GEORGE BROWN PRESS
JOURNALS, BOOKS & PROCEEDINGS

Journal of Computer, Signal, and System Research

Article

Research on the Application of Machine Learning Technology in Hydrological Flood Prediction

Yuan Gao [1,*]

[1] College of Control and Computer Engineering, North China Electric Power University, Baoding, Hebei, China
* Correspondence: Yuan Gao, College of Control and Computer Engineering, North China Electric Power University, Baoding, Hebei, China

Abstract: Urban flooding disasters frequently occur in our country, severely affecting the national development process, anticipating the probability and severity of floods can effectively reduce the negative impacts caused by floods, the rapid progress of hydrology has accelerated the development of flood prediction research. Currently, a lot of machine learning methods are widely applied in the field of flood forecasting based on hydrology, which holds great significance for social development. First, the hydrological models currently used for flood forecasting are introduced. Then, the application of machine learning models in hydrology is elaborated. Finally, the problems and challenges faced by machine learning in flood prediction are analyzed and summarized, and prospects for future flood prediction technologies are discussed.

Keywords: urban floods; flood forecasting; machine learning; hydrology

1. Introduction

Floods are one of the most destructive natural disasters, causing significant negative impacts on human life safety, social infrastructure, agricultural production, and socio-economic systems. Therefore, governments around the world need to conduct accurate and reliable flood forecasting, while also need to develop management strategies centered on preventing flood disasters and reduce flood risks. How to efficiently and accurately simulate floods is an urgent problem that needs to be solved. This paper mainly elaborates on the application of machine learning methods in flood forecasting, analyzing the applications of traditional machine learning methods and neural network methods in flood prediction based on hydrological models.

2. Flood Prediction Based on Machine Learning and Hydrology

The hydrological model follows the principle of water balance and uses various physical equations to describe various hydrological processes such as infiltration and runoff [1]. By integrating hydrological principles, terrain data, meteorological data, and other information, combined with knowledge of physics, mathematics, and hydrology, it is possible to effectively conduct quantitative analysis and prediction of floods. In hydrological models, terrain data is used to describe the topography and water flow paths of the watershed, while meteorological data provides meteorological elements such as rainfall and water evaporation, which are important components of the model input [2]. Hydrological models can simulate the development and evolution of different hydrological processes based on different model structures, such as the formation process of floods and the prediction of peak discharge. By simulating hydrological responses in different scenarios, hydrological models can help decision-makers develop effective response measures, mitigate losses caused by floods, and ensure the safety of people's lives and property.

Received: 13 February 2025
Revised: 25 February 2025
Accepted: 09 March 2025
Published: 14 March 2025

2.1. Hydrological Methods Based on Traditional Machine Learning

Although hydrological models can describe hydrological conditions well, they require analysis of a large amount of terrain data and historical rainfall data [3], and the parameterization process of the model has significant uncertainty [4]. With the continuous development and improvement of machine learning technology, machine learning models have been widely used for flood prediction.

Mathematics et al. proposed using the Autoregressive Moving Average (ARMA) method for hydrological prediction, and applied the ARMA model to study the temporal variation of reservoir water level [5]. The ARMA model performs poorly in handling nonlinear and non-stationary data, and requires data to have distinct seasonal characteristics. In addition, the stability of the model is highly dependent on the selection of parameters, and these limitations may lead to poor predictive performance. To reduce the negative impact of these factors, Mathematics et al. proposed using the Autoregressive Integrated Moving Average (ARIMA) model for prediction. The ARIMA model adds differential processing to the ARMA model, making it more suitable for non-seasonal data and having smaller errors in hydrological forecasting. Although ARMA and ARIMA models can make relatively accurate predictions, these models are suitable for rough seasonal forecasting and have large errors in short-term predictions. In order to make more accurate predictions, Haddad et al. proposed a framework that combines Region Of Influence (ROI) and Bayesian Generalized Least Squares (BGLS) [6]. Based on quantile regression technique (QRT), predicting floods in the eastern region of Australia and improving the selection of predictive variables can lead to more accurate predictions. However, for areas with low frequency of floods, the predicted flood levels are generally higher than the actual values. Kroll et al. assumed floods as random events, described the minimum annual flow using a probability distribution, and predicted floods using the probability distribution of historical water flow data [7]. However, this method is not suitable for short-term forecasting and relies on high-precision measurement data. It cannot make quantitative predictions and requires analysis and processing of a large amount of historical data to obtain reliable long-term forecasting results.

Considering the strong randomness and high complexity of hydrological data, researchers adopted a combined prediction method of wavelet transform and Support Vector Machine (SVM) to train and analyze the data in the Tunxi River Basin [8]. The basic idea of SVM is to use linear models to solve nonlinear problems [9], mapping the input space to a high-dimensional space.

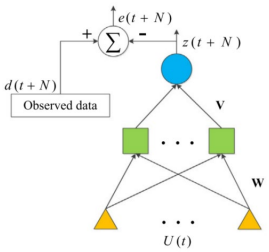
Random forests perform well in processing high-dimensional and large datasets, without the need for feature scaling and data normalization, and have high accuracy and stability. Due to the advantages of the random forest method, Pao Shan Yu et al. used historical data, grid position (latitude and longitude of the study area), and grid elevation data of grid based radar derived rainfall data as input variables, and grid based radar derived data as output variables for model training [10]. Two methods, Single mode Forecasting Model (SMFM) based on random forest and SMFM based on SVM, were used for prediction. It was found that both methods had relatively accurate prediction results for hydrological forecasting 1-3 hours in advance. However, the author did not verify the feasibility of the model in complex environments, nor did they consider the impact of neighboring areas on flood levels in the study area.

2.2. Hydrological Method Based on Neural Network

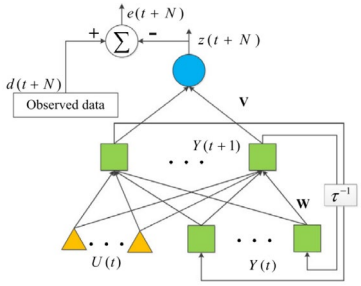
Artificial Neural Network (ANN) is capable of solving complex nonlinear relationships between input and output sets, and has flexible mathematical computing capabilities. It is widely used in the fields of water flow prediction and precipitation prediction. Adamowski et al. used artificial neural networks to perform regression and time series analysis on hydrological data in the Ottawa area of Canada [11]. The results showed that the performance of artificial neural networks in prediction was superior to that of multiple

linear regression methods. However, artificial neural networks have poor ability to process non-stationary data, resulting in significant prediction errors. Shiri et al. proposed combining wavelet analysis (WA) with artificial neural network methods to obtain a WA-ANN prediction model, which can make accurate long-term and short-term predictions [12]. By coupling the two models for prediction, it is usually better than using autoregressive integrated moving average (ARIMA) and artificial neural network separately.

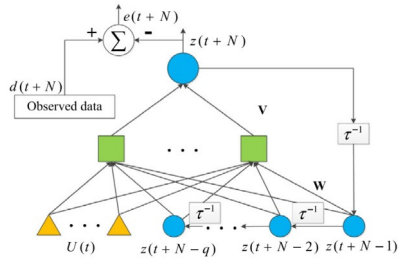
Lu Liu et al. investigated the flood forecasting capabilities of the SIMHYD hydrological model and the Long Short-Term Memory (LSTM) neural network model in 232 basins under different climatic conditions [13]. In typical flood evolution scenarios, compared with traditional hydrological models, LSTM neural networks demonstrated better predictive performance with longer training periods and shorter validation periods. In extreme cases, the predictive ability of LSTM models may be lower than that of SIMHYD hydrological models, because LSTM lacks data for extreme situations during the training process. Chang F-J et al. used rainfall and Floodwater Storage Pond (FSP) data for flood prediction, mainly using static neural networks: Back Propagation Neural Network (BPNN) [14], Dynamic Recursive Neural Network (Elman NN), and Nonlinear Autoregressive with eXogenous Input Network (NARX network) for prediction, the network structure is shown in Figure 1. Artificial neural networks have the ability to approximate nonlinearity, and are therefore often used for research such as flood prediction and rainfall prediction [15].



(a) Backpropagation neural network



(b) Elman neural network



(c) Nonlinear autoregressive with eXogenous input (NARX) network

Figure 1. Three network architectures: BPNN, Elman NN, and NARX [14].

All three types of neural networks can perform well in urban hydrological prediction, but due to the adaptive selection of time delay values and consideration of historical data and external factors, NARX network can effectively alleviate the delay problem, while BPNN and Elman NN cannot effectively alleviate this problem. Compared with the other two network models, the NARX network has a wider applicability and higher prediction accuracy. These three models can ensure a high degree of consistency with the observed values in a shorter prediction time, but the error is larger when the prediction time is longer. Yen Ming et al. used RNN to predict the urban sewage water level at both measured and unmeasured points, which can accurately predict the water level at the measured points [16]. For unmeasured points without historical data support, the Storm Water Management Model (SWMM) is used to predict water levels by combining information from surrounding measured points. Accurate prediction results can also be obtained at unmeasured points. Overall, although RNN can predict the main trend of water level changes, it significantly underestimates the peak water volume. The neural network architecture adopted by Yen Ming et al. is shown in Figure 2.

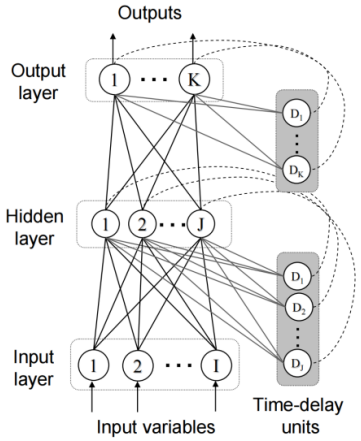


Figure 2. The RNN neural network architecture adopted by Chiang et al [16].

Considering the complexity and dynamics of flash floods, Bui et al. used a deep learning neural network (DLNN) to predict flash floods in a typical mountainous area in northwest Vietnam [17]. The model adopts a network structure containing 3 hidden layers and 192 neurons, and the accuracy of training and prediction stages can reach over 96%. The DLNN model has good flexibility and generalization ability. Through experimental comparison, it was found that the accuracy of prediction using DLNN is higher than that of MLN-NN and SVM methods. The combination of GIS and DLNN can effectively improve the accuracy of flood prediction. The DLNN neural network architecture adopted by Bui et al. is shown in Figure 3.

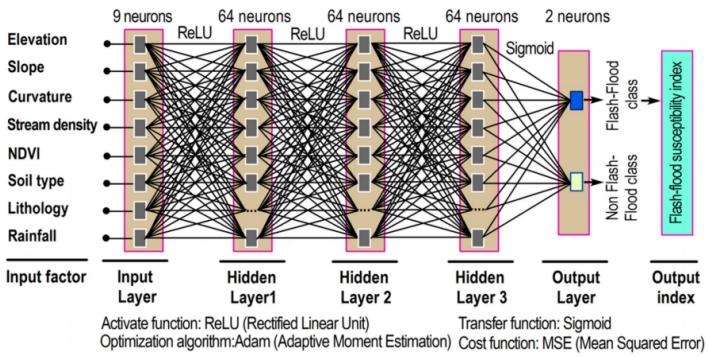


Figure 3. DLNN Neural Network Architecture for Flood Prediction [17].

Due to the problems of low prediction accuracy and overfitting in traditional neural network models for flood prediction. Guo et al. used Convolutional Neural Networks (CNN) to train flood simulation data obtained from 18 rainfall flow maps in three catchment areas. The topographic map and rainfall flow map were divided into several equally sized blocks, and each block was trained instead of training the entire catchment area [18]. The flood prediction time of this method is 0.5% of that of physical model-based flood prediction methods, and it can achieve good prediction results for rainfall events that do not exist in the training set. However, when the terrain being studied changes, it is necessary to retrain the flood prediction model. The neural network architecture adopted by Guo et al. is shown in Figure 4.

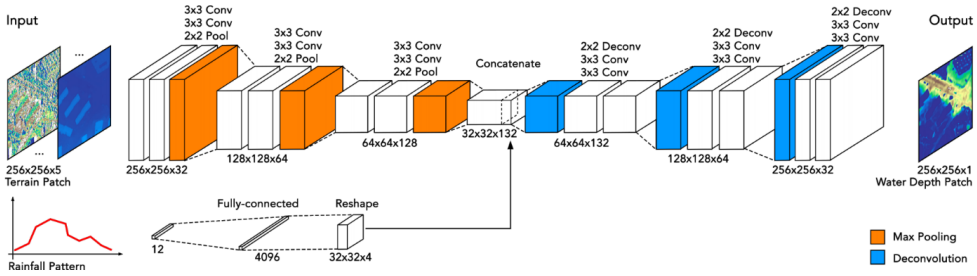


Figure 4. CNN based flood prediction architecture [18].

Lima et al. used a feedforward neural network called Extreme Learning Machine (ELM) for daily water volume prediction [19]. ELM is a single-layer feedforward neural network that has faster training speed and stronger generalization ability compared to traditional neural networks. When the input and output of the model exhibit a non-linear relationship, ELM can achieve good prediction performance through multiple training sessions. However, due to the random generation of weights and biases (Hidden Nodes Or Neurons, HN) from the input layer to the hidden layer at the beginning of ELM, if the initial number of randomly generated HNs is too small, the prediction error will increase as the training data increases. The ELM neural network structure adopted by Lima et al. is shown in Figure 5.

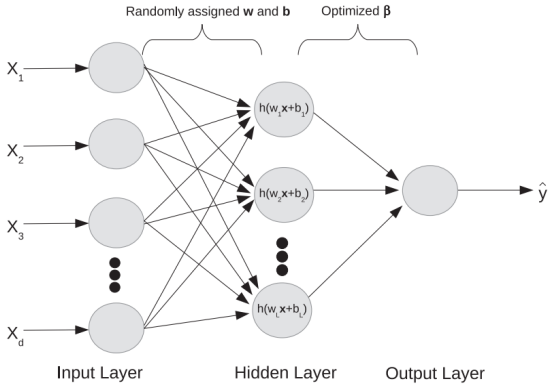


Figure 5. ELM network architecture diagram adopted by Lima et al [19].

In summary, compared with traditional physical prediction models, data-driven hydrological prediction can obtain flood prediction models that meet the requirements in a short period of time. Although this method can accurately describe urban flood water level information, hydrological models often simplify hydrological processes and ignore complex hydrodynamic processes, lacking detailed spatial dynamic information of urban surface floods, such as depth and velocity, and cannot simulate changes in the physical characteristics of floods, such as momentum and mass. And when the training dataset is

limited, the model may have overfitting issues, resulting in poor performance on new, unseen data, thereby increasing prediction errors.

3. Conclusion and discussion

In summary, although the prediction method combining hydrology and machine learning can provide flood prediction to a certain extent, its simulation results lack accurate reflection of the real physical characteristics of floods due to the neglect of the physical properties of fluids in hydrological models. Especially in complex scenarios, hydrological models often cannot capture the complex motion of floods, and when the amount of data is small, the prediction error is large.

Future research should focus on addressing how to achieve accurate predictions on smaller datasets, how to improve the accuracy and efficiency of prediction models in complex scenarios, and how to enhance the generalization ability of prediction models. Further exploration in these areas will provide more reliable and effective methods for flood prediction. Meanwhile, combining the advantages of hydrology and hydrodynamics to develop more comprehensive and accurate prediction methods is also an important direction for future research.

References

1. C. Deng and W. Wang, "A two-stage partitioning monthly model and assessment of its performance on runoff modeling," *J. Hydrol.*, vol. 592, 2021, doi: 10.1016/j.jhydrol.2020.125829.
2. P. A. Mendoza, J. McPhee, and X. Vargas, "Uncertainty in flood forecasting: A distributed modeling approach in a sparse data catchment," *Water Resources Research*, vol. 48, no. 9, 2012, doi: 10.1029/2011WR011089.
3. E. F. Wood, J. K. Roundy, T. J. Troy, et al., "Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water," *Water Resour. Res.*, vol. 47, no. 5, 2011, doi: 10.1029/2010WR010090.
4. C. Paniconi and M. Putti, "Physically based modeling in catchment hydrology at 50: Survey and outlook," *Water Resour. Res.*, vol. 51, pp. 7090-7129, 2015, doi: 10.1002/2015WR017780.
5. M. Valipour, M. E. Banihabib, and S. M. R. Behbahani, "Parameters estimate of autoregressive moving average and autoregressive integrated moving average models and compare their ability for inflow forecasting," *J. Math. Stat.*, vol. 8, no. 3, pp. 330-338, 2012, doi: 10.3844/jmssp.2012.330.338.
6. K. Haddad and A. Rahman, "Regional flood frequency analysis in eastern Australia: Bayesian GLS regression-based methods within fixed region and ROI framework-Quantile Regression vs. Parameter Regression Technique," *J. Hydrol.*, vol. 430-431, pp. 142-161, 2012, doi: 10.1016/j.jhydrol.2012.02.012.
7. C. N. Kroll and R. M. Vogel, "Probability distribution of low streamflow series in the United States," *J. Hydrol. Eng.*, vol. 7, no. 2, pp. 137-146, 2002, doi: 10.1061/(ASCE)1084-0699(2002)7:2(137).
8. M. Pan, H. Zhou, J. Cao, Y. Liu, J. Hao, S. Li, & C.-H. Chen., "Water Level Prediction Model Based on GRU and CNN," *IEEE Access*, vol. 8, pp. 60090-60100, 2020, doi: 10.1109/ACCESS.2020.2982433.
9. Y. B. Dibike, S. Velickov, D. Solomatine, et al., "Model induction with support vector machines: Introduction and applications," *J. Comput. Civil Eng.*, 2001, doi: 10.1061/(ASCE)0887-3801(2001)15:3(208).
10. P.-S. Yu, T.-C. Yang, S.-Y. Chen, C.-M. Kuo, and H.-W. Tseng, "Comparison of random forests and support vector machine for real-time radar-derived rainfall forecasting," *J. Hydrol.*, vol. 552, pp. 92-104, 2017, doi: 10.1016/j.jhydrol.2017.06.020.
11. J. F. Adamowski, "Development of a short-term river flood forecasting method for snowmelt driven floods based on wavelet and cross-wavelet analysis," *J. Hydrol.*, vol. 353, no. 3, pp. 247-266, 2008, doi: 10.1016/j.jhydrol.2008.02.013.
12. J. Shiri and O. Kisi, "Short-term and long-term streamflow forecasting using a wavelet and neuro-fuzzy conjunction model," *J. Hydrol.*, vol. 394, no. 3-4, pp. 486-493, 2010, doi: 10.1016/j.jhydrol.2010.10.008.
13. L. Liu, X. Liu, P. Bai, et al., "Comparison of flood simulation capabilities of a hydrologic model and a machine learning model," *Int. J. Climatol.*, 2023, doi: 10.1002/joc.7738.
14. F. J. Chang, P. A. Chen, Y. R. Lu, et al., "Real-time multi-step-ahead water level forecasting by recurrent neural networks for urban flood control," *J. Hydrol.*, vol. 517, pp. 836-846, 2014, doi: 10.1016/j.jhydrol.2014.06.013.
15. N. Q. Hung, M. S. Babel, S. Weesakul, N. K. J. H. Tripathi, and E. S. Sciences, "An artificial neural network model for rainfall forecasting in Bangkok, Thailand," 2009, vol. 13, doi: 10.5194/hess-13-1413-2009.
16. Y.-M. C. Yen-Ming, L.-C. C. Li-Chiu, T. M.-J. T., et al., "Dynamic neural networks for real-time water level predictions of sewerage systems—covering gauged and ungauged sites," *Hydrol. Earth Syst. Sci.*, vol. 14, no. 7, pp. 2317-2345, 2010, doi: 10.5194/hess-14-1309-2010.

17. D. T. Bui, N. D. Hoang, and M. Martinez-Alvarez, et al., "A novel deep learning neural network approach for predicting flash flood susceptibility: A case study at a high frequency tropical storm area," *Sci. Total Environ.*, vol. 701, p. 134413, 2020, doi: 10.1016/j.scitotenv.2019.134413.
18. J. P. Leito, N. E. Simes, Z. Guo, et al., "Data-driven flood emulation: Speeding up urban flood predictions by deep convolutional neural networks," *J. Flood Risk Manage.*, vol. 14, no. 1, pp. n/a-n/a, 2021, doi: 10.1111/jfr3.12684.
19. A. R. Lima, A. J. Cannon, and W. W. Hsieh, "Forecasting daily streamflow using online sequential extreme learning machines," *J. Hydrol.*, vol. 537, pp. 431-443, 2016, doi: 10.1016/j.jhydrol.2016.03.017.