SCIENCE CHINA
Technological Sciences

• RESEARCH PAPER •

June 2025, Vol. 68, Iss. 6, 1620701:1–1620701:17
https://doi.org/10.1007/s11431-024-2925-8

A novel framework incorporating machine learning into GIS for flood susceptibility prediction of urban metro systems

Haimin LYU[1,2,3], Zhenyu YIN[4*], Shuilong SHEN[5], Xiangsheng CHEN[1,2,3] & Dong SU[1,2,3]

[1] *Key Laboratory for Resilient Infrastructures of Coastal Cities (MOE), Shenzhen University, Shenzhen 518026, China*
[2] *College of Civil and Transportation Engineering, Shenzhen University, Shenzhen 518000, China*
[3] *State Key Laboratory of Intelligent Geotechnics and Tunnelling, Shenzhen 518000, China*
[4] *Department of Civil and Environmental Engineering, the Hong Kong Polytechnic University, Hong Kong 999077, China*
[5] *College of Engineering, Shantou University, Shantou 515063, China*

Received November 8, 2024; accepted March 17, 2025; published online May 8, 2025

Abstract Floods have become increasingly destructive with climate change, resulting in the inundation of urban metro systems. This study complied with global data on flooded metro lines in recent decades. Based on these data, a framework incorporating machine learning (ML) with geographic information system (GIS) was developed to predict flood susceptibility in urban metro systems. To address the scarcity of subway flooding data, this study proposed a novel approach to generate a database for training and testing using ML and GIS. The 7.20 flood event in Zhengzhou, China, was analyzed as a case study. The optimal ML model was selected by comparing predicted flood states with recorded flooded metro stations. Flood susceptibility for the Zhengzhou metro system under future extreme rainfall scenarios was then predicted. Results demonstrated that the number of flooded stations and their flood susceptibility increased with rainfall intensity. These findings highlight the scale and vulnerability of metro systems, providing critical insights for developing resilient underground infrastructure.

Keywords machine learning, GIS, incorporating framework, flood susceptibility, metro system

Citation: Lyu H M, Yin Z Y, Shen S L, et al. A novel framework incorporating machine learning into GIS for flood susceptibility prediction of urban metro systems. Sci China Tech Sci, 2025, 68(6): 1620701, https://doi.org/10.1007/s11431-024-2925-8

1 Introduction

Floods have emerged as the most pervasive and destructive natural hazard under climate change. Catastrophic floods in countries such as Republic of Korea, Japan, South African countries, the United States, and the United Kingdom in 2022 under-score their global threat [1–3]. Urban floods induced by extreme rainfall damage surface buildings and inundate underground infrastructures like metro lines. Statistics reveal numerous flooded metro stations caused by extreme rainfall between 2010 and 2023 [4], with China experiencing severe cases. For instance, the 2021 Zhengzhou 7.20 flood inundated five metro lines and 18 stations, tragically claiming 14 lives [5]. This event drew global attention to flood prevention in urban metro systems.

The threat of catastrophic floods due to climate change is substantial. Additionally, socioeconomic growth will further increase flood susceptibilities in the coming years. The increase in population and assets contributes to a considerable increase in global flood losses with the compounded influences of climate change [6]. The Intergovernmental Panel on Climate Change (IPCC) report published in 2022 was related to mitigation of climate change and adopting appropriate measures [7]. Although policy makers have adopted multiple strategies to mitigate the effects of floods, it has still caused severe damage to infrastructures and property. In addition, many cities have constructed or are constructing metro lines [8–10]. However, extreme rainstorms have become frequent due to global climate change, which has increased the instances of flooded metro lines. The Metro system is a critical mode of transport as it maintains the normal operation of

[*] Corresponding author (email: zhenyu.yin@polyu.edu.hk)

urban infrastructures. Immeasurable loss is caused when a metro line is flooded during rainstorms. Therefore, it is crucial to protect urban metro systems from floods.

Prediction of flooding of metro stations under extreme rainstorms is an efficient technique to assess the flood susceptibilities of urban metro lines. The generation of a flood susceptibility map of a metro system can be used to determine metro stations susceptible to floods. Flood susceptibility mapping is useful for flood occurrence prediction in metro stations and supporting flood management decisions for the safe operation of a metro system. Recent flood events in metro stations have led to studies focused on the flood susceptibilities of metro lines. Lyu et al. [11–13] conducted multiple studies on flood risk assessment of metro system using multi criteria decision making (MCDM) such as analytical hierarchy process (AHP) and fuzzy AHP. However, MCDM methods consist of subjective shortcomings due to expert decisions [14]. Moreover, MCDM-based methods cannot predict the risk of flooding in a metro system under unpredictable rainfall scenarios [15–17]. Therefore, an approach was proposed in this study wherein machine learning (ML) methods were incorporated into geographic information system (GIS) techniques to overcome this limitation and predict flood susceptibilities of metro lines under different rainfall scenarios.

GIS is a powerful tool that can be used to integrate historical flood data as it provides a dominant position to establish a database for ML techniques. The combination of ML and GIS techniques significantly improves flood mitigation, planning, and recovery [18–20]. Several studies have focused on flood risk analysis using ML and GIS techniques [21–24]. Flood susceptibility mapping using ML and GIS techniques can provide accurate flood models and visualize the results in a spatial environment [25–29]. Dodangeh et al. [30] proposed an integrative flood risk model using resampling methods integrated with ML models. Mahato et al. [27] developed ensemble-based ML models to predict flood susceptibilities. It was demonstrated that the ensemble flood susceptibility models provided an optimum accuracy of evaluation performance. These previous studies focusing on flood susceptibilities using ML and GIS techniques provide efficient supports for the prediction of flood susceptibilities of underground metro lines. In contrast to flood risk analysis of a particular study area, the prediction of flood susceptibilities of metro lines requires data related to metro stations, particularly during rainstorms.

Generally, a metro station is designed to be flood resistant using the prevention standard with 50 years rainfall return period (noted as 50-Y rainfall intensity). However, climate change has led to frequent extreme rainstorms, which threaten the safety of urban metro systems such as the flooded metro lines in the 7.20 flood in Zhengzhou. Under this circumstance, this study proposed an approach to predict flood susceptibilities of metro lines when encountering future extreme rainstorms. The major novelty of this study is the following: (i) proposing an approach to incorporate ML models into GIS to predict the flood susceptibilities of metro lines with consideration of anthropogenic factors; (ii) proposing a new insight to generate flood samples to establish a database for training and testing when using ML and GIS to predict flood susceptibility; (iii) comparing and selecting the optimum ML model, which is applied to predict flood susceptibilities of metro lines in Zhengzhou. Different rainfall return periods, such as 50, 100, 200, 500, and 1000 years of rainfall (noted as 50-Y, 100-Y, 200-Y, 500-Y, and 1000-Y rainfall intensities) are designed to analyze the flood susceptibility of metro lines. The metro stations under 50-Y rainfall intensity were considered non-flooded. The flood metro stations in 7.20 flood were used to validate the predicted flood states and flood probabilities obtained from ML models. The database for flood prediction was generated in the GIS platform. The major contribution of this study is that the proposed new insight for the prediction of flood susceptibility of urban metro systems using ML incorporated into GIS and the approach was verified reasonable against the records.

2 Study area and databases

2.1 Study area

The Zhengzhou city is located between longitude 113°–114°E, and latitude 34°–35°N in Henan Province, as shown in Figure 1. The altitude is high in the western region and low in the eastern region, and the metro lines are distributed in the urban center of Zhengzhou. The topographic features facilitate rainwater collection in the urban area from east to west during rainstorms. In 2021, the 7.20 flood caused 398 deaths in Zhengzhou city [31]. Five metro lines with 18 metro stations were flooded by rainwater in 7.20 flood. The Haitansi and Shakoulu stations were flooded and the metro trains were suspended, which trapped passengers and caused 14 deaths [31]. Figure 1 presents the flooded stations in 7.20 flood.

2.2 Rainfall in 7.20 flood

According to China Meteorological Data Network [32], the rainfall that occurred from July 19 to July 21, 2022, in Zhengzhou was considered a rainstorm (a rainstorm is defined as the rainfall > 50 mm in 24 h), wherein the rainfall (up to 200 mm) on July 20 continued for 2 h (Figure 2(a)). The metro stations were flooded within these 2 h. The rainfall process and cumulative precipitation of flood events under 50-Y, 100-Y, 200-Y, 500-Y, and 1000-Y rainfall intensities were determined by the rainstorm intensity formula

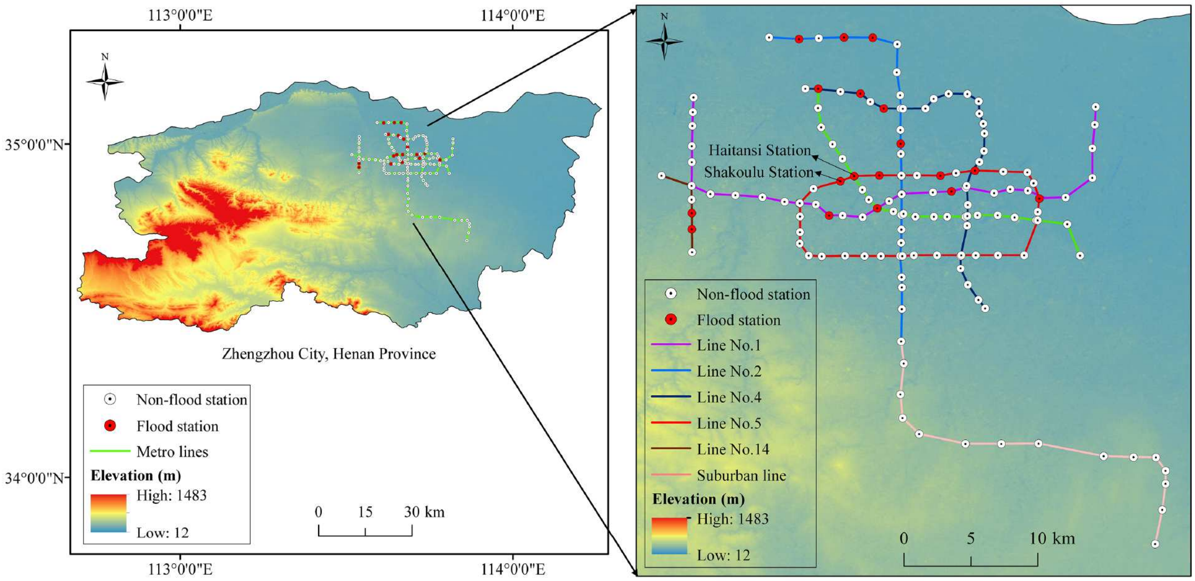

Figure 1 (Color online) Distribution of metro lines in Zhengzhou city.

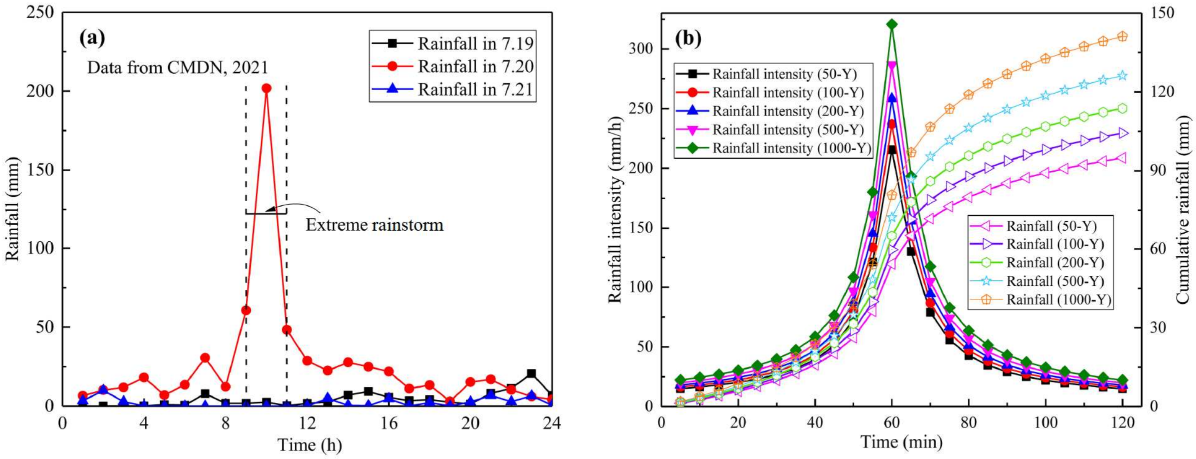

Figure 2 (Color online) Recorded and envisaged rainfall scenarios. (a) Recorded rainfall in 7.20 flood; (b) the envisaged five scenarios.

in Zhengzhou [33], as described in the following formula:

$$Q=\frac{2631.92\times(1+0.751\times\lg P)}{(t+14.2)^{0.779}},\tag{1}$$

where Q is the designed rainfall intensity (L/(hm^2 s)); P is the designed rainfall return period (year); t is the time series (min). Figure 2(b) shows the rainfall process and cumulative precipitation in different rainfall scenarios. The cumulative precipitations of rainfall intensities under 50-Y, 100-Y, 200-Y, 500-Y, and 1000-Y were determined in the range of 90–150 mm of rainfalls in a 2-h period, which were less than that of the 2-h extreme rainstorm on July 20. Therefore, the extreme rainstorm in 7.20 flood exceeded the rainstorm of 1000-Y rainfall intensity, which was the main reason that directly caused the floods of metro lines. Based on the flood metro stations in 7.20 flood, the flood susceptibilities of metro lines under different scenarios such as 50-Y, 100-Y, 200-Y, 500-Y, and 1000-Y rainfall intensities were predicted.

In addition, the spatial distribution of rainfall during the 7.20 flood including the rainfall on July 19 (Figure 3(a)), July 20 (Figure 3(b)), July 21 (Figure 3(c)), and the rainfall on July 20 within 2 h was analyzed. The high rainfalls that occurred on July 19 and July 21 were concentrated in the south-western region with a higher elevation, while the rainfall on 20 July was concentrated in the urban center of Zhengzhou with a lower elevation and dense distribution of metro lines. The rainwater has a tendency to flow from a high-elevation location to a low-elevation location. Topographies with a higher western region and lower urban center

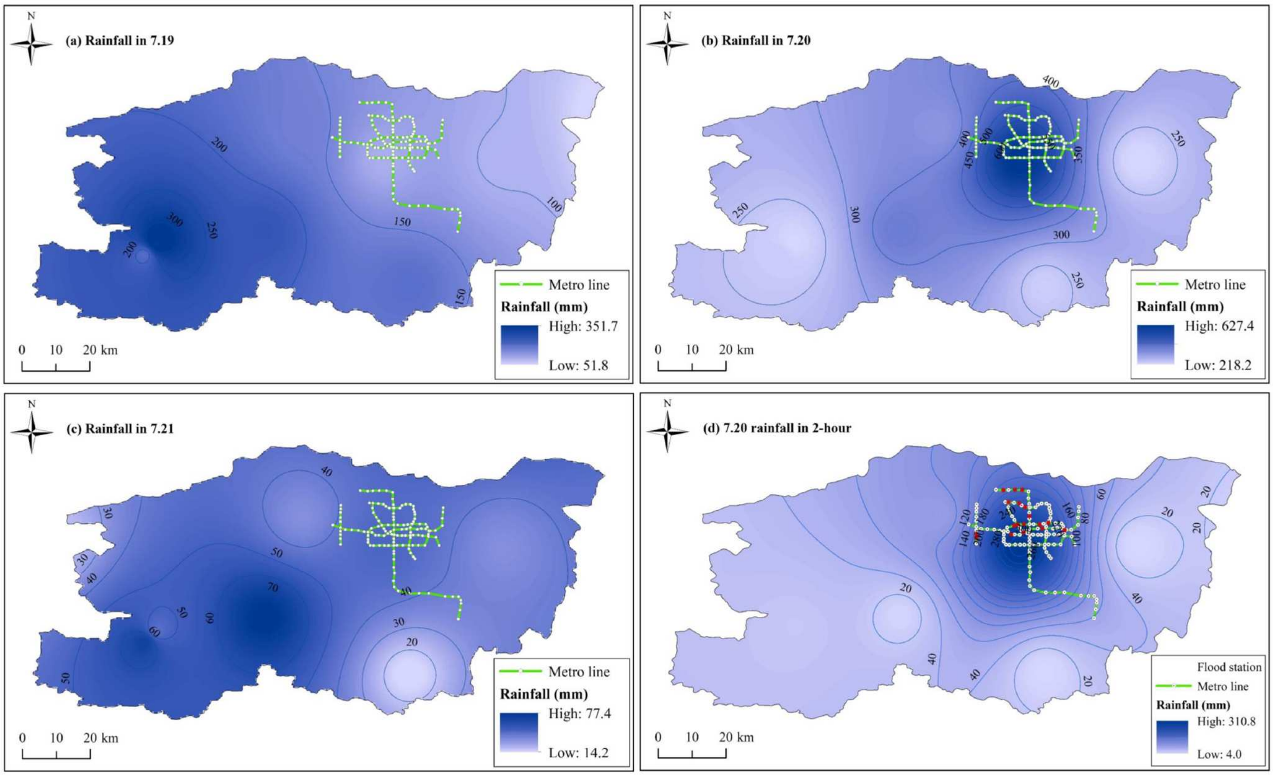

Figure 3 (Color online) Spatial distribution of rainfall in 7.20 flood in Zhengzhou. (a) Rainfall on July 19; (b) rainfall on July 20; (c) rainfall on July 21; (d) rainfall on July 20 within 2 h.

are highly susceptible to flooded metro lines during extreme rainstorms.

2.3 Controlling factors

2.3.1 Topography

The topographic factors include elevation (Figure 4(a)), slope (Figure 4(b)), aspect (Figure 4(c)), hill shade (Figure 4(d)), plane curvature (Figure 4(e)), profile curvature (Figure 4(f)), TWI (Figure 4(g)), NDVI (Figure 4(h)), river density (Figure 4(i)), and river proximity (Figure 4(j)). The data sources of topographical factors are obtained from a digital elevation model (DEM) with 30 m resolution from Geospatial Data Cloud. Based on the DEM data, the elevation, slope, aspect, hill shade, profile, and plane curvatures can be extracted using surface analysis in GIS. The river density and proximity are obtained according to the river network, which are extracted from DEM using hydrological analysis in GIS. The topographic wetness index (TWI) is used to measure water aggregation and surface runoff in a water basin. The TWI can be calculated using eq. (2) in GIS,

$$\mathrm{TWI} = \ln[a / \tan(b\times\pi) / 180], \tag{2}$$

where a refers to the upstream catchment area and b refers to the slope angle in the study area. The normalized difference vegetation index (NDVI) is one of the important parameters that reflect the vegetation information. The NDVI is calculated using eq. (3),

$$\mathrm{NDVI} = \frac{\mathrm{NIR}-R}{\mathrm{NIR}+R}, \tag{3}$$

where NIR refers to the near infrared band, and R refers to the red band of images, $-1 \leqslant \mathrm{NDVI} \leqslant 1$. A detailed description of the data sources is listed in Table 1.

2.3.2 Anthropogenic factors

The anthropogenic factors include land use (Figure 5(a)), population density (Figure 5(b)), gross domestic product (GDP) (Figure 5(c)), and distance to building (Figure 5(d)). The data of land use, population, and GDP were obtained from Resource and Environment Science and Data Center (RESDC) with 1 km raster. The land use types were classified into unused land, forest land, grass land, water body, cultivated land, and residential land with the vulnerability to flood from levels 1 to 6. For population and GDP, flood susceptibility increased with the increase of the population and GDP. To reflect the influence of buildings on flood susceptibilities, the proximities of buildings were produced in GIS using buffer analysis with distances of 200, 400, 600, 800, and 1000 m. The flood risk and the damage caused by flood increase as the distances to buildings decrease. A detailed description of the sources for anthropogenic factors is listed in Table 1.

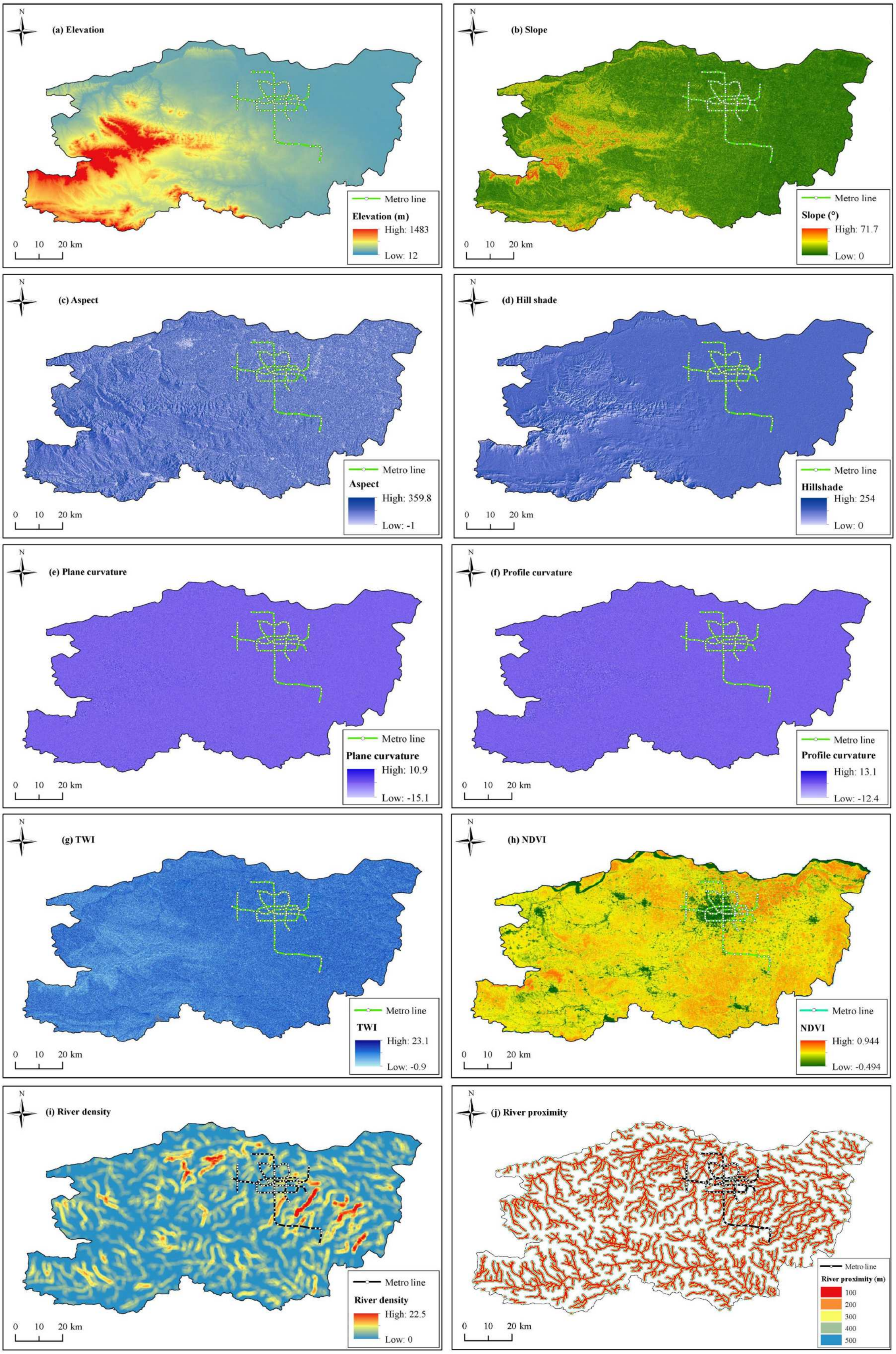

Figure 4 (Color online) Spatial distribution of topographic factors. (a) Elevation; (b) slope; (c) aspect; (d) hill shade; (e) plane curvature; (f) profile curvature; (g) TWI; (h) NDVI; (i) river density; (j) river proximity.

Table 1 Description of the sources for the considered databases[a)]

Factor	Description	Type	Statistics				Data source/ resolution/type
			Min	Max	Mean	STD	
Rainfall720_2h (mm)	Rainfall on July 20 within 2 h	N, S	56.39	310.70	210.13	60.84	Rain gauges and CMDN, 2021/10 m/raster
Elevation (m)	Digital elevation of terrain surface	N, S	63	197	104.33	21.62	DEM/30 m/raster
Slope (°)	Angle of slope inclination	N, S	0	38.57	5.85	3.64	
Aspect	Compass direction of slope exposure	N, S	−1	359.37	175.45	106.28	
Hill shade	Reflect the special terrain characteristics, such as mountain, valley, and canyon	N, S	28	237	178.75	15.39	
Plane curvature	Curvature perpendicular to the slope, indicating concave or convex surfaces	N, S	−2.82	2.76	0	0.56	
Profile curvature	Curvature parallel to the slope, indicating concave or convex surfaces	N, S	−3.29	2.95	0.02	0.62	
TWI	Topographic wetness index, measuring water aggregation and surface runoff	N, S	0	20.08	8.46	3.35	
NDVI	Normalized difference vegetation index reflecting the vegetation information	N, S	0.05	0.87	0.46	0.23	
River density	Distribution of river system in 1 km^2	N, S	0	12.74	2.86	2.47	
River proximity (m)	Distance to the river system	N, S	0	500	260.51	240.54	
Land use	Land cover in the study area	C, S	–	–	–	–	RESDC/1 km/raster
Population density (p/km^2)	Population information in 1 km^2	N, S	40.18	613.48	343.07	157.74	
GDP	Distribution of gross domestic product	N, S	41.16	44558	25330	16010	
Distance to building (m)	Distance to the distributed buildings	N, S	0	1000	365.42	242.79	RESDC/–/point+Polygon

a) N, C, and S stand for numerical, categorical, and static variables, respectively. CMDN stands for China Meteorological Data Network (CMDN), and RESDC stands for Resource and Environment Science and Data Center.

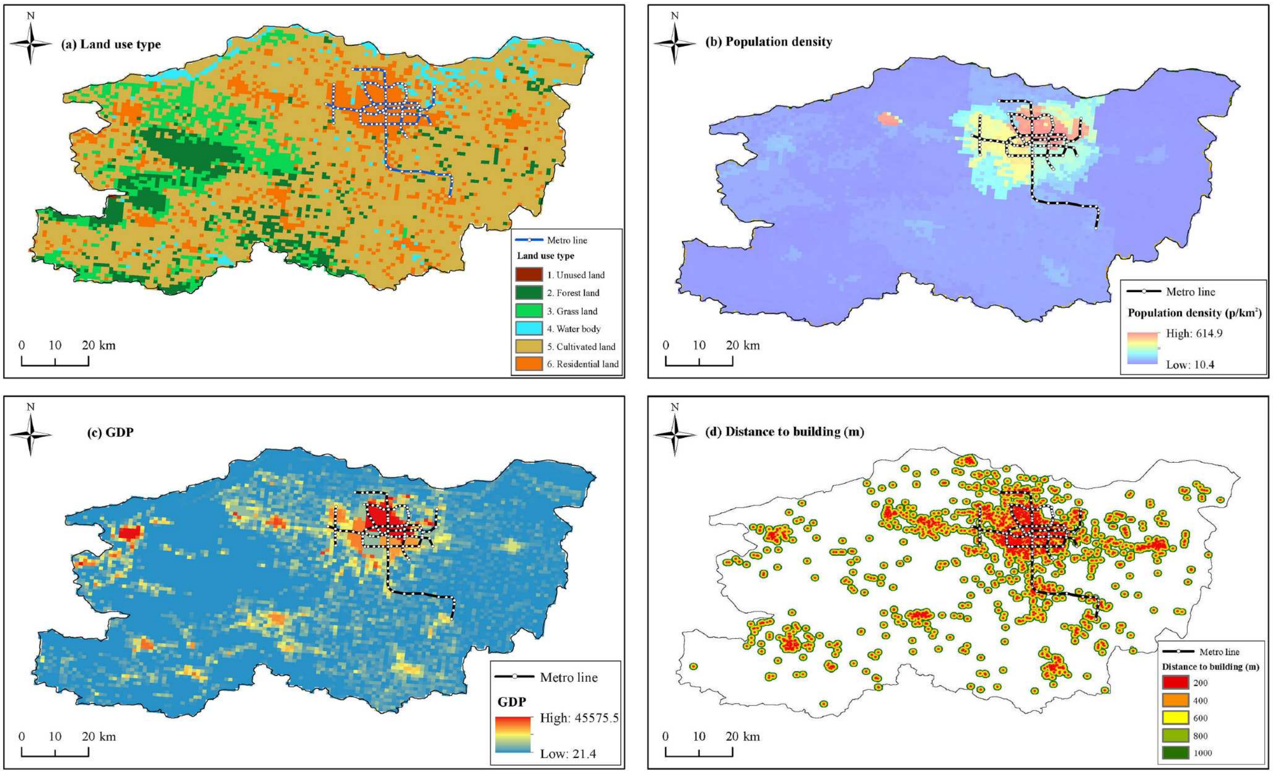

Figure 5 (Color online) Spatial distribution of anthropogenic factors. (a) Land use; (b) population density; (c) GDP; (d) distance to building.

2.4 Feature selection

The features of controlling factors have significant influences on flood susceptibility. To avoid the multicollinearity of the selected factors, the multicollinearity test was conducted by computing the Pearson's correlation coefficients. The Pearson's coefficients represent the correlation relationships among the selected controlling factors, which reflect the features of the database. Figure 6 shows the pairwise correlations among flood controlling factors. The value in the matrix refers to the correlation degrees among controlling factors. The values close to 0 indicate that the two factors are not correlated, while the values close to 1 mean that there is a strong positive correlation between the two controlling factors. According to Tehrany et al. [34], the Pearson's coefficients above than 0.7 may lead to a multicollinearity issue. As shown in Figure 6, the Pearson's coefficients of all controlling factors are less than 0.7, which indicates that these determined 15 controlling factors are independent each other with less values of connection degree.

3 Methodology

3.1 Framework of ML incorporated into GIS

In this study, the flood susceptibility is evaluated by a ratio from the ML algorithm, which represents the probability of flood. Figure 7 shows the framework of the prediction for flood susceptibility of metro systems using ML models. The first part is the integration of the database. The rainfall on July 20 within 2 h was extracted to integrate the topographic factors (including elevation, slope, aspect, hill shade, plane curvature, profile curvature, TWI, NDVI, river density, and river proximity) and anthropogenic factors (including land used, population density, population density, GDP, and distance to building). These data were extracted to generate the non-flood and flood samples. Subsequently, the flood states and probabilities of metro stations were determined using gradient boosting decision tree (GBDT), random forest (RF), adaptive boosting (AB), K-nearest neighbors (KNN), logistic regression (LR), and support vector machine (SVC) models. The performance of these six ML models was evaluated using accuracy, precision, recall, F1-score, receiver operator

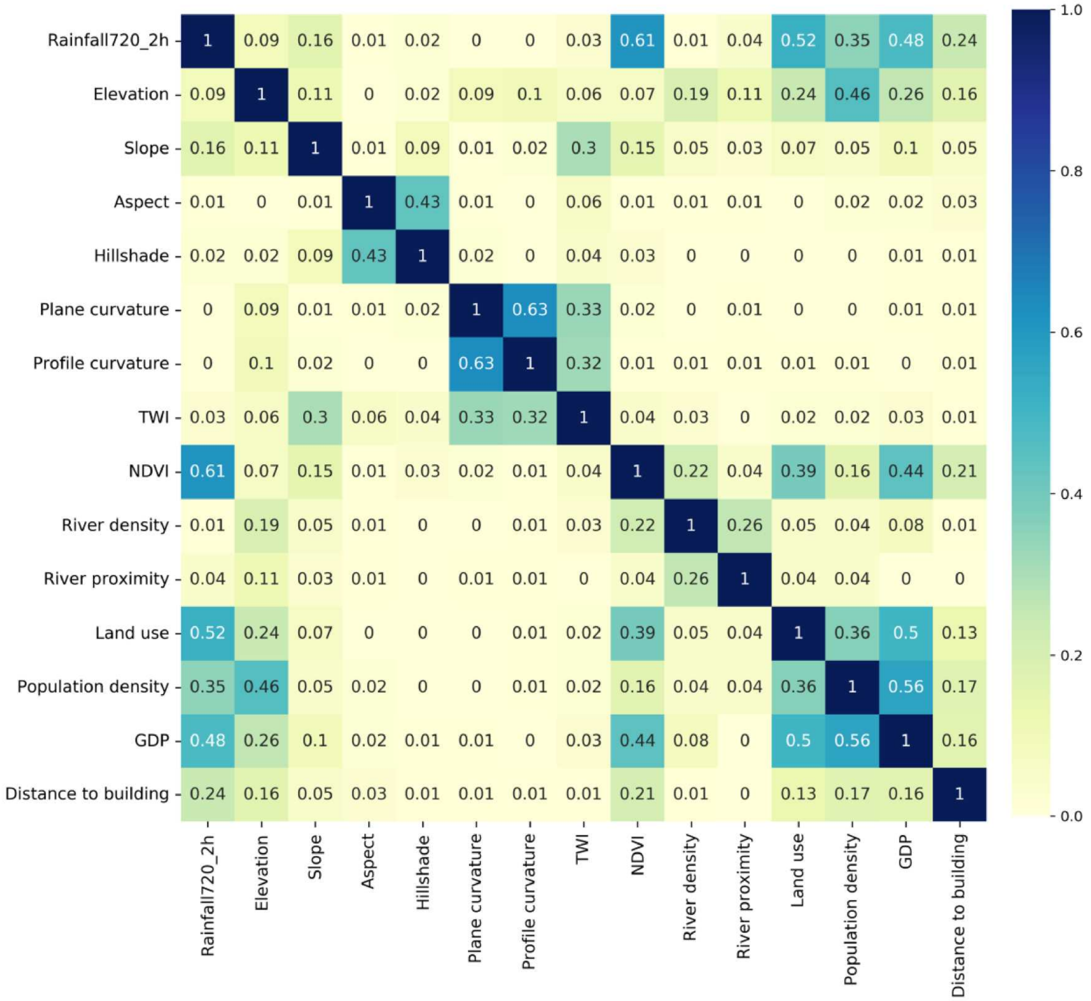

Figure 6 (Color online) Pairwise correlations among flood controlling factors.

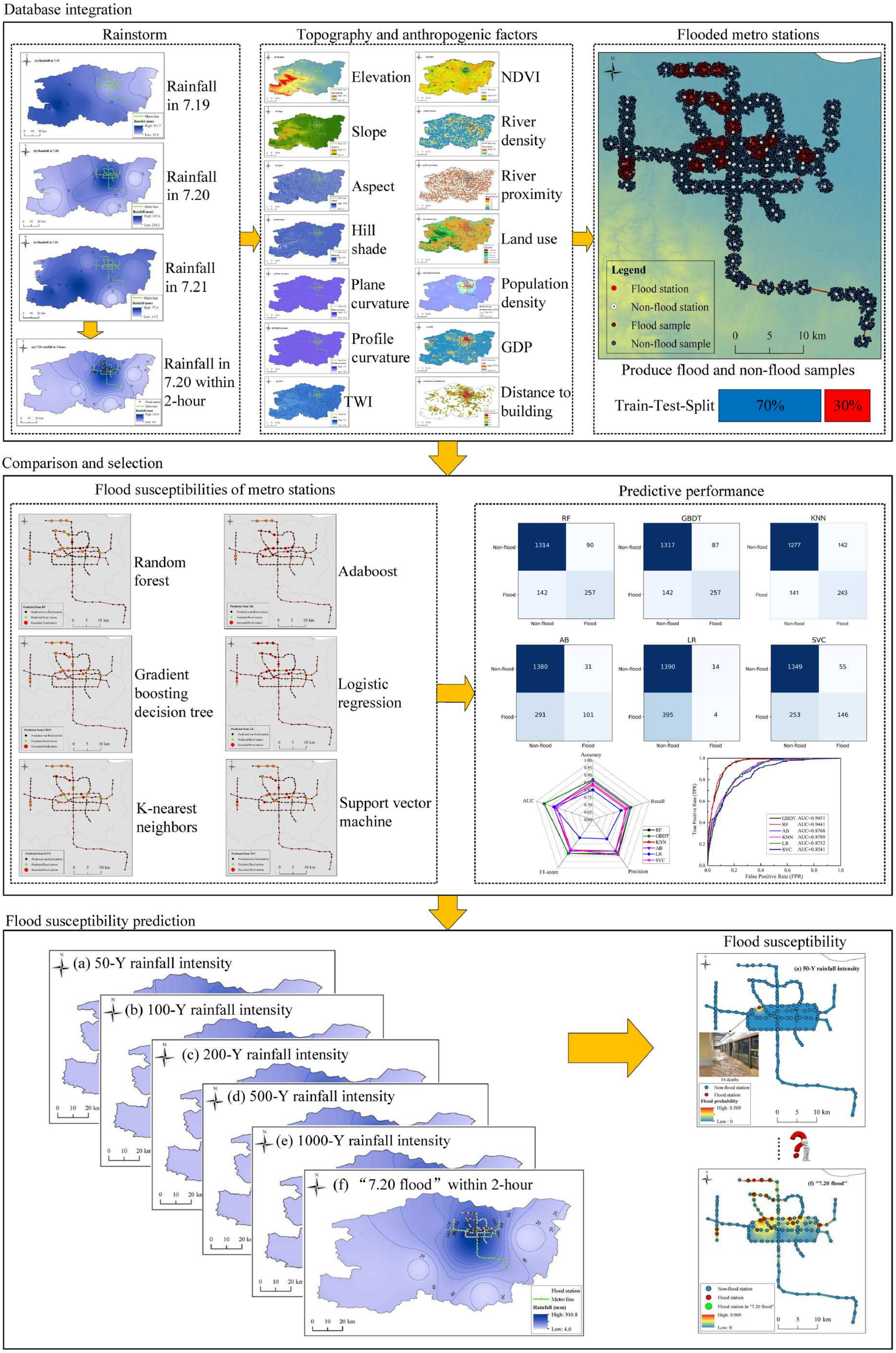

Figure 7 (Color online) Framework of flood susceptibilities prediction using ML.

characteristic (ROC) curve, and area under ROC curve (AUC). Based on the comparison, the model with the optimum performance was selected to predict flood susceptibilities of metro stations under different scenarios including rainfall intensities of 50-Y, 100-Y, 200-Y, 500-Y, and 1000-Y and rainfall during the 7.20 flood in 2 h. The third part is the prediction of flood susceptibilities under different rainfall scenarios using the selected optimum model. The metro stations under the rainfall intensity of 50-Y were designed as non-flood, while 18 stations under the rainfall in “7.20 flood” within 2 h were flooded. How are the flood susceptibilities of metro lines under other rainfall scenarios? The major innovation of this framework is the generation of the database, which is formed by the flood and non-flood samples. The record data only includes 18 flooded metro stations, which represent 18 flood samples. But the 18 flood points are not enough to establish the database for ML training and testing. To overcome this difficulty, the flood samples were generated around each flooded station within 1 km to increase the number of flood samples. The generated flood samples are different in designed rainfall intensity to reflect the reality in different rainstorm scenarios. Section 3.2 gives a detailed description of the generated database.

3.2 Data integration with ML and GIS

During the application of ML and GIS to predict flood susceptibility, it is a critical task to establish a database, since the rare cases of subway flooding. To overcome the difficulty of scarcity of flood data, this study proposed a new way to produce flood samples according to the recorded flood data. Figure 8 shows the samples for training and testing. The flood samples were generated around each flooded station within 1 km to predict flood susceptibility of metro stations under different rainfall scenarios, and the non-flooded samples were generated around each non-flood metro station within 1 km. To reflect the reality of different subway flooding caused by different rainfall intensities, the flood samples were produced differently in different flood scenarios (rainfall intensities of 50-Y, 100-Y, 200-Y, 500-Y, and 1000-Y). The detailed flood samples in the database are introduced as follows.

3.3 Establishment of database

This study considered 15 controlling factors related to flood susceptibility, which included topographical factors (such as elevation, slope, aspect, hill shade, plane curvature, profile curvature, TWI, NDVI, river density, and river proximity), anthropogenic factors (such as land use, population density, GDP, and distance to buildings), and rainfall factors (such as rainfall in 7.20 flood within 2 h, 50-Y, 100-Y, 200-Y, 500-Y, and 1000-Y rainfall intensities). Specifically, the 50-Y rainfall intensity was designed as the flood prevention standard, which was considered non-flooded under this scenario. However, one flood sample should be presented in the datasets during the ML training. Therefore, the Haitansi and Shakoulu stations were designed (wherein 14 deaths occurred during the 7.20 flood) as two flood samples in the datasets under the rainfall intensity of 50-Y. The flood samples (178 flood samples) were generated under a rainfall intensity of 100-Y around these two flood metro stations within 1 km. The flood samples were generated around the flooded stations within 1 km under rainfall scenarios of 200-Y (535 flood samples), 500-Y (890 flood samples), 1000-Y (1146 flood samples), and 7.20 flood (1380 flood samples). Finally, 6010 samples were produced in the GIS database. These samples were adopted to extract the information of rainstorms, and topographic and anthropogenic factors. The flooded metro stations due to the rainfall on July 20 within 2 h were analyzed based on this investigation. Therefore, the rainfall on July 20 in 2 h (denoted as rainfall 720_2h) integrated topographic and anthropogenic factors was used to train and test ML models with a training and testing ratio of 0.7.

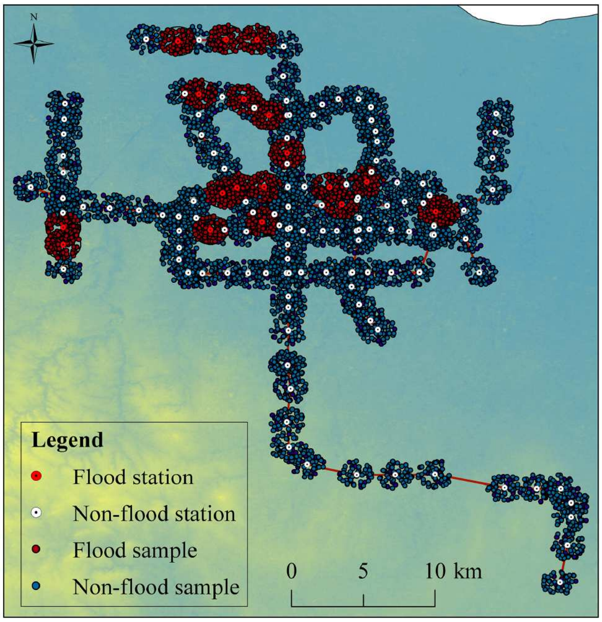

Figure 8 (Color online) Samples for training and testing.

3.4 ML models

In this study, six ML models, namely GBDT, RF, AB, KNN, LR, and SVC were adopted and compared to predict the flood probabilities of metro stations in 7.20 flood. For these six ML models, the GBDT, RF, and AB are typical tree models, and the KNN, LR, and SVC are typical linear models. To compare the performance between tree and liner models on flood susceptibility prediction, these six ML models are selected to predict flood susceptibilities. According to the predicted flood probabilities, the flood susceptibilities of metro lines were mapped by using GIS tools. These analyses were performed using the Python programming environment and machine learning package Scikit-learn.

(1) Gradient boosting decision tree

GBDT was proposed by Breiman [35] and further developed by Friedman [36]. GBDT is an ensemble-based algorithm, which is widely used for classification and regression because it has optimum interpretability and robustness for the discovery of high order relationships among features. The loss function used in boosting trees for a classification problem is an exponential function, and a regression problem is used as a mean squared error. This algorithm does not require pre-processing for data normalization [37,38].

(2) Random forest

RF is an ensemble-based ML technique used for classification and regression, which operates by constructing a multitude of decision trees. The output of the RF for a classification problem is the class selected by most trees. The mean or average prediction of the individual trees for a regression problem is returned [39]. Generally, RF performs better than the decision tree, but its accuracy is lower than that of the gradient boosting tree. However, data features can affect its performance.

(3) Adaptive boosting

AB is a statistical classification metaalgorithm used in conjunction with multiple learning algorithms to improve the performance. Generally, AB is presented for binary classification, even though it can be generalized to multiple classes or bounded intervals on the real line [40]. The main difference between GBDT and AB is that AB modifies the sample weights in each iteration to ensure that the subsequent tree model considers the misclassified samples, whereas GBDT is the latter tree model used to directly fit the residuals.

(4) K-nearest neighbours

KNN is a non-parametric supervised learning algorithm used for classification and regression. If most of the K closest neighbors in the feature space belong to a certain class, the sample belongs to this class. The input in KNN models consists of K closest training examples in a data set, whereas the output depends upon whether KNN is used for classification or regression. KNN determines the class for a classification problem according to the class of the nearest one or several samples. The selection of the K value is the only hyperparameter in the KNN algorithm that has an intuitive and important effect on the prediction results. During application, cross validation is usually used to select the optimal K value [41].

(5) Logistic regression

LR is a statistical model that is generally used for binary classification, which is labeled 0 and 1. The corresponding probability of the value labeled 1 can vary between 0 and 1. Binary classification models are widely used in statistics to evaluate the probability of a certain class or event (e.g., floods and landslides). For a binary classification problem, LR assumed a straight line that completes the linear separation of the sample. LR fits the decision making boundary (including linear and polynomial) and establishes the probability connection between this boundary and classification, thereby obtaining the probability of a binary classification problem.

(6) Support vector classification

SVMs include support vector regression (SVR) and SVC to perform regression and classification problems. SVC is a robust prediction model based on statistical learning frameworks. The basic model of SVC is a linear binary classifier. In addition to conducting linear classification, SVC can efficiently perform non-linear classification using kernel tricks to process the inputs into high dimensional features.

4 Results

4.1 Model evaluation and selection

4.1.1 Confusion matrix

The flood susceptibility of a metro station was considered a binary classification problem, wherein each sample had a positive (flood) or negative (non-flood) prediction. The predicted flood metro stations were compared with that of the recorded flood metro stations to validate the efficiencies of the ML models. The number of actual floods predicted as floods was defined as the true positive (*tp*), the number of actual non-floods predicted as non-floods corresponded to the true negative (*tn*), the number of actual floods predicted as non-floods was defined as false positive (*fp*), and the number of actual non-floods predicted as floods was the false negative (*fn*). The confusion matrices were used to compare the performance of the six ML models to determine the accuracy of the flooded and non-flooded samples in the test subset (see Figure 9). A large value on the diagonal indicated the optimum performance of the model. The GBDT accurately classified 1317 non-flooded and 257 flooded samples, whereas RF accurately classified 1314 non-flooded and 257 flooded samples. The results demonstrated that GBDT and RF accurately identified flooded and non-flooded samples compared with that of the other algorithms.

4.1.2 Performance evaluation

In addition to confusion matrices, the performance of the six ML models was evaluated using the indicators such as ROC curve, AUC, accuracy, recall, precision, and F1-score [42]. The accuracy is used to measure the efficiency of a machine learning model, which is defined using

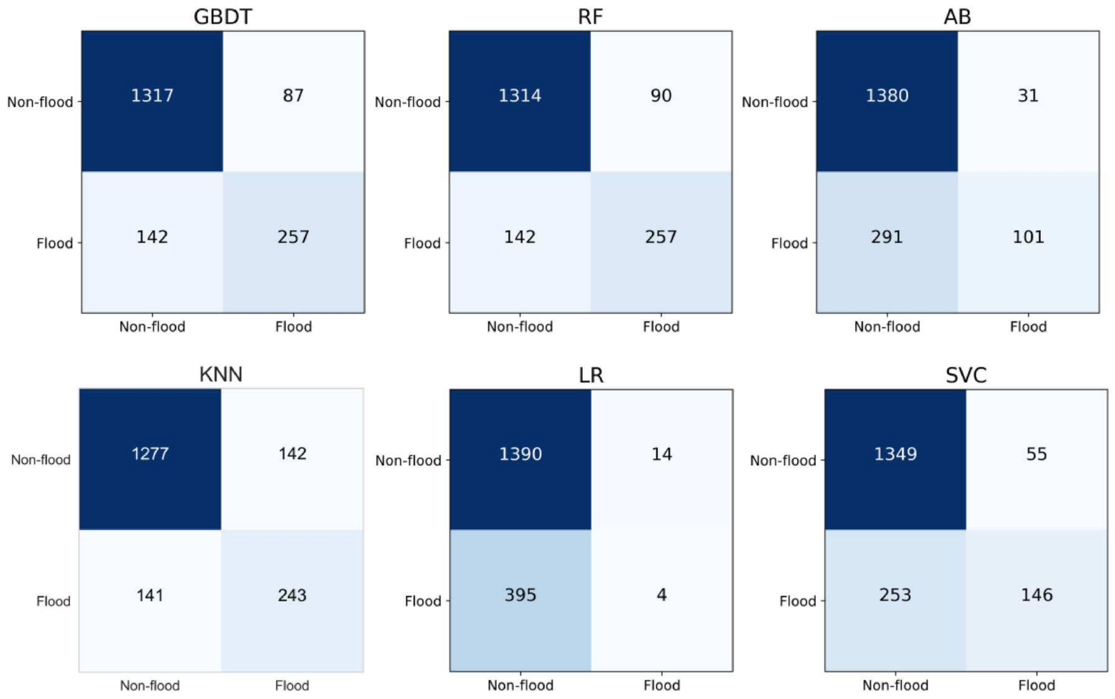

Figure 9 (Color online) Confusion matrices of six ML models.

$$\text{Accuracy} = \frac{tp+tn}{tp+fp+tn+fn}. \tag{4}$$

The precision is used to measure the class consistency between the input data labels and the positive labels produced by machined learning models,

$$\text{Precision} = \frac{tp}{tp+fp}. \tag{5}$$

The recall is used to measure the efficiency of a machine learning model to find positive labels,

$$\text{Recall} = \frac{tp}{tp+fn}. \tag{6}$$

The F1-score is used to express the relationship between positive labels in the input data and those assigned by a machine learning model,

$$\text{F1-score} = \frac{2\times\text{Recall}\times\text{Precision}}{\text{Recall}+\text{Precision}}. \tag{7}$$

The AUC is a comprehensive measurement to reflect the efficiency of a machine learning model to avoid false classification,

$$\text{AUC} = \frac{1}{2}\left(\frac{tp}{tp+fn}+\frac{tn}{tn+fp}\right). \tag{8}$$

Figure 10 shows the ROC curves and performance indicators of the six ML models. The AUC score of RF and GBDT models for the ROC curve was approximately 0.94, which was superior to that of AB, KNN, LR, and SVC models. Moreover, the performance indicators of accuracy, recall, and F1-score of RF and GBDT were greater than those of other algorithms. The comparison indicated RF and GBDT models had a better performance than that of the other ML models.

4.2 Comparison and validation

Figure 11 depicts a comparison between recorded and predicted flood metro stations obtained using the six ML models. It was observed that 18 metro stations were flooded in 7.20 flood. The GBDT algorithm accurately predicted 16 flood metro stations, which was followed by RF (15 flooded stations), KNN (13 flooded stations), SVC (10 flooded stations), AB (6 flooded stations), and LR (2 flooded stations). A comparison between the predicted and recorded flood metro stations exhibited that the GBDT more accurately predicted flooded stations than other models. Therefore, the GBDT model was selected to predict the flood susceptibilities of metro lines under different rainfall scenarios.

4.3 Predicted flood susceptibilities under different rainfall scenarios

During the application of ML to predict flood susceptibility, the ML algorithms can obtain deterministic results. For instance, the GBDT algorithm can obtain the deterministic results labeled as flood and non-flood. In addition, a ratio ranging from 0 to 1 of each label of flood and non-flood can also be output. In this study, the ratios of the labeled flood are

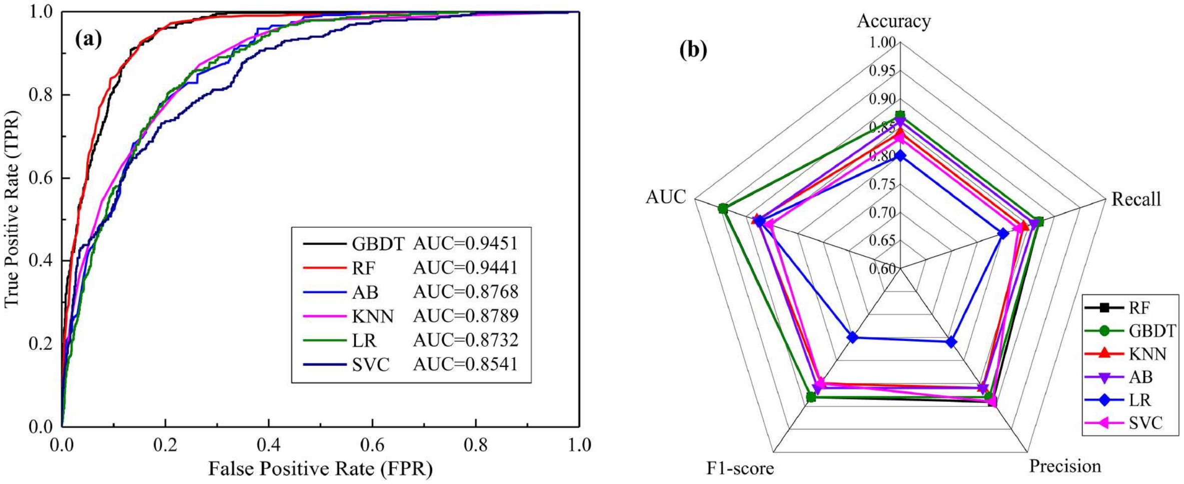

Figure 10 (Color online) Evaluation of ML models with different performance indicators. (a) ROC curves; (b) performance indicators.

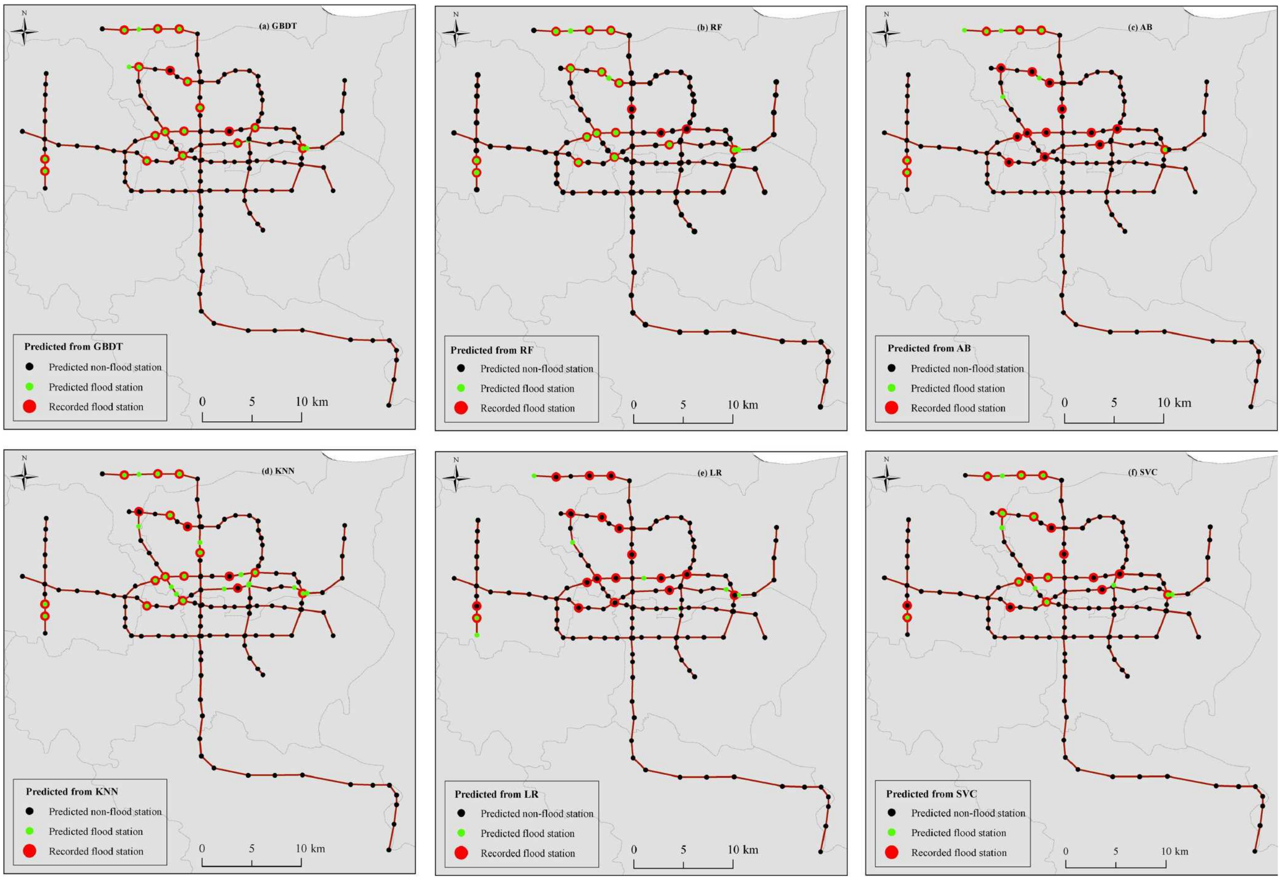

Figure 11 (Color online) Comparison between recorded and predicted flooded stations. (a) GBDT; (b) RF; (c) AB; (d) KNN; (e) LR; (f) SVC.

adopted to represent the flood probabilities. Therefore, the GBDT model was selected to predict the flood probabilities of metro stations under different rainfall intensities of 50-Y (Figure 12(a)), 100-Y (Figure 12(b)), 200-Y (Figure 12(c)), 500-Y (Figure 12(d)), 1000-Y (Figure 12(e)), and rainfall within 2 h in 7.20 flood (Figure 12(f)). Since the standards of flood prevention for metro stations are designed as a 50-Y rainfall return period, thus the metro stations were considered non-flooded under the 50-Y rainfall intensity. The flooded stations in the 7.20 flood were used to validate the predicted flood state of each station. The flood states of metro stations under other rainfall scenarios were analyzed

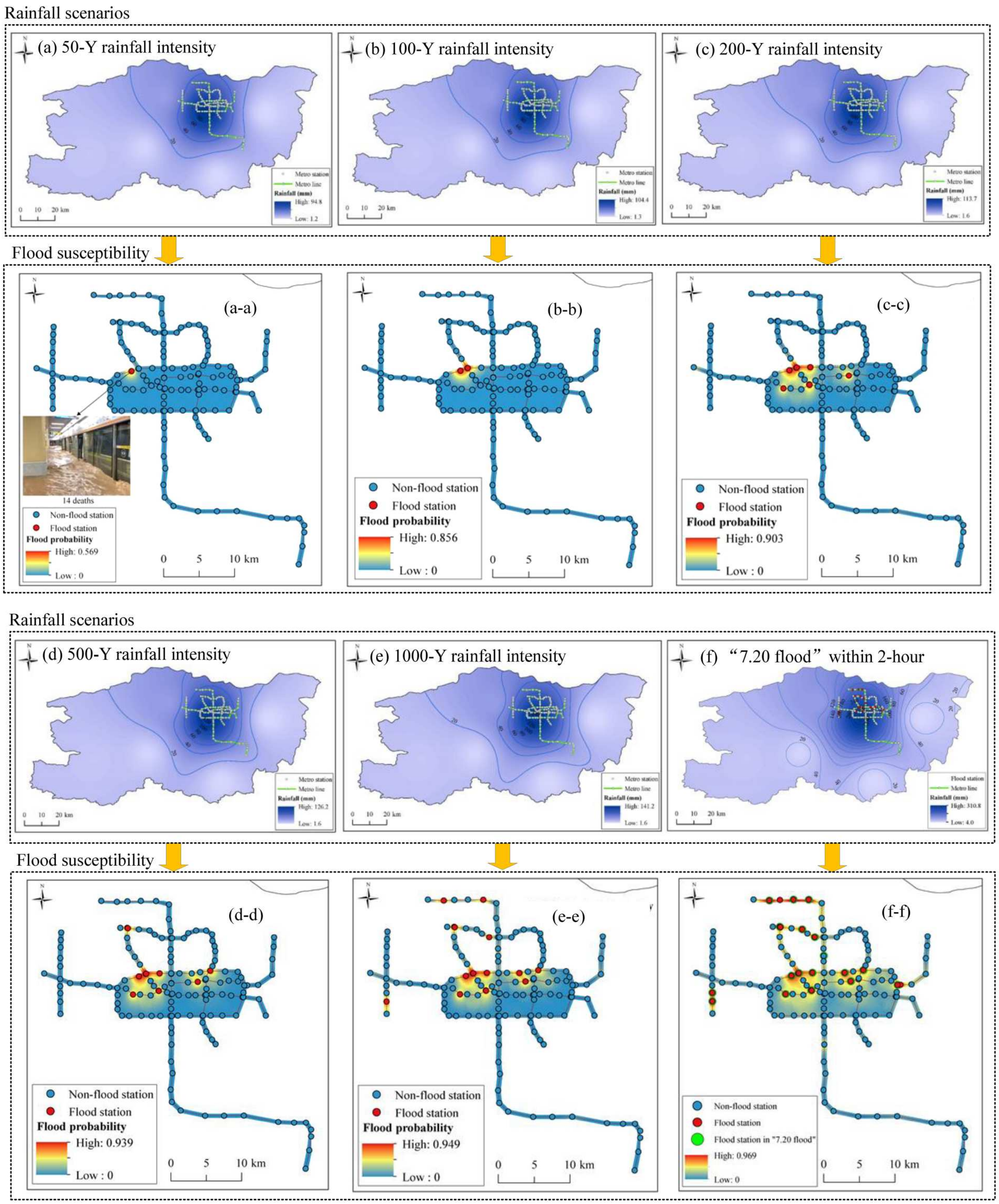

Figure 12 (Color online) Spatial distribution of rainfall scenarios and predicted flood susceptibilities of metro lines under different rainfall scenarios. (a), (a-a) 50-Y; (b), (b-b) 100-Y; (c), (c-c) 200-Y; (d), (d-d) 500-Y; (e), (e-e) 1000-Y; (f), (f-f) rainfall in 7.20 flood within 2 h.

based on the non-flood design for the 50-Y rainfall intensity and flooded stations during the 7.20 flood. According to the predicted flood probability of each station, the flooding susceptibilities along metro lines within 200 m were obtained by interpolation analysis in GIS.

Figure 12(a-a)–(f-f) show the flood susceptibilities of metro lines under the corresponding rainfall scenarios. The results demonstrated that the Shakoulu station was predicted to flood with a probability of 0.569 under 50-Y rainfall intensity. In the case of 100-Y rainfall scenario, the Shakoulu and Haitansi stations were predicted to flood with a probability of 0.856. It was observed that 6, 8, and 13 stations were flooded with an increase in the flooding probabilities from 0.903 to 0.949 under rainfall intensities of 200-Y, 500-Y, and 1000-Y, respectively. The number of flood metro stations and their flooding probabilities increased with an increase in the rainfall intensity. It was observed that 17 flood metro stations were accurately predicted for the rainfall in 7.20 flood within 2 h. One metro station was flooded. However, it was predicted as non-flooded. Conversely, one station was non-flooded. However, it was predicted as flooded. This was due to the uncertainty of the ML and the error was acceptable.

Figure 13 shows the flood probabilities of 136 metro stations under different rainfall scenarios in Zhengzhou. A metro station was prone to flooding when the flooding probability was greater than 0.5. Conversely, a flooding probability of less than 0.5 exhibited a non-flooded metro station. As shown in Figure 13, the majority of the flooded stations were observed in metro lines No. 1, 2, 4, 5, and 14 when the rainfall intensity exceeded 500-Y, which corresponded to flooded stations during the 7.20 flood. The flood probability of metro stations increased with an increase in the rainfall intensity. The flood probabilities under 7.20 flood were the largest followed by flooding probabilities under rainfall intensities of 1000-Y and 500-Y.

5 Discussion

5.1 Contribution of controlling factors to flood susceptibilities

Figure 14 shows the contribution of the considered 15 controlling factors on flood susceptibilities of metro lines. As shown in Figure 14, population density had the largest contribution with a fluctuation of 0.035 on flood susceptibilities, which was followed by the rainfall on July 20 within 2 h (rainfall720_2h). Additionally, river density significantly

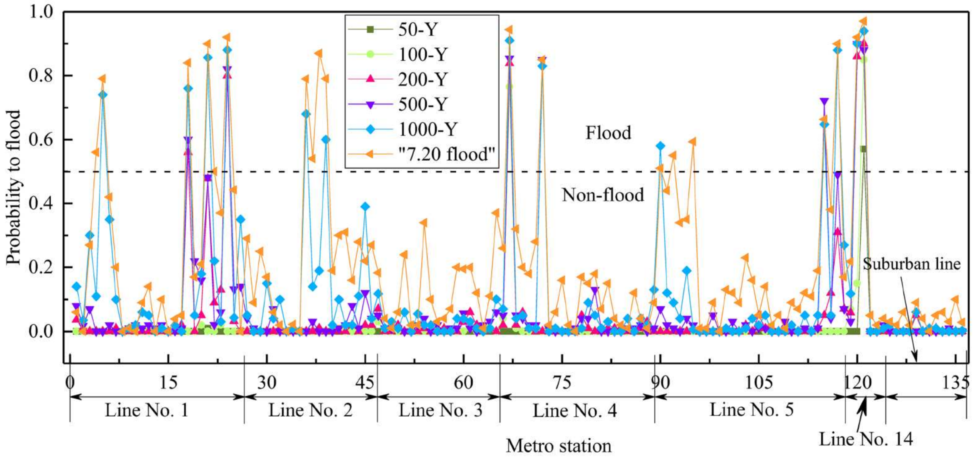

Figure 13 (Color online) Flood probabilities of metro stations under different rainfall scenarios.

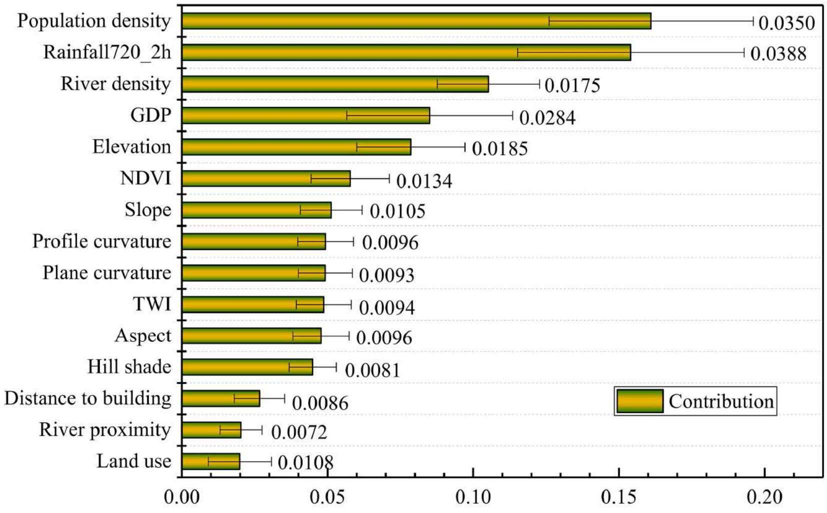

Figure 14 (Color online) Contribution of controlling factors on flood susceptibilities.

affected the flood susceptibilities of metro lines. The rising river level during rainstorms can increase the flood susceptibility. In addition, topographic factors such as elevation, NDVI, slope, profile curvature, plane curvature, TWI, aspect, and hill shade significantly affected the flood susceptibility. The specific topography characterized by the higher altitude in the western region and lower altitude in the urban center with dense distribution of metro lines resulted in the concentration of rainwater during rainstorms. Therefore, special topographical characteristics with high in the western and low in the eastern increased the flood susceptibility of metro lines in Zhengzhou.

5.2 GBDT performance and interpretation

The performance of the GBDT model under different rainfall intensities was evaluated using ROC curves (Figure 15(a)). As shown in Figure 15(a), the AUC values of the GBDT model under different rainfall scenarios are excess than 0.94, which result indicates a better performance of the GBDT model in predicting flood susceptibilities of metro lines. To investigate the effect of controlling factors on prediction outputs from the GBDT model, the Shapley additive explanation (SHAP) summary plot was constructed. Figure 15(b) depicts the SHAP summary plot obtained using the GBDT model. The *y*-axis orders the features according to their importance with a continuous color scale from low to high. Each point denotes a SHAP value and predicted output. The population density, rainfall720_2h, and GDP were the factors that led to SHAP values greater than 1, which indicated these factors considerably affected the predicted output. A high rainfall significantly increased the flood risk, whereas low rainfall reduced the flood risk. A high population density and GDP in a region generally increase the flood susceptibility.

5.3 Limitations

The integration of ML and GIS techniques can effectively predict flood susceptibility of metro lines, which provides suggestive supports for flood prevention and mitigation for urban metro systems. The major obstacle is the collection of critical data related to flood risk. The underground drainage system is difficult to acquire compared with that of meteorological, topographic, and anthropogenic factors owing to data limitation. The present drainage system has a low design standard against rainstorms, which is inconsistent with that of the supposed rainfall scenarios in this study. The underground drainage system was not considered a controlling factor. Moreover, the flood susceptibilities of metro lines were affected by multiple factors. The 15 controlling factors considered in this study did not comprehensively reflect the effects of factors related to flood susceptibility of metro lines, which led to the uncertainties of predicted results.

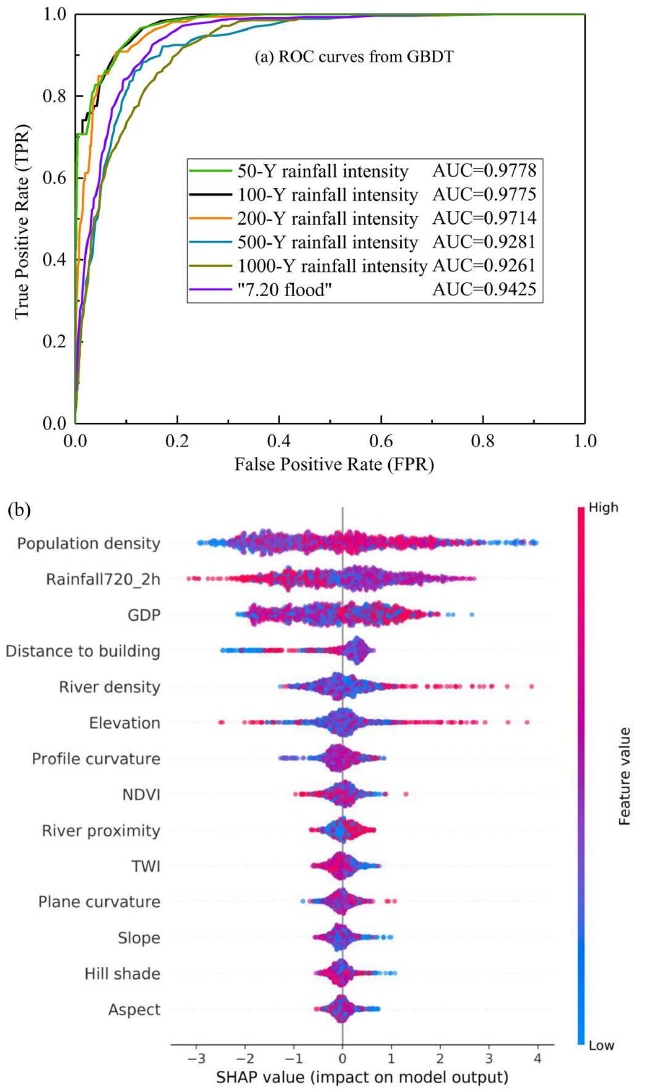

Figure 15 (Color online) GBDT model. (a) ROC curves of GBDT model under different rainfall scenarios; (b) summary plot of SHAP value from GBDT model.

6 Conclusion

This study proposed a new framework to incorporate ML models into GIS to predict flood susceptibilities of metro lines under extreme rainstorms. Six ML models were compared and evaluated according to the flooded metro lines in 7.20 flood in Zhengzhou. The optimum ML model was selected to predict the flood susceptibilities of metro lines under different rainfall scenarios. The major conclusion was summarized as follows.

(1) Floods in underground space are becoming more and more frequent in recent years. It is an urgent request to conduct flood risk assessment and management of urban

metro systems under extreme weather. During the application of ML and GIS to predict flood susceptibility, it is critical to generate a database. This study proposed a new insight to establish a database for training and testing when using ML and GIS.

(2) Machine learning models, including GBDT, RF, KNN, AB, LR, and SVC were incorporated into GIS to predict the flood susceptibility of metro lines. The comparison between predicted flood states of metro stations and recorded data in 7.20 flood indicated that the GBDT model performed the best in predicting flood susceptibility of metro lines.

(3) The GBDT model was selected to predict the flood susceptibilities of metro lines under different rainfall scenarios. The results indicate that the Shakoulu and Haitansi stations are vulnerable to flood. The Shakoulu station will occur flood with a susceptibility of 0.569 under 50-Y rainfall intensity, and the Shakoulu and Haitansi stations will occur flood with a susceptibility of 0.856 under 100-Y rainfall intensity. The numbers of flooded stations were 6, 8, and 13 with flood susceptibilities of 0.856, 0.903, and 0.949 under 200-Y, 500-Y, and 1000-Y rainfall intensities.

(4) The predicted flood susceptibilities of metro lines identified the most vulnerable scales and sections of metro lines, which provide supports to flood risk mitigation and management for the safe operation of the urban metro system.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant No. 42007416), the Key R&D Program of the Ministry of Science and Technology of China (Grant No. 2024YFC3013303), and the Research Grants Council (RGC) of Hong Kong Special Administrative Region Government of China (Grant Nos. E-PolyU501/24, T22-607/24-N).

Funding note Open access funding provided by The Hong Kong Polytechnic University.

References

1 McDermott T K J. Global exposure to flood risk and poverty. Nat Commun, 2022, 13: 3529

2 Aoki Y, Yoshizawa A, Taminato T. Anti-inundation measures for underground stations of Tokyo Metro. Procedia Eng, 2016, 165: 2–10

3 Umar N, Gray A. Flooding in Nigeria: A review of its occurrence and impacts and approaches to modelling flood data. Int J Environ Stud, 2023, 80: 540–561

4 Lyu H M, Yin Z Y, Zhou A, et al. MCDM-based flood risk assessment of metro systems in smart city development: A review. Environ Impact Assess Rev, 2023, 101: 107154

5 Disaster Investigation Team of the State Council (DITSC). Investigation report on "7.20" heavy rainstorm disaster in Zhengzhou, Henan (in Chinese). https://www.sgpjbg.com/baogao/60004.html. Accessed on April 20, 2022

6 Rentschler J, Salhab M, Jafino B A. Flood exposure and poverty in 188 countries. Nat Commun, 2022, 13: 3527

7 Intergovernmental Panel on Climate Change. Climate Change 2022: Mitigation of Climate Change. https://www.ipcc.ch/report/sixth-assessment-report-working-group-3/. Accessed on October 18, 2022

8 Peng F L, Qiao Y K, Sabri S, et al. A collaborative approach for urban underground space development toward sustainable development goals: Critical dimensions and future directions. Front Struct Civ Eng, 2021, 15: 20–45

9 Qiao Y K, Peng F L, Sabri S, et al. Socio-environmental costs of underground space use for urban sustainability. Sustain Cities Soc, 2019, 51: 101757

10 Qiao Y K, Peng F L, Wu X L, et al. Visualization and spatial analysis of socio-environmental externalities of urban underground space use: Part 2 negative externalities. Tunn Undergr Space Tech, 2022, 121: 104326

11 Lyu H M, Sun W J, Shen S L, et al. Flood risk assessment in metro systems of mega-cities using a GIS-based modeling approach. Sci Total Environ, 2018, 626: 1012–1025

12 Lyu H M, Shen S L, Yang J, et al. Inundation analysis of metro systems with the storm water management model incorporated into a geographical information system: A case study in Shanghai. Hydrol Earth Syst Sci, 2019, 23: 4293–4307

13 Lyu H M, Zhou W H, Shen S L, et al. Inundation risk assessment of metro system using AHP and TFN-AHP in Shenzhen. Sustain Cities Soc, 2020, 56: 102103

14 Zheng Q, Shen S L, Zhou A, et al. Inundation risk assessment based on G-DEMATEL-AHP and its application to Zhengzhou flooding disaster. Sustain Cities Soc, 2022, 86: 104138

15 Toosi A S, Calbimonte G H, Nouri H, et al. River basin-scale flood hazard assessment using a modified multi-criteria decision analysis approach: A case study. J Hydrol, 2019, 574: 660–671

16 Balogun A, Quan S, Pradhan B, et al. An improved flood susceptibility model for assessing the correlation of flood hazard and property prices using geospatial technology and Fuzzy-ANP. J Environ Inform, 2020, 37: 107–121

17 Pathan A I, Agnihotri P G, Said S, et al. AHP and TOPSIS based flood risk assessment—a case study of the Navsari city, Gujarat, India. Environ Monit Assess, 2022, 194: 509

18 Espada R, Apan A, McDougall K. Vulnerability assessment of urban community and critical infrastructures for integrated flood risk management and climate adaptation strategies. Int J Dis Res Built Environ, 2017, 8: 375–411

19 Rahman M, Chen N, Islam M M, et al. Location-allocation modeling for emergency evacuation planning with GIS and remote sensing: A case study of northeast bangladesh. Geosci Front, 2021, 12: 101095

20 Luu C, Pham B T, Phong T V, et al. GIS-based ensemble computational models for flood susceptibility prediction in the Quang Binh province, Vietnam. J Hydrol, 2021, 599: 126500

21 Naghibi S A, Hashemi H, Pradhan B. APG: A novel python-based ArcGIS toolbox to generate absence-datasets for geospatial studies. Geosci Front, 2021, 12: 101232

22 Doorga J R S, Magerl L, Bunwaree P, et al. GIS-based multi-criteria modelling of flood risk susceptibility in Port Louis, Mauritius: Towards resilient flood management. Int J Dis Risk Reduct, 2022, 67: 102683

23 Ighile E H, Shirakawa H, Tanikawa H. Application of GIS and machine learning to predict flood areas in Nigeria. Sustainability, 2022, 14: 5039

24 Pham Q B, Ali S A, Bielecka E, et al. Flood vulnerability and buildings' flood exposure assessment in a densely urbanised city: Comparative analysis of three scenarios using a neural network ap-

proach. Nat Hazards, 2022, 113: 1043–1081

25 Pham B T, Bui D T, Prakash I, et al. Hybrid integration of multilayer perceptron neural networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. Catena, 2017, 149: 52–63

26 Tayyab M, Zhang J, Hussain M, et al. GIS-based urban flood resilience assessment using urban flood resilience model: A case study of Peshawar city, Khyber Pakhtunkhwa, Pakistan. Remote Sens, 2021, 13: 1864

27 Mahato S, Pal S, Talukdar S, et al. Field based index of flood vulnerability (IFV): A new validation technique for flood susceptible models. Geosci Front, 2021, 12: 101175

28 Tzepkenlis A, Grammalidis N, Kontopoulos C, et al. An integrated monitoring system for coastal and riparian areas based on remote sensing and machine learning. J Mar Sci Eng, 2022, 10: 1322

29 Tewari A, Kshemkalyani V, Kukreja H, et al. ML and GIS-based approaches to flood prediction: A comparative study. In: Cyber Intelligence and Information Retrieval. Singapore: Springer, 2022. 213–223

30 Dodangeh E, Choubin B, Eigdir A N, et al. Integrated machine learning methods with resampling algorithms for flood susceptibility prediction. Sci Total Environ, 2020, 705: 135983

31 ZZJT (zzjt.zhengzhou.gov.cn). The recovery of metro operation in Zhengzhou. 2021. http://zzjt.zhengzhou.gov.cn/mtjj/5184634.jhtml (accessed on August 25, 2022)

32 China Meteorological Data Network (CMDN). Hour-by-hour observation data of China's ground meteorological stations. 2021. https://data.cma.cn/data/detail/dataCode/A.0012.0001.html

33 Yang Y. Research on the formula of rainstorm intensity of China grid based on remote sensing. Dissertation for Master's Degree. Zhengzhou: Zhengzhou University, 2020

34 Tehrany M S, Jones S, Shabani F. Identifying the essential flood conditioning factors for flood prone area mapping using machine learning techniques. Catena, 2019, 175: 174–192

35 Breiman L. Arcing the Edge. Technical Report. University of California. 1997

36 Friedman J H. Greedy function approximation: A gradient boosting machine. Ann Stat, 2001, 29: 1189–1232

37 Wang Q, Linton O, Härdle W. Semiparametric regression analysis with missing response at random. J Am Stat Assoc, 2004, 99: 334–345

38 Kaiser M, Günnemann S, Disse M. Regional-scale prediction of pluvial and flash flood susceptible areas using tree-based classifiers. J Hydrol, 2022, 612: 128088

39 Ho T K. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Machine Intell, 1998, 20: 832–844

40 Hastie T, Rosset S, Zhu J, et al. Multi-class AdaBoost. Stat Interface, 2009, 2: 349–360

41 Rordam M, Larsen F, Lausten N. An Introduction to K-theory for C*-algebras. Cambridge: Cambridge University Press, 2000. 30–35

42 Deroliya P, Ghosh M, Mohanty M P, et al. A novel flood risk mapping approach with machine learning considering geomorphic and socio-economic vulnerability dimensions. Sci Total Environ, 2022, 851: 158002