. and the model can be applied in the other watershed with a similar climatic environment.

4. Conclusions
A flood forecast model based on CNN for small and medium-sized watershed is proposed in this paper, the experimental results in the Tunxi watershed show that the model can meet the forecasting requirement. But there are still some problems in the model, for example, many parameters of model structure, such as the number of layers of CNN, the number of nodes of FC, etc. are selected based on experience without considering the rainfall during the forecast period. A more reasonable selection method of model structure parameters will be further studied and combined the further rainfall to improve the accuracy of the flood forecast.

Acknowledgments
This work was supported by the National Key Technologies Research and Development Program of China, grant number 2018YFC1508106.

Reference
[1] Crawford NH and Linsley R K 1996 Digital simulation in hydrology: Stanford watershed model (Stanford Univ Press)
[2] Renjun Z 1983 Watershed hydrological model: Xin'anjiang Model and Northern Shanxi model (Water conservancy and electric power press) pp 40-45
[3] Zhijia L, Cheng Y and Ke Zh J 2017 Research and application of the high-resolution rainfall runoff hydrological model in flood forecasting Journal of Hohai (University Natural Sciences) 45 pp 472-480
[4] H.Hidayat, A.J.F.Hoitink and M.G.Sasssi J 2014 Prediction of discharge in a tidal Watershed Using Artificial Neural Networks Journal of Hydrologic Engineering 19 pp 481-486
[5] Z.Z.Latt, Hartmut and Wittenberg J 2014 Improving Flood Forecasting in a Developing Country:A Comparative Study of Stepwise Multiple Linear Regression and Artificial Neural Network. Water Resources Management 28 pp 2109-2128
[6] Kexin L, Weimin B, Jiajia Q, Jiajia L and Huilian Sh J 2015 Application of K-mean cluster method to flood forecasting based on principal component analysis Engineering Journal of Wuhan University 48 pp 447-450
[7] Kaikai M, Shijin L, Jimin W and Yufeng Y J 2016 Comparative study of data-driven intelligent flood forecasting methods for small- and medium-sized watersheds Journal of University of Science and Technology of China 46 pp 774-779
[8] B B Le Cun, J S Denker, D Henderson, R E Howard, W Hubbard and L D Jackel 1989 Proc Handwritten digit recognition with a back-propagation network Proceedings of the Advances in Neural Information Processing Systems pp 396–404
[9] Gu J, Wang Z and Kuen J 2018 Recent Advances in Convolutional Neural Networks J. Pattern Recognition 77 pp 354-377
[10] Elsken, Thomas, Metzen, Jan Hendrik and Hutter 2018 Proc Simple and efficient architecture search for convolutional neural networks. Frank 36th International Conference on Learning Representations
[11] Zhiyu L, Aizhong H, Zhijie Y 2016 Manual on Flood Forecasting and Warning(China Water&Power Press)