

Flood Forecasting Scheme Based on Artificial Neural Networks

Yuxuan Ma

*College of New Energy and Environment, Jilin University, Changchun, China
mayx9923@mails.jlu.edu.cn*

Abstract. Flood prediction is of great significance for safeguarding lives and property. Existing methods mostly employ linear models such as logistic regression, which, despite their simplicity, struggle to capture the nonlinear relationships between multiple factors and flood risk, resulting in limited predictive accuracy. This paper constructs a flood prediction model based on a BP neural network, selecting nine flood impact factors and normalizing them, with logistic regression as a benchmark model. The model performance is comprehensively evaluated on the test set in terms of R^2 , RMSE, and accuracy. Results show that the BP neural network achieves an R^2 of 0.9958 and an RMSE of only 0.0163 on the test set, far superior to the logistic regression model ($R^2=0.7187$, RMSE = 0.1338), verifying the effectiveness of BP neural networks in flood prediction.

Keywords: flood prediction, BP neural network, logistic regression, nonlinear modeling, machine learning

1. Introduction

Under the dual drivers of global climate change and rapid urbanization, the frequency and intensity of flood disasters have shown a significant upward trend [1]. In urban areas, the expansion of impervious surfaces and the reduction of natural storage capacity make rainfall-induced flooding particularly prominent. Flood prediction, as a core component of non-structural flood control measures, can provide emergency management departments with valuable early warning time and plays an irreplaceable role in reducing casualties and economic losses. However, the formation mechanism of floods is extremely complex, influenced by multiple factors such as rainfall intensity and spatiotemporal distribution, topography, antecedent soil moisture, drainage network capacity, and land use type. These factors exhibit strong nonlinearity and spatiotemporal heterogeneity, posing great challenges to accurate prediction.

Traditional flood prediction methods mainly rely on hydrological and hydrodynamic numerical models (e.g., HEC-HMS [2], FLO-2D [3]), which simulate runoff generation and concentration processes based on physical mechanisms and theoretically offer high reliability. Nevertheless, fine-scale two-dimensional hydrodynamic models are computationally expensive, with a single simulation taking tens of minutes to hours, making it difficult to meet the response time requirements for real-time warning [4]. In recent years, data-driven machine learning approaches have provided new technical pathways for flood prediction. Logistic regression, as a classic linear

statistical model, is widely used in flood susceptibility mapping and risk probability assessment due to its simple structure and strong interpretability [5]. However, linear models assume a linear relationship between independent variables and the dependent variable. When faced with the complex nonlinear coupling effects in rainfall-runoff-inundation processes, their fitting capacity is inherently limited, and predictive accuracy often cannot be further improved.

To address these issues, this paper proposes a flood prediction model based on a backpropagation (BP) neural network. With its powerful nonlinear mapping ability and adaptability, the BP neural network can automatically extract high-order interaction information from multi-source input features and has shown potential superior to traditional linear models in hydrological forecasting and flood inundation simulation [6]. Nine key flood impact factors are selected as input features: elevation, slope, drainage density, rainfall, soil moisture, vegetation coverage, drainage capacity index, impervious surface ratio, and distance to reservoir [7]. After normalization to eliminate dimensional effects, a BP neural network with two hidden layers (24-12) is constructed, while logistic regression is introduced as a baseline model for comparative analysis. Model performance is comprehensively evaluated using R^2 , RMSE, and accuracy.

The main contributions of this paper are: (1) verifying the significant performance advantage of BP neural networks over logistic regression in flood risk prediction, providing empirical support for the application of nonlinear models in flood forecasting; (2) revealing the performance gap between linear and nonlinear models in flood prediction tasks through rigorous experimental design and comparative analysis, offering references for neural network parameter configuration; (3) establishing a reusable data preprocessing and model training pipeline that can serve as a methodological reference for related flood forecasting studies.

2. Related concepts and key technologies

2.1. Logistic regression

Logistic regression [8] is a commonly used statistical method for studying the relationship between a binary dependent variable and one or more independent variables. Unlike multiple linear regression, which is suitable for continuous dependent variables, logistic regression applies to response variables following a binomial distribution. Standard linear regression requires a continuous dependent variable, whereas logistic regression takes continuous independent variables as input but outputs discrete binary outcomes (e.g., 0 = not inundated, 1 = inundated). It is essentially a binary classification algorithm that divides samples into two categories. In practice, logistic regression can use whether a grid cell is inundated (1 or 0) to generate a flood probability discriminant for each grid; once conditional factors such as rainfall-induced runoff are input, the flood probability of the target area can be quickly computed.

2.2. Artificial neural network

Artificial neural networks (ANNs) [9] are computational models inspired by biological nervous systems and can effectively handle nonlinear relationships among multiple factors such as rainfall, topography, and soil moisture in flood forecasting. The BP neural network is one of the most widely used ANN architectures, typically consisting of an input layer, hidden layers, and an output layer. The input layer receives feature data of flood impact factors (e.g., rainfall, slope, soil moisture), the hidden layers perform nonlinear transformation and feature extraction on the input information, and the output layer generates the final flood prediction result (e.g., flood occurrence probability or

inundation depth). Neurons between layers are connected by weights, and the magnitude of weights reflects the influence of input features on prediction results. The core of the BP neural network lies in the error backpropagation algorithm, which uses gradient descent to propagate output errors backward from the output layer to the input layer, continuously adjusting connection weights and thresholds so that the network output gradually approaches the desired value. In the field of flood forecasting, BP neural networks have been widely applied to tasks such as flood level prediction, inundation extent mapping, and flood susceptibility mapping, demonstrating higher forecasting accuracy than traditional statistical methods. Therefore, this paper adopts the BP neural network for flood prediction and compares it with logistic regression.

3. Methodology

3.1. Flood prediction model based on logistic regression

Based on flood impact factor data, this study selects nine feature variables as inputs: elevation, slope, drainage density, rainfall, soil moisture, vegetation coverage, drainage capacity index, impervious surface ratio, and distance to reservoir. The outputs are flood risk value (continuous, 0–1) and flood occurrence label (binary, 0/1). A logistic regression prediction model is constructed accordingly.

Because the input features differ greatly in magnitude and range—for example, elevation can reach hundreds to thousands of meters, while soil moisture ranges between 0 and 1—using raw data directly may cause features with larger numerical ranges to dominate model training, affecting parameter estimation stability and predictive performance. Therefore, before model training, the input feature matrix is normalized. This paper applies Min-Max normalization to linearly transform all features to the $[0,1]$ interval, with the formula: $x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$, where x is the original value, and $x_{\min}$ and $x_{\max}$ are the minimum and maximum of that feature. The normalization parameters are computed solely from the training set, and the same transformation is applied to the test set to avoid data leakage and ensure objective model evaluation.

To properly assess generalization ability, the original dataset is split into training and test sets at a 7:3 ratio, with 3,750 samples in the training set and 1,250 in the test set. The training set is used for parameter estimation, and the test set is used for independent performance evaluation.

The logistic regression model combines input variables linearly and maps the result to the $(0,1)$ interval via the Sigmoid function to output flood probability. Let the normalized input feature vector be $\mathbf{x} = (x_1, x_2, \dots, x_9)$. The model is formulated as: $P(y = 1 | \mathbf{x}) = \frac{1}{1 + e^{-(w_0 + w_1x_1 + \dots + w_9x_9)}}$, where $y = 1$ indicates flood occurrence and $y = 0$ indicates no flood; w_0 is the intercept, and $w_1, \dots, w_9$ are the regression coefficients for the input features. Maximum likelihood estimation is used to estimate the parameters, and a default threshold of 0.5 is applied for classification—when $P(y = 1 | \mathbf{x}) \geq 0.5$, the prediction is flood occurrence; otherwise, no flood. To comprehensively evaluate the logistic regression model's performance, we adopt the coefficient of determination R^2 , root mean square error (RMSE), and accuracy as evaluation metrics.

3.2. Flood prediction based on BP neural network

To further exploit potential nonlinear relationships between flood impact factors and risk values, this paper constructs a BP neural network flood prediction model and compares it with logistic regression.

Consistent with the logistic regression model, Min-Max normalization is first applied to the original input features. The normalized features have zero mean and unit standard deviation, which helps improve convergence speed and stability during neural network training. Normalization parameters are also computed solely from the training set and applied to the test set to prevent test-set information from leaking into the training phase.

For dataset splitting, the same 7:3 ratio is adopted, ensuring that both models use identical training and test samples, thus guaranteeing a fair comparison.

The BP neural network model in this study consists of an input layer, hidden layers, and an output layer. The input layer has 9 neurons, corresponding to the nine input features (elevation, slope, drainage density, rainfall, soil moisture, vegetation coverage, drainage capacity index, impervious surface ratio, and distance to reservoir). The output layer has 1 neuron, outputting the continuous flood risk prediction value. Through experimental optimization, the network adopts two hidden layers with 24 and 12 neurons, respectively, resulting in a 9-24-12-1 architecture.

For activation functions, the ReLU (Rectified Linear Unit) function is used in the hidden layers, defined as $f(x)=\max(0,x)$. ReLU effectively mitigates the vanishing gradient problem commonly encountered with the Sigmoid function in deeper networks, while offering computational simplicity and faster convergence; it is widely employed in neural network modeling. The output layer uses a linear activation function (Purelin) to accommodate the regression task where flood risk values are continuous real numbers. During training, the Adam optimization algorithm is employed to update network weights, combining the advantages of momentum and adaptive learning rates, which improves training efficiency and convergence stability. Additionally, regularization parameters are set to reduce overfitting risk; the initial learning rate is 0.001, the maximum number of epochs is 200, and an early stopping strategy is adopted to prevent overfitting.

After network training, the normalized features of the test set are fed into the trained BP neural network to obtain predicted flood risk values for each test sample. Similar to logistic regression, we use R^2 , RMSE, and accuracy to evaluate the BP neural network's performance on the test set. By comparing the predictive performance of the two models on the test set, we analyze the applicability and advantages of traditional linear classification methods versus nonlinear neural network approaches for flood prediction.

4. Experiments

4.1. Experimental setup

Data preprocessing and model training were conducted on Python 3.14, mainly using Pandas and NumPy for data reading and normalization, Scikit-learn for building the logistic regression model, and TensorFlow/Keras for constructing the BP neural network. The original dataset was randomly split into a training set (3,750 samples) and a test set (1,250 samples) at a 7:3 ratio. Min-Max normalization linearly transformed the nine input features to the [0,1] interval to eliminate dimensional differences. Statistical analysis and visualization of the results were carried out on MATLAB R2023a, leveraging its scientific computing and plotting capabilities to generate scatter plots comparing predicted and actual values, and to output performance metrics tables (R^2 , RMSE, accuracy) to meet the quality and layout requirements of the paper. For evaluation metrics, we used R^2 to measure the proportion of variance explained by the model, RMSE to measure prediction deviation, and accuracy computed by converting continuous probabilities to binary labels using a threshold of 0.5.

4.2. Experimental results

4.2.1. Flood prediction results based on logistic regression

Table 1 presents the prediction performance of the logistic regression model on the test set. The logistic regression model achieved an R^2 of 0.7187, an RMSE of 0.1338, and an accuracy of 0.7112. The R^2 of 0.72 indicates that the nine flood impact factors can explain about 72% of the variation in risk values, showing that the linear model has reasonable explanatory power. The RMSE of 0.134 suggests an average deviation of about 0.13 between predicted and true values, which is moderately good on a 0–1 risk scale. However, about 28% of the risk variation remains unexplained by the linear model, implying that logistic regression has limitations in capturing nonlinear interactions among multiple factors.

Table 1. Logistic regression model prediction results

model	R^2	RMSE	Accuracy
Logistic Regression	0.7187	0.1337	0.7112

Figure 1 shows the scatter plot of predicted versus actual risk values for the logistic regression model on the test set. The horizontal axis is the true risk value, and the vertical axis is the predicted probability. The red dashed line represents the perfect prediction reference. The figure shows that predictions aggregate well in low-risk (<0.2) and high-risk (>0.8) regions, but in the middle range (0.2–0.8) there is considerable scatter, with many points deviating from the diagonal. This indicates that although logistic regression can identify extreme risk cases reasonably well, its predictive ability for moderate-risk samples is limited, and the linear decision boundary cannot adequately capture the complex nonlinear evolution patterns in flood risk formation.

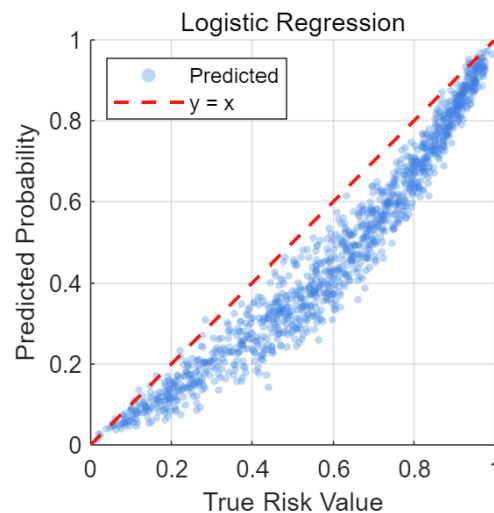


Figure 1. Comparison plot of predicted values and actual values

4.2.2. Flood prediction results based on BP neural network

Table 2 presents the prediction performance of the BP neural network on the test set. The BP neural network achieved an R^2 of 0.9958, an RMSE as low as 0.0163, and an accuracy of 0.6872. Compared with logistic regression, R^2 increased from 0.7187 to 0.9958, an improvement of about

38.6 percentage points; RMSE decreased from 0.1338 to 0.0163, a reduction of approximately 87.8%. This result demonstrates that the BP neural network, through multi-layer nonlinear transformations, can explain almost all the variation in flood risk values, achieving excellent fitting performance on the test set.

Table 2. Prediction performance of BP neural network on the test set

model	Hidden layer structure	R^2	RMSE	Accuracy
BP Neural Network	(24, 12)	0.9958	0.0163	0.6872

Figure 2 shows the scatter plot of predicted versus actual risk values for the BP neural network on the test set. Compared with Figure 1, the prediction points of the BP network almost all cluster tightly around the diagonal, with extremely low dispersion. This indicates that the BP neural network can accurately fit the complex mapping between input features and flood risk, effectively extracting high-order nonlinear interactions among multiple factors. Notably, the predictions remain highly consistent across the entire range (0–1), without any obvious bias concentration region, indicating good full-range fitting ability.

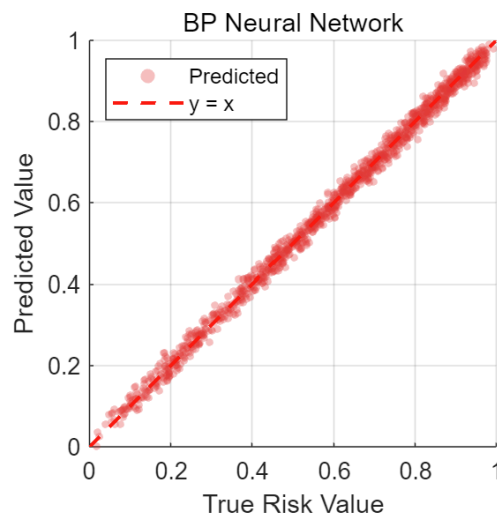


Figure 2. Comparison of BP neural network predicted values and actual values

4.2.3. Model comparison results

Quantitative comparison of the evaluation metrics in Table 3 shows that the BP neural network significantly outperforms logistic regression in continuous risk value prediction. The coefficient of determination R^2 improved from 0.7187 to 0.9958, indicating that the linear framework could explain only about 72% of the risk variation, while the nonlinear neural network explained almost all of it, strongly proving that nonlinear relationships dominate the flood formation mechanism and that linear models suffer from severe underfitting. For RMSE, the BP neural network reduced the error by about 87.8% compared with logistic regression, showing that the BP network's risk predictions are much closer to the true values, with much higher precision.

Table 3. Comparison of prediction performance between two models

Metric	Logistic Regression	BP Neural Network	Improvement
R ²	0.7187	0.9958	+38.6%
RMSE	0.1338	0.0163	-87.8%
Accuracy	0.7112	0.6872	-3.4%

However, for binary classification accuracy, logistic regression (0.7112) slightly outperforms the BP neural network (0.6872). This seemingly contradictory result reveals the fundamental difference between the two models: logistic regression directly optimizes parameters with the classification threshold of 0.5 as the objective, whereas the BP neural network is trained to minimize mean squared error, with continuous risk values as output rather than directly optimizing for classification. When a fixed threshold of 0.5 is used to convert the BP network's continuous outputs into binary labels, the network—designed for precise continuous-value fitting—may not achieve optimal classification performance at that specific threshold. In other words, the BP network "knows" whether the risk is 0.6 or 0.7, but if we only judge based on whether it is greater than 0.5, some information is lost. This suggests that in actual flood early warning, we should not rely solely on binary classification results; instead, we should make full use of the BP network's continuous risk values for refined risk-level early warning—for example, dividing risk into low, medium, high, and very high levels—to maximize the value of the nonlinear model. Future research could also further optimize the BP network's classification performance by adjusting the classification threshold or introducing custom loss functions.

5. Conclusion

This paper addresses the limitation of traditional linear models in capturing nonlinear coupling among multiple factors in flood prediction by constructing a BP neural network-based flood prediction model. Experimental results show that the BP neural network achieves an R² of 0.9958 and an RMSE as low as 0.0163 on the test set, compared with logistic regression (R²= 0.7187, RMSE = 0.1338), representing an improvement of about 38.6% in goodness-of-fit and a reduction of about 87.8% in prediction error. These results fully verify the superior capability of BP neural networks in capturing deep nonlinear synergistic effects among flood impact factors. This study provides an effective technical pathway for data-driven flood prediction and holds significant academic reference value and engineering practical significance for promoting the application of machine learning methods in urban flood control and disaster mitigation.

References

- [1] Nirupama, N., & Simonovic, S. P. (2007). Increase of flood risk due to urbanisation: A Canadian example. *Natural Hazards*, 40(1), 25–41. doi: 10.1007/s11069-006-0003-0
- [2] Khan, I. R., Elmahdy, S. I., Rustum, R., et al. (2024). Floods modeling and analysis for Dubai using HEC-HMS model and remote sensing using GIS. *Scientific Reports*, 14(1), 25039. doi: 10.1038/s41598-024-74736-3
- [3] Mitra, D., Das, D., & Ghosh, A. R. (2021). Flood mapping and prediction using FLO-2D basic model. In *Advances in Water Resources Management for Sustainable Use* (pp. 109–120). Springer Singapore. Singapore. doi: 10.1007/978-981-33-6412-7_9.
- [4] Lee, J., & Kim, B. (2021). Scenario-based real-time flood prediction with logistic regression. *Water*, 13(9), 1191. doi: 10.3390/w13091191

- [5] Waiyasusri, K., Wetchayont, P., Tananonchai, A., et al. (2023). Flood susceptibility mapping using logistic regression analysis in Lam Khan Chu watershed, Chaiyaphum Province, Thailand. *Geography, Environment, Sustainability*, 16(1), 123–134. doi: 10.24057/2071-9388-2022-159
- [6] Liu, F., Xu, F., & Yang, S. (2017). A flood forecasting model based on deep learning algorithm via integrating stacked autoencoders with BP neural network. In *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)* (pp. 58-61). IEEE. doi: 10.1109/BigMM.2017.29
- [7] Tella, A., Mustafa, M. R. U., Balogun, A. O., et al. (2023). Spatial prediction of flood in Kuala Lumpur City of Malaysia using logistic regression. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 363-369. doi: 10.5194/isprs-archives-XLVIII-4-W6-2022-363-2023
- [8] Huang, X., Fang, Y., He, X., et al. (2026). Diversity and differences of flood evacuation strategies among rural households in the Hengduan Mountains, Southwest China. *Geomatics, Natural Hazards and Risk*, 17(1), 2608248. doi: 10.1080/19475705.2025.2608248
- [9] Al-Juaidi, A. E. M. (2026). Integrating hydrological, physical, and chemical factors for soil moisture prediction using advanced machine learning models. *Journal of Hydrologic Engineering*, 31(4), 04026025. doi: 10.1061/JHYEFF.HEENG-6811